Collaborative Project

Integrated Project
Reflective Learning at Work

European Commission Seventh Framework Project (IST-257617)

Deliverable	***D10.3***
Summative Evaluation of MIRROR Appsphere usage and effectiveness at test beds	
Editor	*Bettina Renner, Gudrun Wesiak*
Work Package	*10*
Dissemination Level	*Public*
Status	*Final*
Date	*01 July 2014*

The MIRROR Consortium

Beneficiary Number	Beneficiary name	Beneficiary short name	Country
1	imc information multimedia communication AG	IMC	Germany
2	Know-Center (Kompetenzzentrum für wissensbasierte Anwendungen und Systeme Forschungs. Und Entwicklungs GmbH) Graz	KNOW	Austria
3	Imaginary srl	IMA	Italy
4	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH Saarbrücken	DFKI	Germany
5	Ruhr-Universität Bochum	RUB	Germany
6	The City University	CITY	UK
7	Forschungszentrum Informatik an der Universität Karlsruhe	FZI	Germany
8	Norges Teknisk-Naturvitenskapelige Universitet	NTNU	Norway
9	British Telecommunications Public Limited Company	BT	UK
10	Tracoin Quality BV	TQ	Netherlands
11	Infoman AG	INFOM	Germany
12	Regola srl	REG	Italy
13	Registered Nursing Home Association Limited	RNHA	UK
14	Neurologische Klinik GmbH Bad Neustadt	NBN	Germany
15	Medien in der Bildung Stiftung	KMRC	Germany

Amendment History

Version	Date	Author/Editor	Description/Comments
V0.1	20.05.2014	Bettina Renner	Document generation with headings for sections
V0.2	28.05.2014	Bettina Renner, Gudrun Wesiak	Version for internal review
V1.0	01.07.2014	Bettina Renner, Gudrun Wesiak	Final version

Contributors

Name	Institution
Bettina Renner (editor)	KMRC
Gudrun Wesiak (editor)	KNOW
Marina Bratic	KNOW
Martin Degeling	RUB
Monica Divitini	NTNU
Tobias Dumont	DFKI
Angela Fessl	KNOW
Sandra Feyertag	KNOW
Thomas Kleinert	DFKI
Neil Maiden	CITY
Simone Mora	NTNU
Dalia Morosini	IMA
Lars Müller	FZI
Kristine Pitts	CITY
Michael Prilla	RUB
Veronica Rivera	FZI

Reviewer

Name	Institution
Roy Ackema	TQ
Nils Faltin	IMC
Viktoria Pammer-Schindler	KNOW
Kevin Pudney	RNHA

Legal Notices

The information in this document is subject to change without notice.

The Members of the MIRROR Consortium make no warranty of any kind with regard to this document, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The Members of the MIRROR Consortium shall not be held liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.

Executive Summary

This deliverable reports the methodology and results of 19 summative evaluations carried out during year 4 of the MIRROR project. All of them incorporated methods from the summative evaluation framework presented in the Evaluation Toolbox in D1.5. Four more evaluations conducted during year four are reported in D10.2, Chapter 4 as they are formative evaluations with summative aspects. The conclusions derived from the individual evaluation reports are shortly summarized in this deliverable, whereas cross-cuttings analyses and related insights are described in detail in deliverable D1.7.

All five application partners have been involved in demonstrating the applicability and usability of MIRROR apps for different user groups in different domain scenarios. For some application partners, this has involved supporting evaluations in a number of different testbed sites. In addition some evaluations have been conducted in external testbeds which shows external interest in our apps and thus exploitation potential.

Each of the evaluations has provided its own findings regarding the potential for the reflective learning approaches trialled to bring about positive change within the organisations in which evaluations were conducted. From the evaluations we gained insights regarding all four levels of the Kirkpatrick model, i.e. reaction, learning, behaviour, and results on an organisational level. The reports describe usage data, analyse different aspects of learning such as development of reflection practices and learning outcomes, and describe if and how behaviour of app users changed. Furthermore, in some evaluations also the content of user notes entered into the apps was analysed in line with the reflection coding scheme developed by WP1 and WP 6 during year 4. In every evaluation, partners looked at specific KPIs on which the approaches and apps evaluated may have an effect. Although it was often difficult to prove a direct impact of the reflection approaches on these KPIs there are positive indications that computer-supported reflective learning could increase quality of work, individual work performance of employees, employee satisfaction, and client satisfaction.

Further insights could be gained regarding supporting factors and barriers for the successful introduction of reflective learning. One potential influencing factor is for example the work experience of app users. While some approaches are more appropriate for newcomers in their job, as was already shown in year 3 for the Serious Games, there are other approaches from which more experienced workers seem to benefit more (CaReflect, WATCHiT).

The apps developed during this project do not only target different levels of reflection such as individual, collaborative and organisational, they also focus on different stages in the reflection process according to the Computer Supported Reflective Learning model developed in WP1. During the evaluation it became apparent that it is often important to support users during the complete reflection process, i.e. starting from the collection of relevant data in the plan and do work stage until applying potential reflection outcomes to existing work processes. Although, not every single app was developed to support the whole reflection cycle, it could be shown that users can benefit from using a combination of apps. An evaluation combining KnowSelf and ARA, which focus on different stages of the reflection process, and included a coaching approach showed that user can well benefit from such a socio-technical approach.

Overall, the conducted summative evaluations helped the project to gain valuable insights about how reflective learning can be introduced in organisations and which factors may affect the success. An aggregation of these insights and results from overall analyses of the data presented in this deliverable can be found in D1.7.

Table of Contents

Table of Figures

List of Tables

1 Introduction

This deliverable, D10.3, is produced as part of WP10, and reports evaluations of the designed and implemented MIRROR concepts, reflective learning methods, and corresponding applications. All reported evaluations have been conducted in application domain test beds in the fourth project year. During the project we distinguished three types of evaluations, formative evaluations, formative evaluations with summative aspects, and summative evaluations (for definitions see D10.1). Whereas D10.2 and its update report the results of formative evaluations (with summative aspects) carried out in the project years 3 and 4, all the summative evaluations carried out in year 4 are reported in the deliverable at hand. Results from these individual summative evaluations have been integrated and overall analyses of the data have been carried out. The outcome of this cross-cutting analysis and related insights are summarized as independent report in deliverable D1.7.

This deliverable reports 19 summative evaluations, 18 completed and one on-going, which are grouped according to the following criteria:

- Duration of the intervention: long-term (continuous app-support for several weeks) vs. short-term (interventions set-up as reflection campaigns of one to a couple of days)
- Type of intervention: integrated into the work process vs. training contexts

All five application partners have been involved in demonstrating MIRROR's applicability and usability for different user groups in different domain scenarios, by participating in and actively supporting the evaluation of at least two different apps. Additionally, IMC served as a testbed and also several external partners could be engaged to test MIRROR apps.

This document is structured as follows:

- Section 2 provides a description of the overall methodology of summative evaluations. This includes the summative evaluation toolbox, in detail described in D1.5, as well as the reflection coding scheme developed by WP1 and 6 during year 4;
- Section 3 provides short descriptions of all project internal testbeds, that is the five application partners as well as one additional partner;
- Section 4 provides an overview of the evaluations reported in this deliverable including the classification of long-term vs. short-term, work vs. training, and the business sectors in which the evaluations took place;
- Sections 5 and 6 present the long-term and short-term evaluations, respectively; in section 7 the interim report of one ongoing evaluations is presented;
- Section 8 presents some conclusions gained across all evaluations;
- The Annex contains the items of the summative evaluation toolbox (section 9), the templates used to collect information for this deliverable from all partners (section 10), data overview tables of the evaluations (section 11), and finally further material of interviews from two evaluations (section 12);

2 General methodology of summative evaluations in MIRROR

In section 2.1 we shortly describe the summative evaluation toolbox developed during year 2 (D1.5). The summative evaluation toolbox is based on a modification of the Kirkpatrick model. It allows us to assess relevant indicators for reflective learning and their impact for individuals and teams, as well as the organisation as a whole.

In the following we shortly describe the main components of the toolbox including the four levels corresponding to Kirkpatrick's evaluation model. The exact item formulations can be found in Annex 1. For a more detailed explanation of the content as well as the development of the toolbox we refer to D1.5 'Evaluation Framework'.

Section 2.2 describes the coding scheme for reflective content which was developed by WP 1 and 6 during year 4. The coding scheme was used in several evaluations in order to categorize the written content produced by the users in a consistent and systematic way. For details about the development of the coding scheme we refer to D6.4.

2.1 Summative evaluation toolbox

The toolbox contains a set of core questions, which are applicable to all conducted evaluations and therefore allow comparisons across different apps and/or testbeds. Additionally, a large set of optional questions is provided by the toolbox, which covers IT-attitudes, barriers for app usage, general app effects, behavioural intentions, and many more. The selection of these questions in terms of amount and content was carried out by the app developers doing the evaluations. Thus they are not common to all evaluations and therefore not reported in this section, but in the respective results sections (for exact item formulations we refer to D1.5).

2.1.1 Demographic Information

Demographic data about the participants was used to describe the sample in which an evaluation took place and to connect the participant data across different times and types of measurement. Each question in this and the following sections is labelled with question identifiers to allow for data integration. Here, **CD** stands for **C**ore Question **D**emographic items.

This section contains questions about classic demographic information of participants such as age, gender, job etc.

2.1.2 Level 1: Reaction (Usage)

Although usability is firmly an issue of formative evaluation (see Deliverable 10.2), app usage is a precondition for any higher level in the evaluation model. Thus, we are concerned about the actual usage numbers and qualities.

Objective usage data was received from log files, perceived usage from self-report questions.

Core Question Log **F**ile (CF)

Core Question **U**sage (CU)

2.1.3 Level 2: Learning

The second level of evaluation can be divided into questions concerning the learning process (app-specific reflection questions and short-reflection scale) and questions concerning the learning outcomes.

App specific reflection questions: Core Question App-Specific Reflection Question (CA)

A central aspect of our summative evaluation framework is the actual support of reflection that the apps provide. Thus these questions refer to the four phases of the theoretical reflection model developed within the project (CSRL model, see Deliverable 1.4 and its update D1.6). Each app has been designed to support certain processes during reflective learning. The CA questions are provided to evaluate how the apps addressed these areas of support.

The amount of questions chosen from this section depends on the breadth of functions the app in question provides to support reflection. The relevant questions (functions the app provides) have been chosen by the developers for each app.

Additionally some questions, for which no effect was expected, have been used as control questions. These could either be CA questions where no changes were expected or other question about functions the app does not support. The control questions helped to identify whether the participants only reported positive reactions because they want to please the designers or evaluators (social desirability/demand characteristics). If positive effects are seen only for the supported functions, this gives confidence that it actually is a true effect.

Short Reflection Scale

The Short Reflection Scale (SRS; developed in collaboration with WP4 and WP6) assesses participants' general tendency to reflect and the importance they place on reflection. This scale is different from the App-Specific Reflection Questions in that one does not need to use an app to answer these questions. Rather, this scale allowed us to see whether using the apps changed participants' general tendency or inclination to reflect both individually and in teams. Thus, for long-term evaluations, this scale was used both pre- and post-implementation. In short-term evaluations the SRS was presented only once, because changes in one's general tendency to reflect need some time to become evident. Note that organisational reflection is not included in this scale, as it is intended to assess factors directly related to the individual employee. **CR** stands for **C**ore Question Short **R**eflection Scale item.

The scale contains questions for individual as well as team reflection which can be seen as subscales.

Learning Outcomes

The toolbox contains two mandatory subjective questions regarding the outcome level of reflective learning. **CL** stands for **C**ore Question **L**earning item.

2.1.4 Level 3: Behaviour

To keep the amount of questions as short as possible, we only asked participants a single core question regarding the behaviour level. This core question assesses the central aspect: Did the work behaviour improve?

With "work behaviour" specific behaviours that are the target of the reflection support (e.g., time management) are meant and therefore replaced the more general term "work behaviour" in the item. **CB** stands for **C**ore Question **B**ehaviour item.

If possible we tried to get additional data about KPIs in interviews, with manager ratings, or about KPIs on an individual level.

2.1.5 Level 4: Results

Change over time in test bed-relevant organizational-level KPIs was assessed. The KPIs have been measured starting at or before the first implementation until the end of (or even shortly after) all implementations. Care has been taken that only the relevant unit(s)/personnel who used the apps were assessed, in order to more accurately detect changes. Nevertheless it is important to note that context factors that could not be controlled or influenced by the apps during the project might have dampened eventual changes. As KPIs were to be defined according to the specific test bed and the apps tested and were highly individual to the constellation of the test bed and app we decided not to show a list of possible KPIs here. KPIs used in the evaluations are reported in the respective methods section.

2.2 Scheme for coding of reflection elements

In order to provide a common and systematic way to analyse the qualitative data obtained from users' textual entries into the apps (e.g. in reflection journals, as reflection notes, etc.), a coding schema was developed by WP1 and WP6. The schema considers only reflection elements, i.e. coding sentences or unities of meaning, which are divided into several categories of reflection starting with the simple description of an experience or problem (category 1) to drawing conclusions and implications from reflection (category 9). Text entries that do not contain any reflective element are a priori excluded from the analysis.

IMPORTANT: Codes should be used only based on what is in the text, meaning that no interpretation by the coder should be given. For example, if there is no explicit mentioning of an experience, coders should not assume that someone makes a statement from their experience etc.

1) Description of an experience and mentioning of an issue/problem, including adding to the description (in a comment)
2) Mentioning emotions: reporting how oneself or others felt during an experience (eg. "Was not fun man" or "this made me really angry")
3) Interpretation or justification of actions: explanation or reasons for actions of persons involved in the experience, assessment of the situation (e.g., explanations why the situation is problematic or relevant for work but not only describing the problem (→1)), hypotheses for problems / success (individuals: interpretation added in initial statement or additional comment; collaborative: interpretation added by other participant), for example "Person A started to act nasty on me (code 1). As far as I am aware I had done nothing to deserve this (code 3)"
4) Linking an experience explicitly to other experiences (own or from other persons) (e.g. I made a similar experience or I was told about a similar experience by XY → explicit reference to past experience needed)
5) Linking an experience to different pieces of (own, collective) knowledge, rules, values, organisational documents etc. OR giving advice/solution suggestions not explicitly linked to a particular experience (e.g. never accept blame for another's mistake; google it if you can't get any help)

6) Responding to interpretation of the action (individuals: “inner dialogue”, collaborative: response by other participants)
   a) Inquiry/different perspectives: giving possible alternate perspectives (for individuals mentioning more than one perspective, adding perspectives), for example “you could also do …”, without further explanation / more speculating than in 7a (could, would, ..)
   b) Challenging or supporting (probing) assumptions / opinions / attributions (own and/or others’): “against the backdrop of rationalizing action” (Boland and Tenkasi 1995); for example “Agreed!” or “Hmmm. Is this really different from …”
7) Working on a solution
   a) Explanation of reasons: background, going beyond standard attributions (for solutions: differentiates advice or standard solutions from reasoning) → reasons not given ‘flat’ but based on assumptions, insights, … (Boland and Tenkasi 1995); e.g., ”It is good to do X, because it helps to …”
   b) Giving solution suggestions: Giving possible solutions without proposing to set them in practice (step before thinking about implementation), referring to an experience (e.g. from my experience a list of FAQ’s is useful) – can be the result from past reflection about own experiences, reporting the trial of solutions (e.g. “I suggested that he could do…” or “From my experience you should …”)
8) Insights / learning from reflection
   a) Different / better understanding of experience (single-loop learning): reporting insights (“It is good to know that I personally haven’t done something wrong” or “I realised that I shouldn’t have been so worried about this”), also reporting that a better understanding has happened without explicit mentioning the content of the insight
   b) Generalising from experiences, finding patterns across experiences, considering further aspects beyond the immediate context → “critical reflection” (Hatton and Smith 1995), looking for the roots of a problem (double-loop), e.g. “The best way I have found out to deal with this is …”, differentiation from 7b – insights have to come from the current reflection process but not from past experiences
9) Drawing conclusions and implications from reflection (not from own experiences or knowledge); suggestion to apply new practice (may be on different levels → general or for experience only) – more concrete and final than just giving solution suggestions, discussing how to implement a change, e.g. “Will definitely try and do … in the future”, also changes that already have taken place (e.g. “I have used your advice …”

3 Testbed descriptions

In this section we provide a short description of the organisations and the organisational units which tested apps for the application partners of the MIRROR consortium. For BT also test users and their job roles are reported here, because they apply to all BT evaluations. For the remaining testsbeds, description of users and their job roles differed and thus are reported in the individual evaluation reports. Sections 5.1 -5.7 refer to these descriptions in order to prevent redundancy. Characteristics which were special to a certain evaluation as well as testbeds outside the consortium are described in the according evaluation reports.

3.1 BT

Test bed organisation and the organisational unit

Two call centres of the large telecommunication company British Telecom (BT) in Great Britain served as test-bed for the long-term summative evaluation of the MoodMap App. The two call centres are situated in Scotland and together they have about 450 people. British Telecom is a large telecommunications company, serving customers in more than 170 countries. BT employs more than 300 people (divided in 19 teams) in its Dundee centre and 157 people (divided in 7 teams) in its Alness centre in a range of functions from directory enquiries to residential and business broadband. Such call centres can handle an average of 27.000 calls from all over the UK every day – almost 10 million calls over a year. The task of the call-centre staff is to support incoming product support or information inquiries from customers in an efficient, professional and friendly way.

Test users and their job roles

In the evaluations at BT three job-roles were involved namely advisors, coaches and managers.

The **advisors** are responsible for taking the calls and solving any issues directly with the customers. Their main task is to provide customer service with a high quality, which increases the company‘s reputation and boost sales. Thus, advisors have to handle complex calls and situations, need to have excellent listening and problem solving skills as well as excellent communication skills. Generally it is a stressful job and they face high demanding situations with a daily occurrence.

The task of the **coaches** is to support and train the advisors for their work, mainly in terms of behaviour and sales psychology. The coaches regularly conduct coaching sessions with their advisors. They ensure that the advisors have the necessary skills, knowledge and behaviours to deal with any aspect of a customer call and resolve queries first time. Coaches play an important role on the improvement of the advisors’ performance and satisfaction. Through their coaching, problems and challenges of the advisors can be identified and accordingly actions to improve this can be taken. To this respect, advisors’ mood can be a very good indicator.

The **manager**’s task is to ensure that the advisors are performing against targets, they are reviewing the advisors’ performance and they supervise the advisors coaching and training sessions. Furthermore the manager has to provide feedback back to his/her superior about his/her team as well as has to transfer to the advisors important organisational decisions and instructions he/she receives from his/her superiors. They additionally monitor multiple KPIs regarding the performance related to customers and advisors.

3.2 IMC

Test bed organisation and the organisational unit

The evaluation at IMC was conducted with employees from several departments at the IMC headquarter in Saarbruecken, Germany. With its products IMC offers its clients a wide range of solutions for all business processes in training and education. Also, IMC provides professional services covering the content design and production as well as consulting and managed learning services helping clients to (re)organize their learning processes and to select, implement, adapt and integrate suitable software systems and technologies. IMC operates in the corporate segment as well as in the higher education and schools area.

3.3 Infoman

Test bed organisation and the organisational unit

Infoman AG is a consulting company that consults, sells, and personalizes Microsoft Customer Relationship Management (CRM) Software to help analyse and optimise the marketing, sales and service processes of their customer companies (small and medium enterprises (SME)).

People mainly work in small teams of two to three people. Altogether, the company has about 60 employees, most of them based in the headquarters. However, they have a lot of meetings with customers at the customers' site which require internal preparation and post-processing. Daily work is heavily focused on customers' needs which require a high degree of flexibility and the development of individual best practice. Consulting and sales thus involve a high degree of reflection on interaction with the customer. Therefore, knowledge management and sharing is considered to be a major challenge at this test bed.

3.4 NBN

Test bed organisation and the organisational unit

The Neurological Clinic Bad Neustadt (NBN) is one of the largest neurological centers of excellence in Europe. NBN is a specialty hospital for neurological acute and rehabilitation medicine. Located in the middle of Germany the clinic has in total 284 beds. The teams are specialized in the wide range of neurological diagnostic and therapy to provide our patients the best care.

For the MIRROR Project the NBN selected the Stroke Unit as testbed. The Stroke Unit is a specialized entity of NBN that deals with acute cases of strokes. The time pressure and the daily work with emergencies and their results are a burden for all employees on a stroke unit.

The Stroke Unit at NBN has 10 certified beds, and employs 50-80 mainly female staff, and part-timers. At the Stroke Unit there are working about 40 nurses, split up into three shifts (morning, afternoon and night shift). Their task is to care for the patients and the patient's special needs after a stroke.

3.5 RNHA

Test bed organisation and the organisational unit

The Registered Nursing Home Association (RNHA) is an umbrella organisation for some 1,200 UK Nursing and Care Homes, providing residential and nursing care for a largely elderly population of residents. RNHA homes are spread throughout the UK, and are often run as a single enterprise, although the sector has a number of business 'chains' or brands e.g. BUPA which are run as a group of care homes. As in the rest of Europe, the life expectancy of the average care home resident in the UK is rising, with a concomitant increase in the incidence of dementia. Around two-thirds of nursing and care home residents have some form of dementia, putting an additional strain on the care staff due to the unique and complex challenges such a disability can cause.

A typical RNHA home has 40-50 beds, and employs 50-80 mainly female staff, and many part-timers. With a high turnover of care staff - around 20-25% annually is common for homes in the sector - most homes always have a significant number of inexperienced and new staff.

3.6 Regola

Test bed organisation and the organisational unit

Regola is an Italian company who is leading in the development of software- technology and cloud services in the emergency domain, including ICT systems for emergency centres and volunteering associations. The company is located in Turin and consists of altogether 35 employees distributed over 5 departments. Summative evaluations within the testbed Regola addressed different user groups from the employees at Regola itself to volunteers for emergencies. A more detailed description about the users will therefore be provided in the individual evaluation reports.

4 Overview of Evaluations

This section gives an overview of all summative evaluations done in year 4 and reported in this deliverable. Table 4.1 shows which app was tested in which testbed and also contains information about the duration of the evaluation, i.e. if the evaluation was over a period of some weeks – long-term – over if the evaluation took place as kind of reflection campaign of one or several days – short-term. We also differentiated between the sector the app users worked in, business, healthcare, and emergency. Furthermore, the table shows if the evaluations happened as part of the daily work of users or if it was used within a training context.

The individual evaluations are described according to a general template (see Annex 2), which structures the reports in the following way: each study starts with a short description of the organisational context, i.e. a short description of the target users, their organisation, their job roles and the identified need and potential for reflective learning in their respective work settings. This is followed by some theoretical assumptions regarding the selected approach for reflective learning and how the respective apps relate to the CSRL and the transition model (Prilla, Pammer, & Balzert, 2012; see Figure 10.3.1 in Annex 2). All apps have been developed with the purpose to support the process of work-related reflective learning at one or more stages. In D1.6 the developer's perspective on how technology can be used to promote the reflective learning process is described. This framework includes a conceptual model which outlines the main stages of a reflection cycle and how multiple cycles can be interconnected as well as a set of conceptual tools supporting app development and their use in test-beds. An empirical part reports on how well the model can be used to describe the function of different apps from a developer perspective. In the deliverable at hand the focus is on the user perspective, i.e. for each app there is a concrete description of how the given functionalities can support the users in each of the four model stages, as there are plan and do work, initiate reflection, conduct reflection session, and apply outcome. In the results section of each evaluation it is reported how well users actually feel to be supported by the app for different stages and transitions of the model. An integrative view on how MIRROR apps as a whole support the CSRL model and how this is perceived by the users is given in D1.7. After the theoretical assumptions, the main part of the evaluation reports describe the selected research approach and used methods, as well as the results obtained for each level of Kirkpatrick's evaluation model, namely reaction, learning, behaviour, and results. A discussion section with the inferred implications concludes each evaluation report.

Despite this common structure, there is still some variability regarding the length and depth of individual reports, as the contributions have been provided by different authors.

Table 4.0.1. Overview of the reported evaluations

App	Testbed	Business Sector	Duration	Context	Section
KnowSelf	Infoman	business	long-term	work	5.1
KnowSelf, ARA	IMC	business	long-term	work	5.2
MoodMap App	BT, Regola[a]	business	long-term	work	5.3 and 5.4
TalkReflect	RNHA, NBN, RBKC[b]	healthcare, business	long-term	work	5.5
DoWeKnow	Infoman	business	long-term	work	5.6
IAA, IMA	BT	business	long-term	work	5.7
MedicalQuiz	NBN WS, SU[c]	healthcare	long-term	training	5.8
CaReflect	RNHA	healthcare	short-term	work	6.1
WATCHiT, WATCHiT Procedure Trainer	Regola (Cuneo[d])	emergency	short-term	training	6.2
The Virtual Tutor and Rescue League Serious Games	Regola, RNHA, extern	healthcare, emergency	short-term	training	6.3 – 6.5
Yammer[e]	RNHA	healthcare	long-term	work	7.1

[a] as the users were employees of Regola and not emergency volunteers we counted them as belonging to business

[b] RBKC = London Royal Boroughs of Kensington and Chelsea

[c] WS = Workshop, SU = Stroke Unit

[d] Cuneo = big emergency training event in Italy

[e] evaluation still running to the time this deliverable was created therefore only interim report

5 Evaluation reports of long-term interventions

This section contains the reports of all completed summative evaluations which are long-term, that is the apps were used for several weeks or months. Subsections 5.1 through 5.6 report about evaluations at work. 5.7 describes a training evaluation.

5.1 The KnowSelf Evaluation at Infoman

This and the next section report long-term summative evaluations of the KnowSelf App and the combination of KnowSelf App and ARA app. The KnowSelf application is intended to support reflective learning regarding time management and self-organisation by logging the activities on a personal computer and providing an “AS IS analysis” of how users spend their time at work.

5.1.1 Organisational context

Test users and their job roles

The study participants at Infoman were full-time employees working in six different departments (e.g. CRM consulting, Business Development & Innovation, marketing, etc.) mostly on the management level (e.g. research manager, marketing manager, team leader).

They can be described as knowledge workers conducting a majority of their work using a computer. The rest of the time is structured by meetings, spontaneous or planned, organized with a calendar tool or by communication with customers by phone and e-mail.

Dependent of the status of the project they are working on, the distribution of time spent with different activities (computer work, meetings, phone calls etc.) can vary significantly. Accordingly, they have to evince high flexibility and to organise a broad range of different tasks and responsibilities, to keep track of important dates and deadlines and have to be self-responsible for their own productivity.

Identified need and potential for reflective learning

During a normal workday there are some possibilities for reflective learning to take place, e.g. when the participants exchange their views on different experiences they had (e.g. after having a telephone call with a customer) or discuss with each other on the hallway or in meetings difficult projects or situations. User studies in Y1 showed that although they acknowledge reflection about their work as a good thing, it was said that their daily work does not allow them enough time to do it regularly. This could be especially important when thinking about time management, a basic skill that knowledge workers need to have. As the study participants are required in their jobs to self-organise their work, and manage their time, investing energy in improving time management means creating possibilities for reflective learning which could result in optimised workflow and higher work performance.

Potential organizational impact

The KnowSelf App is focused on capturing data about individual work practice and supports individual learning by reflection. Therefore, we expect that using the KnowSelf App will give to individual participants new insights regarding their personal time management (e.g. time fragmentation, time spent per project or document); it will help them to assess how they spend time at work, motivate them to introduce some changes if they interpret some behaviour as inappropriate or if they cannot fulfil their working tasks. The individual employee may decide to transfer these insights based on individual reflection to team or organizational level, which

could result in higher work satisfaction of a team (if they together resolve a shared time management issue) or in higher work productivity (as a result of an optimized time management) of individual workers as of whole teams.

5.1.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

The KnowSelf application is intended to support reflective learning regarding time management and self-organisation. The current version of the app logs the activities on a personal computer (PC), provides an “AS IS analysis” of how users spend their time at work, and help them to reflect about their past working days and how they develop over time. This personal time management analysis tool is targeted towards knowledge workers who work mainly on a PC.

The participants in our evaluation study were asked to start the KnowSelf App every day for six weeks on their computer at work (or to use “Auto start” option). The app was running in the background while they worked and captured automatically window focus and focus switching on a PC. For each window in focus, it also determined the window title and if applicable the path to the window resource.

The participants were asked additionally to track manually their task-related activities via the stop-watch functionality whenever it seemed appropriate, and at the end of each working day to review their time use during the day, and write down observations, insights, and plans for change (approx. time investment: 10 min per day).

Documenting observations on work practices and time management, as well as solutions, experiments etc. was possible via the “Reflection diary” window in the app.

The timeline visualization in the KnowSelf App, which shows all the resources used within one day or some other time range selected, should stimulate individual reflection. When looking at all applications and resources used that day the user could compare it with planned activities, could ask questions about the best (personal) way to complete assignments, if there were enough pauses, how fragmented the job actually was etc.

Relation to MIRROR CSRL Model

The KnowSelf App can be related to all four phases of the CSRL Model. In the 1st phase ‘plan and do work’ the participants conduct their work in the usual manner whereby the app records these activities. The captured data enable consequent reconstruction of individual work time and review of their work and learning history (transition ‘data’).

The 2nd phase ‘initiate reflection’ focuses mainly on setting the goal for reflection. The user is free to set an objective for reflection and to decide how much time he/she will invest. The app helps to set new objectives in that it enables gaining insights regarding personal time management, e.g. comparison between personal calendar (should do) and actual recorded activities (has been done) or tracking the time management goal achievement with the help of the app. The data recorded and visualized by the KnowSelf App should serve as a trigger for the next phase.

In the 3rd phase ‘conduct reflection session’ the participants make use of the data the KnowSelf App captured about their work behaviour and reflect on the data having in mind the goals they set themselves in the previous phase. By selecting different working days and comparing them with each other, it is possible to compare similar work experiences and situations. All insights

gained in this phase can be inserted into the Reflection Diary and serve as a basis for planning the future changes (transition 'outcome').

In the 4th phase 'apply outcome' the user decides, based on the reflection outcomes from the previous phase, whether some changes in the work life should be implemented and in what way. Again, Reflection Diary can be used to set goals for behavioural changes which can be reviewed at a later point and all goals checked as soon as they have been accomplished.

If a user realises that some other aspects of personal time management are suboptimal he/she could decide to invest energy to implement various changes which led to subsequent reflection cycles.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

The focus of the KnowSelf App is to capture data about individual work practice and to foster individual reflection. Considering the transition model, it can be said that the KnowSelf App provides relevant material which can trigger individual reflection process (step 1). The reflection outcomes may lead to consecutive reflection processes (recursion into step 2) and to individual application of outcomes (step 3). Naturally, the individual employee may decide to transfer insights and outcomes based on individual reflection to team or organisational level, but this transfer is not directly supported by the app.

5.1.3 Research approach

Design and procedure

An Infoman representative introduced the app at a kick-off meeting to the participants, presented the MIRROR project and explained the concept of reflective learning. KNOW prepared for the participants instructions for the study, installation files for the KnowSelf App, user guide and a 'cheat sheet' (a short overview of the planned activities, timeline, and the basic app information: installation, the main app functionalities, contact in case of problems).

They filled in the pre-questionnaire which served as a baseline measure before the intervention in order to compare it with the post-intervention measures. The participants were asked to use the app over a period of 6 weeks (January 16th – beginning of March 2014) and to reflect on the collected data on a daily basis.

By means of progress monitoring during the 6 weeks the following measures were conducted: a weekly reminder email with a link to an online questionnaire was sent (questions regarding app usage, reflection, learning effect) which on one side supported the regular use of the app reminding the participants to reflect on the captured data and to enter observations, and on the other side provided us with continuous feedback.

After using the Know Self App for 6 weeks the study participants were asked to fill in the online end-questionnaire. The users were asked as well to send us their anonymised log files documenting their app usage. In order to discuss their feedback on the app individually and deepen our understanding of the results we conducted follow-up interviews with some of the participants.

Participants

Twelve Infoman employees, four females and eight males, participated in the summative evaluation of the KnowSelf App. The median age was 20 to 29 years, with the youngest participant younger than 19 and the oldest 40 to 49.

At the time of the evaluation all of them were employed full-time in different departments with three-quarter of them on the management level (e.g. research manager, marketing manager, team leader etc.). On average the participants have been working in their current positions somewhat longer than one year with total work experience in this field of about 4 years.

Summative evaluation methods used

Out of the summative evaluation toolbox (see chapter 3) core-questions from all four levels were chosen (reaction, learning, behaviour, results); more specifically these sections were covered: demographic items, level of participation, short reflection scale, app-specific questions, learning outcomes, work-self assessment, and loyalty metric. Some questions beyond the Core Questions were as well used: Usage (USE 01, 03, 06, 07), Usefulness/Satisfaction (SAT 01, 03), Inclination Long-Term Usage (LT 01, 02), Knowledge/Skills (KS 01, 02, 04), Work (WK 01, 05, 06, 08, 09, 13, 14), and work satisfaction. Additionally, in the post-questionnaire we addressed the KnowSelf prompts with a few questions while usage and some learning and behaviour aspects were assessed throughout the evaluation by use of weekly questionnaires as well as by the analysis of the reflection diary entries.

As an individual KPI measure we used a self-assessment of their personal time management measured with the time management scale by Hansen (2001)[1] before and after our intervention. We also asked participants to assess their time wasters before and after the intervention to see whether this assessment will change. Time-wasters in the everyday work were identified using the scale by Seifert (1999)[2].

With respect to the organisational impact of time management tool, the following KPIs were measured: subjective assessment of personal time management and time wasters as well as the job satisfaction in a pre/post comparison.

5.1.4 Results

In the following the results are presented according to four Kirkpatrick levels (Reaction, Learning, Behaviour, Results). Different number of participants provided answers to the evaluation tools we applied: pre-questionnaire (n=12), post-questionnaire (n=10), interviews (n=7) and app log data (n=7).

5.1.4.1 Level 1: Reaction (Usage)

According to the self-reports of the users at Infoman they used the tracking tool on average 3.7 days per week and the time investment was about 25 minutes weekly (see Figure 5.1.1 for more detailed depiction of subjective usage of the app).

[1] Hansen, Katrin: Zeit- und Selbstmanagement, 1. Aufl., Berlin, Cornelsen, 2001
[2] Seifert, L. J., Das 1x1 des Zeitmanagements, 1999

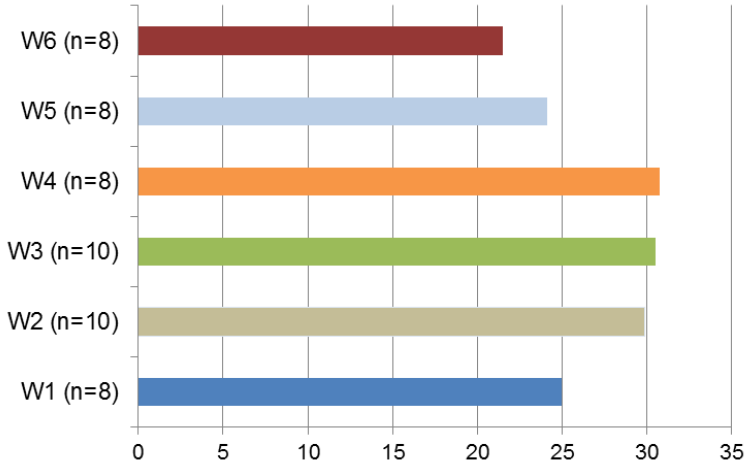

Figure 5.1.1. Subjective usage over time (week 1 to week 6) in minutes

Using the Friedman's ANOVA test we compared their weekly-questionnaire self-reports of the time they invested in the app and we found a decrease over time: the usage was significantly lower in the last two weeks than in the time before ($\chi^2(1,N=8) = 4.5$, $p=0.034$).

Interestingly, the evaluation of the log data we received from seven out of twelve participants showed an average app usage of 1.14 days a week and an average weekly time investment of 6.42 minutes. In the Figure 5.1.2 we compared the subjective usage based on the self-ratings collected in the weekly questionnaires and the objective usage based on the log data we received. Different number of participants provided data for particular weeks (e.g. Week 1 (n=8/6): 8 participants provided their subjective rating while 6 participants shared their log data for that week).

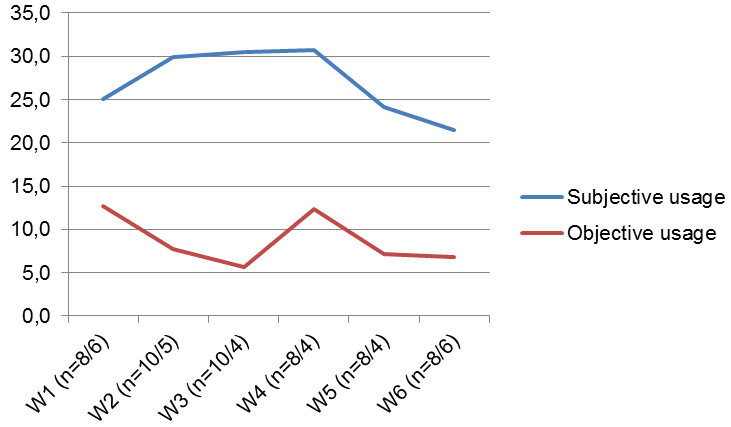

Figure 5.1.2. Comparison of subjective and objective usage (min/weekly)

One possible explanation for the discrepancy between objective and subjective usage could be that the evaluation duration was six weeks, but not all participants had the app running for the whole time frame (e.g. due to absence from work, etc.). Also some participants claimed in the interviews as well as in the questionnaires that the app didn't capture the data properly and that their recordings for several days were incomplete or missing completely. This was confirmed by a comparison of their activity log data to the reflection diary entries they entered, because on several days it occurred that they wrote diary entries (which is a proof for using the app), but their activity recordings including data about their app usage were lost due to

technical reasons. This finding could provide the most realistic explanation why the objective app usage calculated from the log data is distinctly smaller than the subjective usage the participants assessed themselves, because even when the activity recording didn't currently work properly for us to check the exact usage later, participants could still use the app for reflecting over existing data, tracking tasks and storing their reflective thoughts.

We asked additional questions regarding possible issues that could have stopped them from using the app more intensively. Feedback from the questionnaires as well as the feedback received from interviews showed that some of the participants experienced either lack of time (USE1: I did not have the time to use the app; M=3.10 SD=1.2) for reflection during their busy working hours, especially as analyzing the amount of detailed data or manual task tracking was perceived as very complex by a third of the participants, or lack of motivation because about a half of them did not see a clear advantage for themselves from using the app (USE03: I did see no advantage in using the app; M=3.40 SD=1.08). Additionally, some of them felt uncomfortable with the fact that the tracking tool recorded "every move" on the computer (M=3.40 SD=1.27).

One third of the participants experienced in the beginning a few technical problems with the KnowSelf App (e.g. problems with Auto Start and required Java version, web-view of the app etc.) but some of these issues were resolved in the first weeks. Also some of the features that were actually implemented in the app, could not be found intuitively and were thus not used by the participants. According to participants' statements these issues did affect their inclination to use the app.

Satisfaction was measured with two items using the 5-point Likert scale from strongly disagree (1) to strongly agree (5). The participants were moderately satisfied with the app (SAT01: M=3.00 SD=0.82). Seven of 10 participants reported they believe that the app can be used to complement professional training (SAT03: M=4.10 SD=0.87).

We collected also feedback to the app on a general level: what was good or what needs to be improved. The participants reported in the first weeks (in the weekly questionnaires) that their first impression of the app was that it has a good design and structure and they saw its potential for giving data overviews which trigger reflection and for supporting the awareness of their time management. However, they asked for some kind of help with the interpretation of the captured data in order to be able to draw some conclusions from it as well as recommendations what could be changed to optimise the behaviour patterns and some ideas how to do it. As already mentioned above, some specific features could not be used by the participants because of a lack of visibility, e.g. synchronization with their calendar for the sake of a comparison and an automatic takeover of the meetings they entered.

What they really liked are the visualizations and having the overview of their work routine. And in general most of them stated they believe that time management apps can help improving one's time management to some extent mainly by raising the users' awareness for different aspects of their time management.

We also asked which employees could especially profit from such an app: in their opinion these are the employees who are involved in several projects and have assignments from several areas. The app could help them to get an overview of their tasks and the time they invest in each of them. If an employee works customer-dependent and often cannot influence how the day will be organised (many external influences) the app will be less helpful or sometimes even annoying if it reminds the user to change or influence happenings that are beyond their power. Also in the interviews some participants stated that the app is especially useful for people who are rather unexperienced regarding their time management. For people who have already

been engaging in reflection about their work patterns before, the collected data will not necessarily deliver any new insights.

When asked about their inclination of long-term usage the majority was undecided whether the app can provide some long-term advantages in their work-life (LT01: M=2.90 SD=0.74) while only one person would like to use the app continuously as part of his/her work life in the future (LT02: M=2.30 SD=1.01) mostly because of the possibility to track the computer activities and resources used for a specific project. Reasons for not using the app included having already enough knowledge about one's working behaviour, following a very structured work process, or having a job with high spontaneity and a lack of structure.

5.1.4.2 Level 2: Learning

The participants gained new insights regarding their time management on the one side through the usage of the app and on the other side because of the request to think about their time management on a regular basis.

About half of the participants stated in the interviews that the data collected by the KnowSelf app most intensively triggered their reflection in the beginning (in the first days and weeks) but that the amount of gained insights and the added value decreased after using the app for some weeks and already having applied some changes. Not being able to draw new conclusions out of the data was especially a problem for the participants who already had practiced reflection before and are familiar with their own working patterns.

In the next two sections we report on the learning process and learning outcomes in more detail.

Learning Process

The App-Specific Reflection Questions (CA statements) regarding the KnowSelf App showed that the app helped the participants the most by collecting information relevant to reconstructing experiences from work (CA1), by reminding them to reflect (CA10) and by providing relevant content for reflection (CA40). With the help of this relevant material they reflected about various issues, such as interruptions, work fragmentation and their working processes in general as well as the question how they could optimize those.

The control item where no change was expected (the app providing information about related experiences) was rated the lowest and hence confirmed the true effect of other statements (CA16).

Average of all five CA statements showed that the participants were moderately satisfied with the reflection support the KnowSelf App provided (see Table 5.1.1 for detailed statistics).

Table 5.1.1. App-Specific Reflection Questions

	Mean	SD
CA1	3.40	0.96
CA7	3.20	0.92
CA10	3.40	1.08
CA12	3.00	1,16
CA16 (control item)	2.10	0.88
CA40	3.40	0.70
CA_all (without control item)	3.28	0.66

The short reflection scale using a 5-point Likert scale was applied to measure the participants' tendency to reflect before and after the evaluation period. The ANOVA test for repeated measures showed that the main effect of type of reflection (individual vs. team) was statistically significant (F(1,9)=14.10 p=0.005); specifically, it showed that the participants' tendency towards individual reflection was significantly higher than towards team reflection in our data sample ($M_{ind.}$=3.72; M_{team}=2.91). However, the second factor (time: pre vs. post) was not statistically significant (F(1,9)=0.123 p=0.734) as well as the interaction between the two factors (F(1,9)=0.310 p=0.591).

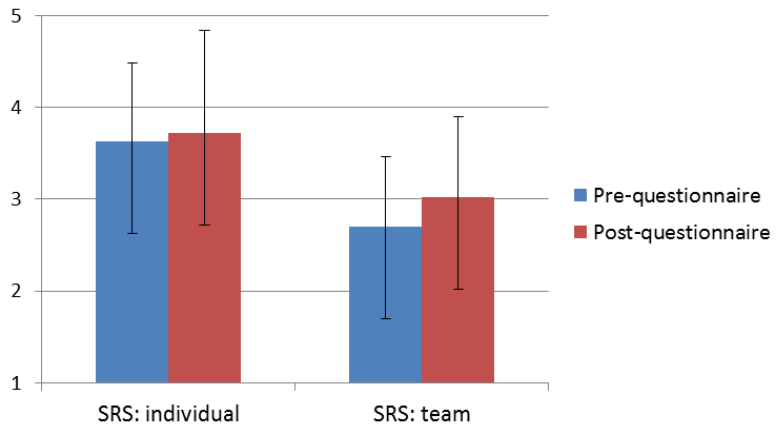

Figure 5.1.3. Mean reflection scores (individual and team reflection) in pre-post comparison

As we can see the reflection tendency before and after the evaluation period did not significantly change over time. This was confirmed as well in the weekly questionnaires: we collected every week their subjective ratings of their tendency to reflect on time management and way of working, and these ratings remained stable during the evaluation period (Friedman test: $\chi^2(2,N=8) = 2.61$, $p=0.88$). However, in the interviews when talking about reflection in more detail we found out that one third of them report they reflect more, mainly in the beginning and end of their work days as well as when interruptions or unusual events occur.

The reflection process was supported additionally by one reflection guidance component of the KnowSelf App: the KnowSelf Prompts. The current version of the KnowSelf App included

an automatic notification system which reminded the participants to reflect at a particular time of a day or after some unusual event was recognised by the app (e.g. high number of switches or long periods of idle time). We asked the participants to what extent were these automatic notifications helpful for their reflection process (see Table 5.1.2). The results show that the users did not profit from the prompts as we have expected it; they perceived the notifications often as disruptive during work and as additional source of work fragmentation. Most positively were evaluated the reminder regarding the most used resources and the general reminder to reflect about the data in the app. In the interviews it was suggested that the prompts could pop up each day at the same time or even better a summary of all notifications could be listed somewhere directly inside the app so a person could look at it when the time was right.

Table 5.1.2. Assessment of Know-Self Prompts

	Mean	SD
Please indicate your agreement with the following statement.		
The KnowSelf Prompts motivated me to reflect.	2.00	1.06
To what extent were the following categories of KnowSelf Prompts helpful to you?		
Reminder for using KnowSelf generally (e.g. look at heatmap, write diary entries)	3.22	1.09
Reminder of project recording	2.75	1.28
Notification about specific amount of switches	2.30	1.34
Notification about unusual amount of idle time	2.57	0.79
Notification about most used resources	3.30	1.25

5.1.4.3 Learning Outcomes

After using the KnowSelf for a period of six weeks, more than half of the participants indicated that they have improved their understanding in the area they wanted to improve (KS02: M=3.60 SD=0.52). Four of them improved also their work related skills in that area (KS04: M=3.40 SD=0.52).

Further, six of ten persons agreed that they made a conscious decision about how to behave in the future regarding their time management (CL01: M=3.30 SD=1.06) and gained a deeper understanding of their work life (CL02: M=3.50 SD=0.71).

The participants reported in the post-questionnaire and in the interviews several ‘lessons learned’, such as: to daily set priorities, handle interruptions in a more conscious way, reserve time for the tasks with the highest priority and to plan some time for the unplanned occurrences etc. When talking about “time wasters” most of them mentioned that Outlook causes constant interruptions and that handling e-mails and appointments in Outlook consumes a lot of their time. Other problems they identified were high work-fragmentation, high number of distractions and the need to frequently reprioritize tasks due to new unpredictable ones or bad planning.

Some participants claimed that gaining these insights motivated them to set goals for themselves and stick to these plans without getting distracted or switching to other tasks so

often. The acceptance of documenting these goals with the help of the KnowSelf's Reflection Diary grew over the first half of the evaluation, as can be seen by only 16 diary entries stored during the first two weeks, but already 48 entries written during week three and four. This number then remained constant for the end phase of the evaluation with another 47 entries made in the last two weeks. Of course these numbers again refer only to the seven participants who made their log data available to us. In total we received 139 diary entries, but 28 of them were excluded from the following analysis, because they were stored after the end of the official evaluation time, while the participants still continued using the app.

Besides goals, the participants documented in the Reflection Diary their experiences, insights and comments concerning the app. We analysed those entries using the qualitative analysis schema developed by WP1 and WP6 (see schema description in Deliverable D6.4). The content was analysed with respect to whether reflection has happened, how deep the reflection was and (if available) which app-specific aspects the content contains.

The coding was conducted by three independent coders who in the first round of coding had consistency regarding different categories of reflection elements between 62% (category 3) and 95% (category 4). A possible reason why the intercoder consistency varied between different categories up to 30% could be the relative novelty of this coding schema and overlapping of a few categories or different understanding of categories by different coders.

There were 103 statements by 6 participants, who were willing to share their explicit data. The reflection diary data of the seventh participant who shared their log data with us was hashed, so its content was not available to be considered in this analysis. In total 33 diary notes were classified as non-reflective (no reasons or critical interpretation; senseless or non-answerable content) while for another 11 entries the coders couldn't agree on whether they were to be interpreted as reflective or non-reflective (these items were excluded from the dataset for later analysis). The remaining statements (59) were classified as individual reflection items (no other actors involved) containing various reflective elements.

The final results are based on the coders' agreement: after independently coding the material the coders discussed the differences and tried to come to an agreement. The final codes are presented in the Table 5.1.3. It was possible to assign more than one code to one entry since some statements contained more than one reflection element (see Table 5.1.3: Column Frequency). This table only contains 57 of 59 reflective entries, because for two entries the coders could not reach an agreement. For each entry we also specified the highest category reached (the highest level of reflection reached by this entry) which was assigned to basic, medium or high level of reflection (see Table 5.1.3: Column Level of reflection containing learning).

Table 5.1.3. Analysis of reflective content

Categories of reflection elements	Frequency	Level of reflection containing learning

1. Description of an experience	54	Basic level (13)
2. Mentioning emotions	0	
3. Interpretation or justification of actions	33	Medium level (28)
4. Linking an experience explicitly to other experiences	1	
5. Linking an experience to different pieces of knowledge, rules, values, organisational documents	0	
6a. Responding to interpretation of the action (inquiry/different/alternate perspectives)	1	
6b. Responding to interpretation of the action (challenging *or supporting* assumptions / opinions / attributions)	0	
7a. Working on a solution based on assumptions, insights (explanation of reasons)	7	
7b. Working on a solution (giving suggestions without proposing to set them in practice/referring to an experience)	4	
8a. Insights / learning from reflection (different / better understanding of experience)	5	High level (16)
8b. Insights/learning from reflection (generalising from experiences, finding patterns across experiences)	6	
9. Drawing conclusions and implications from reflection	8	

Analysing the individual categories of reflection elements from Table 5.3 we can conclude that most of the participants in their reflection diary entries described their work experiences and occurred issues (category 1). According to the qualitative analysis schema this corresponds to the basic level of reflection (provision and description of experience, but no (explicit) traces of reflection).

The second most documented category was ‘interpretation or justification of actions’ (category 3): we found a lot of interpretations and reasons for their successful as well as problematic experiences.

As the assignment of more than one category to one diary entry was possible, for calculating the level of reflection containing learning only the highest category assigned to one entry was considered. Resulting from this, most of the reflection outcomes reported in the reflection diaries can be assigned to the medium level of reflection: participants didn’t only describe their experiences, but they were mostly providing explanations of reasons for their experiences as well as solution suggestions to observed problems too. Also more than a quarter of the reflective entries documented insights gained and conclusions drawn from personal experiences.

These findings confirm that the participants gained new insights and a better understanding of their work experiences and provided possible reasons for those experiences. These new

insights and conclusions could be then used as a basis for application of changes in order to improve their time management.

5.1.5 Level 3: Behaviour

As described above, the participants not only gained insights regarding their time management (level 2), one half of them also improved their time management with the help of the KnowSelf App (M_{CB01}=3.40, SD=1.08).

Additionally, about 70% of the participants reported that they have used their learning regarding time management on the job (M_{WK01}=3.5, SD=1.18), focused more on their work tasks with the help of the KnowSelf App (M_{WK09}=3.90, SD=0.74) and kept up their change of behaviour (M_{WK05}=3.67, SD=0.50).

These questionnaire data were confirmed in the interviews: half of the participants claimed to have changed their behaviour at work to some extent in order to work more efficiently. Here are some of the changes they reported to have implemented in their work life:

- Time planning
  - more aware when making plans (e.g. plan time for the high-priority tasks);
  - more realistic time planning (not dealing with too much tasks at once);
  - planning the day consequently and assigning certain amount of time to specific tasks;
  - handling e-mails only in time-slots defined for it;
  - planning time for the unpredictable occurrences.
- Interruptions
  - saying “no” to tasks and people;
  - dealing with interruptions more consciously (e.g. not always instantly switching tasks when a colleague asks for something or reacting instantly on urgent mail).
- Work fragmentation
  - work more focused on one task.

Although the participants reported numerous positive changes regarding their time management and the higher awareness for these topics we did not find a statistically significant change when analysing the pre- and post-measure of the Time management (pre: M=3.60 SD=0.6, post: M=3.58 SD=0.40; t(9)=-0.10 p=0.92) and Time wasters scale (pre: M=1.89 SD=0.33, post: M=2.10 SD=0.45; t(9)=1.30 p=0.23). This result means that they assess their time management and dealing with time wasters the same as before. On the one side it is possible that they were already very skilful regarding their time management and handling the time wasters what limited their area for improvement or motivation to achieve significant changes in this area. On the other side, we collected various statements about gained insights and implemented changes; however, it is possible that these changes were not global enough to change the appraisal of the whole set of skills time management represents. Another relevant aspect is the limited time of the evaluation and the importance of the long-term aspect when dealing with behavioural changes such as these.

Asked about the effect of the applied changes on their work one quarter of the participants claimed in the interviews that the quality of their work had improved and two participants stated that the changes had at least enhanced their efficiency. In general most of them were convinced that a more structured and conscious way of working over longer periods of time will lead to a better quality of work.

Concerning the influence of the changes on their satisfaction at work, most of them claimed they are in general more satisfied when they have less stress and fulfil their planned tasks during a normal work day. Work satisfaction is for most of them a complex issue depending on many factors (internal or external) and cannot be changed very easily. Nevertheless, one quarter of the participants said in the interviews that the changes they applied also had a positive effect on their satisfaction with their work. There was also one statement about decreased satisfaction because as the participant finished his tasks faster he also got more tasks, which he didn´t want. For the majority the work satisfaction (measured with the questionnaire item) remained stable during the evaluation period.

5.1.5.1 Level 4: Results

When asked how likely is it that they would recommend the KnowSelf App to a friend or colleague we found out that one person out of 10 would actively recommend it (promoters: scores 9-10) and passively another 30% (scores 7-8). The rest 60% of the participants would not recommend it (detractors) which of course influenced negatively the Net Promoter Score (accounts to -50%).

Possible explanations why not more participants would recommend the app could be that some of the participants did not see a personal benefit in using the app because they missed guidance and some app features and therefore would not recommend it. However, asked about their future behaviour five participants claimed that they will try keep up the positive changes regarding their work behaviour and time management as effectively as possible in the future, too.

Asked about the effects on organizational level one third of the participants is of the opinion that all improvements achieved on individual level will aggregate to an improvement of the whole organization. Also four out of twelve participants stated that more employees should improve themselves by practicing reflection or participating in time management measures. Two participants pointed out that improving individuals' time management and efficiency is not enough, but that collaborative communication has to be considered too (e.g. handling of meetings and interruptions, expectance of constant availability).

5.1.6 Conclusion & Discussion

The intervention in this test bed included KnowSelf App usage by the 12 participants on a daily basis during 6 weeks. As the log data show the app was not used as intensively as expected which could be caused by technical problems in the first weeks or lack of clear benefits resulting from this experience.

If we look back at their expectations stated in the pre-questionnaire we find that they expected gaining deeper insights on their time management and work routines as well as an improvement of their time and task management which should result in improved work performance. While we can confirm that the participants gained new valuable insights with the help of the app and that these insights motivated some of them to implement different changes, such as more consequent and realistic planning, better handling of interruptions and e-mails etc., this was not sufficient to affect their assessment of time and task management on a more general level.

Further, one third of the participants stated before the intervention they wish for support with task prioritizing and management and gaining a deeper understanding of how they spend their time at work. The latter expectation was met for most of the participants since the majority expressed that they improved their understanding in the area they wanted to optimise. The

former expectation regarding the support was apparently missing because the results show clearly that the data tracking is sufficient for getting an overview, for triggering the reflection about these issues and getting the first idea what are the areas for personal improvement. It is also clear that without a sort of either technical support with the data interpretation or human support (e.g. coach or trainer) it is challenging for the users to understand this amount of data they are confronted with, to draw some conclusions out of it and decide what actions could be taken. Some of them were overwhelmed, some of them lacked motivation after a while because they were not getting further and had a feeling they are stuck in this situation without an idea how to proceed.

What they would find helpful is on the one side a support with how to understand and interpret the data (e.g. to see some patterns or trends over time) and on the other side to receive some recommendations for actions (suggestions, tips) for their specific situation.

It is of great importance as well that the app requires only a minimum of interaction during the day since a lot of them find it very disruptive to be interrupted during the normal work flow when they try to work focused on their task.

If the app would manage to do all that and support them profoundly with their time management they would recognise it by less stress in their everyday work life, more spare time and improved work efficiency. As another result of an improved time management the participants named the ability to prioritize tasks better and working according to those priorities effectively as well as experiencing positive insights when looking back at a work day.

In summary it can be said that the KnowSelf App supported the participants mostly by providing them relevant material for reflection and helping them to reconstruct their work history. These captured activity data served as trigger for initiating reflection and provided basis for the reflection session (CSRL Model phases 2 and 3, see 5.1.2). This enabled the participants to reflect more profoundly on their work behaviour and time use which was shown by many examples of gained insights and implemented behavioural changes. The remaining phases were also supported to some extent, since some participants used the app's Reflection Diary to make plans about how to slightly change their working behaviour and several times also documented there the changes they had applied and the resulting experiences. In this regard our expectations were met and as well those of most of the participants. However, we see need for additional work in the future in the area of comprehension and interpretation of the data and understanding better how the participants could profit even more from the analyses and visualisations the app can provide. One possibility could be to implement additional interpretation help in the app in the form of tips and recommendations based on the specific set of data. Human support (coach or a person supporting them during the time of app usage) could also be helpful to derive explanations for captured behaviours together and to determine possible actions which could lead to more productive work life and improved work life balance.

5.2 The Knowself/ARA App (part of the Time Management Coaching Approach) Evaluation at IMC

In this evaluation the Knowself (see Section 5.1) was used together with the ARA App, which enables documenting learning and reflection outcomes, goal achievement, application of time management rules and experiences regarding that matter.

5.2.1 Organisational context

Test users and their job roles

The study participants at IMC are mostly full-time employees working in different departments with half of them working on the management level. They can be described as knowledge workers conducting a majority of their work using a computer. For most of them, their workdays are highly dependent on the status of the project they are working on. Days in the office are structured by meetings, spontaneous or planned, with both colleagues and customers. They have to organise daily a broad range of different tasks and responsibilities, to keep track of important dates and deadlines and are self-responsible for their own time management, productivity and results.

Identified need and potential for reflective learning

Most IMC employees are self-responsible knowledge workers and thus need good time management skills. Improving time management skills is not easy, as it requires new techniques to be learned as a theoretical concept and then adapted and applied well in work practice. Like other behaviour changes this change relies on own motivation and self-control. Both reflection and external support by a coach can be of a great help to support the transfer from knowledge to work practice. The need for improvement of time management skills was confirmed in the pre-interviews. In these interviews several topics were addressed: current work situation regarding time management, challenges regarding time management and expectations towards this coaching experience.

One time management issue they experience daily are interruptions at work. Although it is sometimes inevitable to be interrupted, the interviewees admit that interruptions degrade personal productivity as it forces them to take additional time to focus their thoughts again in order to resume the task. Another major challenge is working overtime. Almost all of them experience it on a weekly basis and they would like to better understand the possible causes and resolve this problem. The goal is to learn to work more focused and to be more in control of one's own time. Planning skills were mentioned since several of them have problems with realistically planning time for their tasks, appointments etc. This leads often to another challenge and that is working under time pressure. They expressed their wish to reduce this time-pressure at work and learn some strategies and methods which could help to overcome this problem. As these issues have consequences on their private life and work-life balance, they anticipate that if they would improve their time management and self-organisational skills it would positively affect both their personal and professional life.

Potential organizational impact

This coaching approach targeted time management skills of individual employees. As such no direct organizational impact was anticipated. But time management skills have a strong impact on personal productivity. In case the time management coaching is introduced to a large number of employees, the individual productivity gains could add up to a substantial

organizational productivity gain. In addition, employees with good time management skills would have better control over their work life and by this a stronger feeling of self-determination. Better productivity and self-determination are expected to lead to higher employee satisfaction. Again, if this happens for many employees, it becomes an organizational KPI impact. Satisfaction will lead to a better culture and higher employee retention. They have an impact on organizational productivity, innovation and saved cost.

But there are also risks. Individual reflection can reveal problems on the team and organizational level (like having too much work assigned). If they are addressed well they can in turn have a positive organizational impact. If not, they can increase personal frustration, and if it happens for many employees, with negative impact on the organization. It would be expected that the positive effects would be much stronger than the negative ones.

5.2.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

The time management coaching approach combined the usage of a computer activity tracking tool (KnowSelf or ManicTime) and the Activity Recommendation App (ARA) with a weekly coaching session. The participants were asked to use the tracking tool every day for six weeks on their computer and at the end of each working day to individually review their time use during the day, to reflect on their time management skills and to assess whether they achieved the goals they set themselves (e.g. work more focused on one task). The tracking tool provided them various visualisations of the resources and applications they used within one day or some other time range selected, and allowed an overview of their work activities. The ARA App enabled documenting learning and reflection outcomes, goal achievement, application of time management rules and experiences regarding that matter.

In the weekly coaching sessions the coach and the coachee reviewed the individual progress, adjusted the time management rules if not appropriate anymore or at some point established that the particular goal has been achieved, the new behaviour has been adopted and no further practice regarding that goal is needed.

Relation to MIRROR CSRL Model

This approach can be related to all four phases of the CSRL model. In the *1st phase: plan and do work* the coachees perform their normal work activities but try to follow time management rules they selected in the introductory session with the coach. The KnowSelf app records these computer activities automatically. The captured data enable consequent reconstruction of individual work time and review of their work and learning history (transition 'data'). The *2nd phase: initiate reflection* focuses mainly on setting the goal for reflection. In our case the coachees reflect at the end of each workday on their data and review the usefulness of the chosen time management rules they try to apply in their everyday work. They are free to set an objective for reflection and to decide how much time he/she will invest. The KnowSelf App helps to set new objectives in that it enables gaining insights regarding time management or tracking the time management goal achievement with the help of the app. The data recorded and visualized by the KnowSelf App should serve as a trigger for the next phase. In the *3rd phase: conduct reflection session* the apps support the coachee to reconstruct and review her work experiences and re-evaluate them. The coachee evaluates how helpful were the time management rules and can decide to change the rule, to reject it, to choose a new rule etc. All insights and reflection outcomes gained in this phase can be documented in the ARA App and

serve as a basis for planning the future changes *(transition 'outcome')*. In the *4th phase: apply outcome* the time management rules are assessed once again with the help of the ARA App and the coachee decides what should be changed in the usual workflow or time management habits, how to implement these changes and what is needed to succeed in it. Reflection Diary in the KnowSelf app can be used to set goals for behavioural changes which can be reviewed at a later point and all goals checked as soon as they have been accomplished. This phase can be conducted by the individual coachee or with the help of the coach. If a coachee realises that some other aspects of personal time management are suboptimal he/she could decide to invest energy to implement various changes which led to subsequent reflection cycles.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

The focus of the used apps was to capture data about individual work practice and to foster individual reflection. Considering the transition model, it can be said that the tracking tool provided relevant material which could trigger individual reflection process and the ARA App enabled documentation of reflection outcomes for present or future assessment (step 1). These reflection outcomes could lead to consecutive reflection processes (recursion into step 2) and to individual application of outcomes (step 3). It was possible for the individual coachee to decide to transfer insights and outcomes based on individual reflection to team or organisational level, but this transfer was not supported nor was it object of this evaluation.

5.2.3 Research approach

Design and procedure

Before the evaluation started, the participants filled in an online-questionnaire and each participant was invited to an introductory session which had the following goals: for coachees to get to know the coach and the applications that will be used (KnowSelf, ARA, ManicTime), to learn about the procedure and time plan, to talk about their current situation regarding their time management. The baseline measure was introduced: one week long they should behave as they normally do regarding their time management and in parallel run the tracking app to record their work activities and at the end of the day assess the goal achievement in the ARA App. Based on the result of the baseline measure the coach and the coachee determined the time management rule(s) the coachee should practice and try to internalize in her work behaviour. The apps were used as a support in the whole process. The tracking tool helped to understand where the time was invested, what prevented a goal to be achieved etc. The ARA App enabled documenting learning and reflection outcomes, goal achievement, experiences regarding application of time management rules. The coach and the coachee met every week to talk about the progress, adjust the rules if not appropriate or to establish that the goal had been achieved, the new behaviour had been adopted and no further practice was needed. In the sixth week coachees conducted the final measuring (comparable to the baseline measure) without being coached anymore and assessing daily to what extent they achieved their time management goals. They evaluated the coaching experience using a "Check the coach" questionnaire and discussed it in more detail in the final discussion with the coach. Additionally, they filled in the post-questionnaire and participated in the follow-up interview conducted by the research partner (KNOW).

Participants

The group consisted of two female and eight male participants. The median age was 30.39 years, with a range from 20-29 to 40-49. At the time of the evaluation nine of them were employed full-time and only one person worked on a part-time basis. They worked in different departments (e.g. Technical Services, Sales, Innovation Labs etc.) and half of them worked on the management level (e.g. project leader, finance director, business development manager etc.). On average the participants have been working in their current positions for almost three years with total work experience in this field of more than four years.

Summative evaluation methods used

Out of the summative evaluation toolbox (see chapter 3) core-questions from all four levels were chosen (reaction, learning, behaviour, results); more specifically these sections were covered: demographic items, level of participation, short reflection scale (CR), app-specific questions (CA), learning outcomes (CL), work-self assessment (CB), and loyalty metric (LM). Some questions beyond the Core Questions were as well used: Usage (USE 01, 03, 06, 07), Usefulness/Satisfaction (SAT 01, 03), Inclination Long-Term Usage (LT 01, 02), Knowledge/Skills (KS 01, 02, 04), Work (WK 01, 05, 06, 08, 09, 13, 14), and work satisfaction. Additionally, in the post-questionnaire we asked some questions regarding the importance of particular components of the whole approach and we addressed the KnowSelf prompts with a few questions.

The coaching part of this approach was evaluated using the questionnaire “Check-the-Coach” by Bachmann, Jansen and Mäthner (2004)[3] which covered these areas: structure, process, result, assessment and overall rating.

With respect to the organisational impact of time management tool, the following KPIs were measured: subjective assessment of personal time management (measured with the time management scale by Hansen (2001)[4]) as well as the job satisfaction in a pre/post comparison.

5.2.4 Results

In the following the results are presented according to the four Kirkpatrick levels. The ARA App was used by the all participants while the tracking tools were used by 5 participants each (KnowSelf and ManicTime App).

5.2.4.1 Level 1: Reaction (Usage)

According to the self-reports in the post-questionnaire the users at IMC used the tracking tool on average 3.4 days per week and the ARA App on 3.8 days (over the course of 6 weeks). When reporting the average weekly usage they stated they used both the tracking tool and the ARA App about 15 minutes each on a weekly basis. We did not receive any log data from the participants (most probably due to privacy concerns) thus it was not possible to evaluate the objective app usage and compare it to the self-report measures.

We asked additional questions regarding possible barriers to the tracking tool usage. These questions as well as the feedback received from interviews showed that the participants experienced either lack of time for reflection during their busy working hours (USE1: I did not have the time to use the app; M=3.50 SD=1.27) or lack of motivation for extra effort after a long work day. It was a positive observation that they felt comfortable with the fact the tracking

[3] Bachmann, T., Jansen, A. and Mäthner, E.: Check-the-Coach: Fragebogen zur Evaluation von Coaching, Goldenstedt, Christopher Rauen, 2004.
[4] Hansen, K.: Zeit- und Selbstmanagement, 1. Aufl., Berlin, Cornelsen, 2001

tool records their “every move” on the computer (I was uncomfortable having my data recorded in such detail; M=1.50 SD=0.71) but one half of them stated they don´t see an advantage in using it (USE03: I did see no advantage in using the app; M=3.50 SD=1.35). Besides a few technical problems with the KnowSelf App the participants stated they missed in both tracking tools a sort of recommendation how to proceed after seeing the captured data in the app. They wished for automatic recognition of some behavioural patterns as well as more support regarding data interpretation and possible future actions. In the interviews some reasoned that the substantial effort of analysing such detailed data would outweigh the resulting benefit. Also some participants claimed that the guidelines they received in the beginning were not clear enough regarding what exactly they should look for when investigating their data.

Satisfaction was measured with two items (SAT01 and SAT03) using the 5-point Likert scale from strongly disagree (1) to strongly agree (5). The participants were moderately satisfied with the apps (tracking tool KnowSelf: M=3.00 SD=0.71; ManicTime: M=3.60 SD=1.14; ARA M=3.35 SD=0.91). (SAT01: M=3.00 SD=0.82).

The data collected by the tracking tools was fascinating for the participants in the first days and weeks, such as the possibility to see each day exactly where time was invested. But after a while some of the participants experienced difficulties how to proceed from there: as already mentioned, the time investment would be too large to examine such data in more detail and after receiving an overview of their activities they missed more complex analyses or recognition of behavioural patterns which could give them input on a higher level.

Concerning the ARA App it was reported that the app was mainly used for documentation of goals and experiences without being able to draw some conclusions from that material. The participants expressed they missed an opportunity to reflect collaboratively and exchange experiences which can be supported with the ARA App but was not offered in this scenario.

When asked about their inclination of long-term usage about two thirds of the study participants stated they would in the future again take part in a time management coaching such as this which was also confirmed by the Check-the-Coach questionnaire regarding this question. Further, 40% participants see the long-term advantage of this approach in their work life.

Table 5.2.1. Inclination Long-Term Usage

	Mean	**SD**
LT01: I see the long-term advantage this time management coaching approach has in the work-life.	3.50	0.71
LT02: I would like to participate in the future in a time management coaching again.	3.70	0.95

In the interviews we found out that one participant plans to continue using the ARA and KnowSelf App. Four participants stated that a collaborative approach (group coaching, group discussions, collaborative use of ARA) would be helpful and interesting to test in the future. And the most positively, all participants were determined to keep up reflection and applied changes in the future or at least try to do so.

5.2.4.2 Level 2: Learning

Asked about the effect of the time management coaching approach on their reflection 80% of the participants stated in the interviews that the regular constraint to occupy themselves with their time and task management on a regular basis motivated them to reflect and become more aware of those issues. The lesson learned for most of them was that more intense and realistic planning enhances their ability to manage their time and to meet deadlines. Also they gained insights in how constant interruptions of any kind (mails, switching to unexpectedly incoming tasks) cost them time and that they should try to reduce or avoid them if they are not of the highest priority.

Learning Process

The app-specific reflection questions regarding the KnowSelf App showed that the app helped the participants the most by providing accurate information about their work (M_{CA12}) and by collecting information relevant to reconstructing experiences from work (M_{CA1}). The control item where no change was expected (the app providing information about related experiences) was rated the lowest and hence confirmed the true effect of other statements (M_{CA16}). Average of all five Core Question App-Specific Reflection Questions (CA statements) showed that the KnowSelf App provided better reflection support when compared to the commercial tracking tool ManicTime App.

The ARA App supported well reflecting on experiences from work (M_{CA2}) and capturing reflection outcomes (M_{CA7}). Again, the control item confirmed the true effect: help with simulating the work process was not supported by the app (M_{CA42}).

Table 5.2.2. Descriptive statistic for all CA questions asked

Questions	KnowSelf (n=5) M; SD; Md	ManicTime (n=5) M;SD; Md	ARA (n=10) M; SD
CA1	3.60;0.89; 3.00	3.20;1.10; 4.00	
CA2			3.70; 0.82
CA6			3.00; 1.16
CA7	3.00; 1.23; 3.00	2.00; 1.23; 2.00	3.50; 1.18
CA8			3.00; 1.56
CA10	3.20; 1.30; 3.00	2.20; 0.84; 2.00	
CA12	3.60; 1.14; 4.00	3.60; 0.89; 3.00	
CA16 (control item)	2.20; 1.30; 2.00	1.00; 0.00; 1.00	
CA40	3.00; 1.00; 3.00	2.40; 1.52; 2.00	
CA42 (control item)			1.50; 0.85
CA_all (without the control item)	3.28; 0.81	2.68; 0.67	3.30; 0.83

The short reflection scale using a 5-point Likert scale was applied to measure the participants' tendency to reflect before and after the evaluation period. The ANOVA test for repeated

measures showed that the main effect of type of reflection (individual vs. team) was statistically significant (F(1,9)=22.12 p=0.001); specifically, it showed that the participants' tendency towards individual reflection was significantly higher than towards team reflection in our data sample ($M_{ind.}$=3.64 M_{team}=2.44). However, the second factor (time: pre vs. post) was not statistically significant (F(1,9)=0.77 p=0.404) as well as the interaction between the two factors (F(1,9)=1.17 p=0.307).

This finding was confirmed as well by the results of the Check-the-Coach questionnaire where the participants stated they gained much more individual understanding of their work situation through the coaching than they learned about their colleagues.

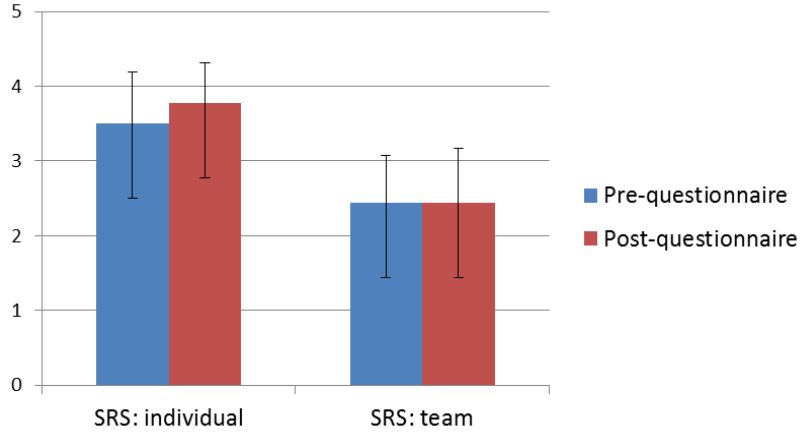

Figure 5.2.1. Mean reflection scores (individual and team reflection) in pre-post comparison

When comparing the reflection tendency before and after the evaluation period we did not find a statistically significant change over time regarding both individual and team reflection.

However, 7 out of 10 participants stated in the interviews that they reflect more after this coaching and app usage experience and that their awareness of how they spend and plan their time at work is now higher. The extent of this new gained reflection level varies from simply making themselves more aware of certain situations over to reconsidering interruptions and time wasters to a more conscious planning.

Regarding the app support of the reflection process half of the participants confirmed in the interviews that the detailed data provided by the tracking tools (both automatic recordings and manual task tracking) as well as the daily entries into ARA supported their reflection. For 80% of the participants the coaching provided equally important triggers for reflection and for gaining new insights.

The reflection process was supported additionally by one reflection guidance component of the KnowSelf App: the KnowSelf Prompts. The current version of the KnowSelf App included an automatic notification system which reminded the participants to reflect at a particular time of a day or after some unusual event was recognised by the app (e.g. high number of switches or long periods of idle time). We asked the participants whether these automatic notifications were helpful for their reflection process. The results show that the users did not profit from the prompts as we have expected it; they perceived the notifications often as disruptive because they popped up at times when they were working intensively on a task or simply did not want another window they had to read or think about. Most positively were evaluated the general reminder to reflect about the data in the app and the reminder to record projects.

Table 5.2.3. Assessment of Know-Self Prompts (n=5)

	Mean	SD	Median
Please indicate your agreement with the following statement. (scale: 1: strongly disagree to 5: strongly agree)			
The KnowSelf Prompts motivated me to reflect.	1.60	0.89	1.00
To what extent were the following categories of KnowSelf Prompts helpful to you? (scale: 1: not helpful at all to 5: very helpful)			
Reminder for using KnowSelf generally (e.g. look at heatmap, write diary entries)	2.80	1.30	3.00
Reminder of project recording	2.80	0.84	3.00
Notification about specific amount of switches	2.25	1.50	2.00
Notification about unusual amount of idle time	2.25	1.50	2.00
Notification about most used resources	2.75	0.96	2.50

We asked in the post-questionnaire for a few suggestions how this automatic prompting could be changed in order to be more useful to them, and collected the following ideas e.g. a person could define some web-sites they characterise as “non-productive work” (e.g. Facebook, Gmail or something else) and the app would remind them if they were longer than e.g. 10 minutes on that site; the app could connect the activities and the set goals and provide reminders about that; the app could automatically recognise when it is convenient to release prompts (e.g. not at the time when a person is working intensively and has a high frequency of switches).

Since the KnowSelf prompts were experienced as rather disruptive it was interesting to hear that one ManicTime user missed notifications reminding him of using the app for reflection.

Learning Outcomes

After taking part in the time management coaching nine out of ten persons agreed or strongly agreed that they made a conscious decision about how to behave in the future regarding their time management (CL01: M=4.30 SD=0.68).

The participation in the time management coaching and usage of the tools motivated the participants on the one side to think about their current behavior and habits and on the other side to decide to change something about it. According to the Check-the-Coach questionnaire 8 of 10 coachees stated they completely or at least partially achieved the time management goal they set themselves in the beginning of the coaching. These are some of goals they achieved: more realistic and structured time and task planning, reduced effect of time wasters, adjusted working hours, better dealing with interruptions, and in general higher awareness regarding issues in this area.

One third of the participants claimed they also gained a deeper understanding of their work life (CL02: M=2.80 SD=1.03) while around two third of them improved their understanding (KS02: M=3.70 SD=0.68) and work-related skills in the area they wanted to improve (KS04: M=3.40 SD=1.08).

In the interviews three participants decidedly claimed that their time and task management was unstructured before (if existing at all) and many of them could identify the potential for improvement in their way of working. They confirmed that a more structured and realistic time and task management is helpful with reducing stress and meeting deadlines. However, two participants pointed out that they also learned that some aspects of work cannot be changed if they are part of the job or defined by others.

Each coachee had in the ARA App a list of time management techniques she worked on that was updated in each coaching session. At the end of the coaching the following techniques had been applied and learned successfully (self-reports by the participants):

1. Keep a prioritized list of tasks and work by this list (8 tried, 8 succeeded)
2. Process emails to keep inbox empty (5 tried, 4 succeeded)
3. Only check emails a few times a day (3 tried, 2 succeeded)
4. Set aside a pre-planned amount of time for planned activities (3 tried, 1 succeeded)
5. Read a fixed number of pages a day to stay informed (1 tried, 1 succeeded)
6. Split large tasks into smaller tasks (1 tried, 1 succeeded)

There were five more techniques that were tried by one coachee each, but they could not apply and learn them.

We analysed as well the notes the coach was taking during the weekly coaching sessions regarding coachees' experiences with the apps, insights and learning outcomes, etc. There were 40 statements collected.

Most of these statements could be described as reflections on the tool usage (e.g. "*Although I cannot motivate myself to reflect every day I use the app on a daily basis to publish my experiences*"; "*It is hard to reconstruct the whole work day with the help of the app*"; "*It could be useful to use the tracking tool also for personal goals but it should be mobile, not on the computer*"; "*It is interesting to see in the tool when the shifts from one task to another happen and to realize how long were you really working on one task*").

The reactions we analysed showed that on the one side they seemed to experience some difficulties with making the most of the apps due to various reasons but on the other side they could derive some insights and conclusions from their interaction with the apps.

In the following we listed some of the insights or new understandings they reported during the coaching sessions: improved understanding of the personal way of working and why something is not functioning in the desired way; better and more realistic time planning; improved awareness and understanding of the project and tasks status, etc.

5.2.4.3 Level 3: Behaviour

Not only did the participants gain insights regarding their time management (level 2), eight of ten participants also improved their time management with the help of our approach (CB01: M=4.10 SD=0.74). It was confirmed also by the Check-the-Coach questionnaire that 8 of 10 persons learned some new behavioural patterns.

Additionally, the participants reported they used their learning regarding time management on the job (WK01: M=4.10 SD=0.57), focused more on their work tasks with the help of our approach (WK09: M=3.90 SD=0.57) and kept up their change of behaviour (WK05: M=4.20 SD=0.79).

As an individual KPI measure we used a self-assessment of their personal time management before and after our intervention. We found a statistically significant change in their assessment: after using the apps and participation in the coaching they evaluated their time management skills more positively ($t(9)=-3,4$; $p=0.007$). This result goes along with the previous observations that during the coaching period they gained new insights regarding their time management, optimised it in one or another aspect, applied what they learned on the job and as a result they evaluated their time management skills better than before the intervention.

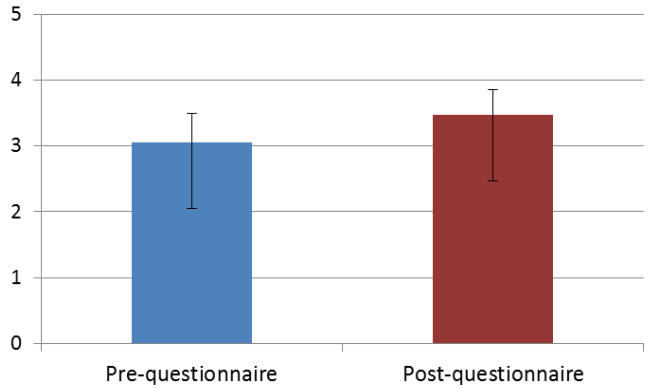

Figure 5.2.2. Mean over all questions regarding time management (pre-post)

In the interviews seven participants reported that they changed or at least intensified their time and task planning (e.g. prioritizing their tasks, estimating time effort for tasks more realistically and including time buffers for unexpected happenings). Some of them also claimed to apply the method of an empty inbox they learned in the coaching, meaning that emails are instantly transformed into tasks or filed away in a structured way. Two third of the participants also stated that they handle time wasters (interruptions of all kinds, unnecessary task switching, some meetings) differently and more consciously now.

We also asked them whether they feel that these positive changes affect their work quality and work satisfaction. While one third of the participants stated although they feel they work more efficiently this does not affect their work quality, another third was convinced that a more structured and conscious way of working does indeed lead to a better quality of work because one is less prone to errors.

Concerning the influence of the applied changes on their satisfaction at work, six out of ten participants admitted that they are more satisfied when they have less stress and fulfil their planned tasks due to a better planning and time management. On the other hand one third of the participants stated that they were less satisfied because they realized either that despite applying time management improvements the workload was still too high or that fulfilling tasks faster and more efficiently causes new tasks to follow sooner. On average the work satisfaction (measured with the questionnaire item) remained stable during the evaluation period (pre: M= 3.20 SD= 0.79; post: M= 3.30 SD= 0.82; $t(9)=-0.246$, $p=.811$).

5.2.4.4 Level 4: Results

When asked how likely it is that they would recommend the whole approach or particular components to a friend or colleague we found that the highest chance of being recommended has the coaching part. This was also confirmed in interviews, where 80% of the participants named the coaching as the crucial part of the whole approach for triggering reflection, gaining insights and learning new techniques. They explained that the regularity of the sessions

motivated them to reflect and supported the learning effect, that discussions with another person are more helpful than reflecting on their own and that the coach introduced them to interesting new methods for improving their time and task management.

The whole approach would actively be recommended by 30% of participants (promoters: scores 9-10) and passively by another 30% (scores 7-8). The net promoter score amounts to –10%.

Regarding the tracking tool we found that the ManicTime has a better chance to be recommended when compared to the KnowSelf App: 40% of ManicTime users would recommend it actively (promoters) while most participants are neutral about recommending the KnowSelf App. This is surprising when considering that the app-specific reflection questions showed that the KnowSelf App was better evaluated when compared to the ManicTime App.

The ARA App would be actively recommended by one person out of 10 while three can be regarded as passives. The coaching part would actively recommend 40% of participants (active promoters) and another 30% (passives). The net promoter score amounts to +10%.

Table 5.2.4. Percentages of promoters, passives, detractors and NPS

	Approach	KnowSelf	ManicTime	ARA	Coaching
Promoters	30%	0%	40%	10%	40%
Passives	30%	0%	0%	30%	30%
Detractors	40%	100%	60%	60%	30%
Net Promoter Score	-10%	-100%	-20%	-50%	+10%

According to the Check-the-Coach questionnaire 8 of 10 participants agreed or strongly agreed they would recommend coaching to another person.

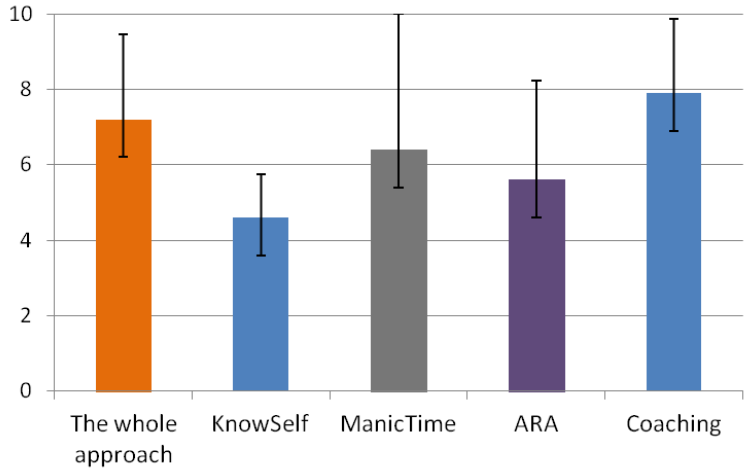

Figure 5.2.3. Mean scores of loyalty measure

When asked about the effect on organizational level most participants stated that an improved time and task management could influence the whole organization. Twenty percent of the participants explicitly stated that they try sharing their learning outcomes with colleagues. One

third of them said that the improvement of time management on the individual level would also improve the general productivity and collaboration, while four participants emphasised that processes and collaboration would have to be optimized on a more global level (e.g. meetings, responsibilities, communication) in order to induce an improvement on an organizational level.

5.2.5 Conclusion & Discussion

The approach applied in this test bed included usage of the apps combined with weekly coaching sessions. It was expected on the one side that the apps will provide relevant material for reflection and on the other side that the coaching will help the participants to work with these data, to follow his or her goals and improve the aspects of time management they wanted to change.

The evaluation confirmed that the participants gained valuable insights regarding their time management skills and way of working, made a conscious decision to change some aspects of it, applied the lessons learned on the job and as a result in general they improved their time management.

Although all components of this approach influenced this positive result, it is clear that in the eyes of the participants the coaching part was the most important. The ARA app was the 2nd and the tracking tool the 3rd responsible factor. The apps served well as a support for coaching sessions in the sense that the goals set together with the coach and progress regarding them were documented in the ARA App and the tracking tool provided a reality check to the participants how they spent their time and showed them whether they behaved according to the goal they set themselves. The participants stated that the apps were surely important for reflection and focusing on the goals because of the functions mentioned above but the coaching session gave a sort of a frame for all of it and a motivational boost to really devote oneself to it on a daily basis. The input from the coach in the introductory session was very important for definition of the expectations, goals and the whole coaching setting. Further, they were supported in their learning progress and trying out new time management techniques, they could discuss with the coach why they experienced difficulties with some goals and how the current work situation was affecting the goal achievement.

The apps alone could be used for giving the overview of the activities and tracking of tasks and projects but a person would have to be very self-disciplined and self-reflective already to use the apps regularly for the reflection and learning purposes without being coached and guided. This strengthens the thesis based on the evaluation results at Infoman that additional support is needed to make sense and use of the recorded data, because the combination of app usage with the coaching lead to an improvement of time management at IMC, while the usage of KnowSelf alone at Infoman did trigger some deeper understanding and insights in working situations, but didn't lead to a generally improved self-assessment of time and task management.

Another aspect that affects this kind of learning at work is the current job situation. On the one side: how high is the work load at the moment, does the nature of the current job allow to apply certain time management techniques, etc., and the organisational factors on the other side: to what extent is the employee expected to self-organize its work (or is she instead expected to follow pre-defined work or react to messages), is it possible to apply time management techniques that increase one's productivity, when the service quality of customers or work partners is at risk, should time management techniques be trained in all levels (management and worker) instead of only one level (management) of a team to yield real change, etc.

Quote from one coaching participant regarding the potential future plans: “I think it is important for the managers here to know of this possibility. I participated as head of department myself and therefore know of it and already talked to our human resources department what sort of implementation for other members of my team could be possible.”

One occurrence that was not expected was the fact that the whole approach, which aimed at improvement of time management, would exert additional stress on participants instead of reducing the stress level. Similar to the statements collected at Infoman, one third of the participants at IMC claimed that their participation was a temporal burden and another third stated that they didn’t even have the possibility to apply all suggested changes because of their high workload. Also four out of ten participants pointed out that some aspects of time and task management simply are not changeable, because they are defined by their job roles or current tasks (e.g. constant interruptions by newly incoming tasks, phone calls or emails as well as lack of necessity/possibility to structure their tasks).

An important insight gained was that the evaluation of the coaching and app experience should be separated from the coaching itself and should be conducted after the coaching has finished. This would allow the coachees and the coach to focus only on the coaching process and the goals they want to achieve without having to assess at the same time the applied methodology.

We also collected some ideas how the approach in general could be improved: the coaching should have a longer and more extensive introductory phase where the reflective learning approach, the methods and goal setting are discussed in more detail. Further, the usage of apps and the sessions’ length and frequency could be adjusted a bit more to suit the coachees’ individual needs. A possibility how everybody would profit more from the app usage could be the app usage in the group setting, which was suggested by both the coach and coachees. This could strengthen the effects of reflection by sharing or comparing insights with those of others, finding solutions for mutual issues etc.

Summarizing these results it can be said that all phases of the CSRL model were supported to some extent by the time management coaching approach: The combination of data analysis and coaching sessions provided the participants with input on how to improve their planning and conducting of their normal work activities (phase 1). Reflection might not have been initiated like planned at the end of each work day for each participant due to timely and motivational reasons, but it was initiated on a regular basis (phase 2) by looking at the data recorded by the tracking tool and documentation of goals and progress in ARA as well as by the discussions with the coach, which especially served as a motivation for the participants to occupy themselves with their time management. Also these sessions with the coach granted the conduction of reflection sessions (phase 3) and the re-evaluation of work experiences and changes already applied to work behaviour and time management. Changes were applied following some time management methods introduced by the coach in the beginning as well as following insights gained from the coaching and data analysis during the evaluation (phase 4). Although the CSRL model is sufficiently supported by the time management coaching approach, we do of course see the potential for further improvement. One possibility could be to not mainly focus on general time management issues and rules, but to target a more personalized approach dealing with more individual data interpretation, issues and goals.

5.3 The MoodMap App evaluation at BT

This and the following section describe two evaluations of the MoodMap App. To avoid redundancy all aspects which refer to both evaluations are described only once in this section 5.3.

The MoodMap App is a tool for mood tracking, which allows users to capture their moods as well as their related notes and context during their working shifts. Different visualisations on an individual as well as team level support the re-evaluation of their work experiences during a day or a week in order to reflect on them and gain new insights.

5.3.1 Organisational context

Test bed organisation and the organisational unit

The general description of the test bed as well as user and job role description can be found in section 3.1.

The MMA evaluation was set up for 4 teams. Each of the teams is led by a manager and has one or two coaches, who support together the advisors of the team. In each team there are between 10 and 20 advisors.

Identified need and potential for reflective learning

The use of the MoodMap App was initially identified by managers of the call centre, who were willing to be aware of how their advisors feel at work. The managers and coaches in the centre were also introducing a new model of coaching in order to allow employees to take more decisions regarding their learning process at work. Until that moment, the model they had followed had been quite passive from the employees' side. This change was a decisive aspect that pointed towards the necessity of supporting learning by reflection. With this, a new culture is being created in order to encourage employees to take their own decisions and take into account how they feel and what they need for themselves and in order to improve their work performance.

During their daily work, emotions and reflection play a very important role. This applies not only for the advisors but also for the coaches and managers. Advisors have to face many challenging situations during their calls with customers. They have little time between one call and the next one and they need time to think about how they feel or how they are performing. The relationship between coaches and coachees is very important in their daily work. Coaches and also managers have major interest to know how the advisors, they are responsible for, are feeling in order to provide appropriate support to them. In many cases, managers and coaches are not aware of the issues that the advisors are dealing with and how these affect their work and emotional well-being.

In these different settings, the MoodMap App should give the individual the possibility to reflect about her own mood as well as the mood of the whole team. With the help of some guidance offered by the app, the participants should be activated to re-evaluate his/her mood or mood changes with the goal to learn from it. This includes learning from critical reflective thinking or his/her own working behaviour during a certain meeting, a customer call or any other challenging working situation. Additionally it should give managers a new perspective to take into account and reflect about the team's mood, if there might be some problems or if everything runs smoothly within the team. According to these results and the insights gained after reflecting on it, possible actions could be taken.

The potential for reflective learning can be considered from two different perspectives. From the individual point of view with regard to all three roles, the MoodMap App should help to

- provide an easy possibility to re-experience their working day upon their moods and the corresponding context and notes
- make them aware of their own mood development during the day with regard to their conducted calls

With regard to the perspective of coaches and managers and their coaching with the advisors, the MoodMap App should contribute to:

- identify possible skill gaps on how to deal with difficult customers and difficult situations.
- find out, if someone feels badly because of troubles with customers or with solving some technical issues e.g. internet protocol (IP) issues.
- discover calls or patterns of calls which reduce the energy of the advisors.
- identify reoccurring periods of a working week with low feeling and energy levels.
- improve coaching sessions, by reflecting together with the advisors on the mood development and individually adapt the sessions to the advisors need.
- support coaching sessions with the advisors which are aimed at raising the energy level of the advisor again.
- make aware of the team mood and try to take some actions to improve it.

Potential organizational impact

The BT call centres are currently running a program which focuses on how they speak to their customers and the effects this can have. Furthermore they are currently restructuring their coaching cultures and wanted to use the trial of the MoodMap application as complementary approach to their ongoing efforts. With the trial we wanted to find out if the usage of the MoodMap App might have any impact on the coaching sessions and on the employee's self-reflection and satisfaction.

The overall goals from the organisational point of view was that this evaluation should

i. create awareness that the organisation is caring about their staff
ii. create awareness that for managers and coaches the individuals are important
iii. create self-awareness about the advisors own work performance

As a result the call centres management would like to see their advisors becoming more active as well as constructive during work. In the end, they would like to motivate their managers and coaches to be more proactive and more visible and participating with their advisors in the floors.

5.3.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

Within the MoodMap App the user (regardless of the user's role as advisor, coach or manager) has the possibility to capture her individual mood states, which she has during a working day or during a meeting. The mood is composed of two levels namely the valence level (feeling

good – feeling bad) and the arousal level (high energy – low energy). Additionally, the user is motivated to think about her current mood with the help of so-called reflection interventions. These interventions ask compulsory to select a context for the currently inserted mood e.g. after a call, after a coaching session, after a break or whenever they thought their mood has changed. Furthermore users have to insert a note to each mood, in order to be able to better reconstruct their working experiences during the day. The compulsory insertion of a context and a note should trigger the users to think about their mood and the relationship with their work during the usage of the application.

Moods, notes and contexts are presented in an individual timeline visualization, which enables the user to recall past working experiences more easily. On a collaborative level, two visualisations provide the possibility to see the average team mood and compare it to the own mood. When a working day is over, two different reports are available which cover a summary of the day as well as the average mood development of the team during that day. Additionally, there exist special team views, which are only accessible by coaches and managers. These views give them more insights about the moods of each single advisor of his/her team and serve as triggers to reflect upon the mood development of their team members as well as the mood development of the whole team. This reflection should lead to outcomes that make them more proactive with regard to their advisors. A detailed description of the MoodMap App including all its features and different visualisations can be found in D9.5.

Within this evaluation the MoodMap App v3 (version 3) was used. This version of the MoodMap App includes additional features especially implemented for BT in order to fulfil the needs of their working setting. The MoodMap App v3 is described in detail in Chapter 7 of D9.5. Whenever the term MoodMap App is used in this document, we refer to MoodMap App v3. Only when referring to another version of the MoodMap App we explicitly mention the version number.

In order to smoothly integrate the MoodMap App into the daily working routine of an advisor, the application was integrated into BT's call taking system. So whenever they have finished a call with a customer or when they are doing any task, they could just click on a button in their system to go to the MoodMap App and capture their mood.

Relation to MIRROR CSRL Model

Plan and do work

The MoodMap App supports capturing one's own moods during work. The captured moods, positive as well as negative ones, can be used for reconstructing and reflecting on experiences from work or motivate users to reflect. The app follows a user-initiated capturing approach and learners can state their mood through a coloured bi-dimensional map which represents emotional states. Additionally, to each inserted mood, a compulsory context and a personal note have to be attached. The MoodMap App also supports the sharing of experiences, as moods, notes and data are being shared among participants of the same team (transition "data").

Initiate reflection

In order to initiate reflection, the application follows two approaches.

First, the MoodMap App has a reflection guidance mechanism, which motivates users to reflect about the individual mood or the team mood during the usage of the app. To each inserted mood, a compulsory context and note have to be inserted. This leads to a better

contextualisation of the inserted mood and can later be used to reconstruct past working experiences easily. On the other hand while selecting a context and inserting a note, they were more or less forced to reflect or at least to think about the currently inserted mood.

Secondly, reflection can also be initialized by exploring and analysing the different visualizations and data reports. By analysing this data and its context, the user can set the objective for reflection. As moods are shared, the aggregated data is available for exploration and thereby it facilitates reflection on it. The app does not plan or organize the reflection session (i.e. does not setup the "frame"), but mainly makes available the data which can support it. The MoodMap App provides reports on past working days and for coaches and managers additional team visualisations to follow the team's mood development during the day or a whole week. Additionally, debriefing sessions can be organized where several members of the team and/or the manager or coach discuss the results collaboratively.

Conduct a reflection session

When reflecting during the usage of the MoodMap App, the users have to select a suitable context to indicate in which situation they are and they can attach their comments directly to the mood in form of a note. Several visualizations are offered to support the reconstructing of work experiences, e.g. a timeline presenting the individual mood development, the comparison of individual and collaborative moods, and a collaborative map showing all individual anonymised moods. After the shift, the reports also show a summary of the captured moods as well as the evolution of the moods in a timeline. Moods are contextualized through the user's comments and personal notes. These are also available in the visualizations in order to support the reflection process. The captured data is shared by all participants of a meeting (or all members of a team during a shift). The app supports the re-evaluation passively, relying in a data explorative approach. Users can keep track of their insights and reflection outcomes in the Reflection Journal, which is available in each shift report. This preservation of outcomes corresponds to the transition "outcome".

Apply reflection outcome

In the Reflection Journal the user can find the reflection outcomes that he/she has entered in each meeting report. Based on the insights gained, the users may apply them during their following working days (transition "change"). When reflecting again on the individual mood or the mood development the users can check whether they were able to improve their work and emotional state.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

The MoodMap App supports reflective learning on an individual as well as on a collaborative level. The user can insert one's own mood, context and corresponding note during work. The user can use the different visualisations available within the MoodMap App to individually reflect about her mood development during the day or comparing her mood to the average team mood. On the other hand the data is shared anonymously within teams and their meetings or shifts. Only users with a manager or coach role have also access to the individual moods and other data of their team members. This allows them to better assess how their employees are feeling as well as to discuss their moods with them and reflect on the data collaboratively. Outcomes from this reflection may have an impact on their work performance and therefore on the organizational goals and results.

Regarding the transition model (see Figure 10.3.1) it can be stated that the MoodMap App has different types of triggers to initiate reflection and keep up the reflection process (Steps 1 and 2). First, by compulsory asking for a context and a note to each mood reflection can be triggered (Step 1). Selecting the corresponding context and adding a note to the mood could lead to a recursive reflection process (Step 2). Second, by reviewing the shift reports on an individual or team level at the end of the day reflection can be initiated both at individual and collaborative level (Step 1). Thereby, depending on the situation of the user the reports can be explored by oneself, or collaboratively in a meeting or coaching session. Re-experiencing the mood development during one working day by using the context and notes could lead to a recursive reflection process (Step 2). The gained insights or outcomes can be stored within the reflection journal. Third, for managers and coaches, revisiting the team views might also lead to initiate a reflection process (Step 1). A recursive reflection process might be triggered, if the managers and coaches came up with unusual or significant mood developments on an individual or team level (Step 2). Applying these insights or outcomes during work (Step 3a, Step 3b) can follow from the reflection processes, but is not explicitly supported by the application.

Since the MoodMap App was designed to support individual reflection and in this setting also collaborative reflection to a certain extent (managers and coaches together with their advisors), push- or pull mechanisms to initiate communication and collaborative reflection are not directly implemented in the application, but can be instantiated when using the application and reflecting on the captured data. For example, it can be seen as a pull-mechanism when a manager initiates a team meeting because he/she sees in the MoodMap App that the average team mood is very negative. On the contrary, a pull-mechanism occurs when the team members themselves share their data and reflect together and pro-actively contact their manager, in order to present him/her their insights after reflection. The outcomes resulting from individual reflection can be stored in the reflection journal, but not shared with others. The outcomes resulting from the collaborative reflective processes are shared and communicated through the already established work processes of the participants (e.g. in a coaching session or in team meetings).

5.3.3 Research approach

Design and procedure

Before we started the summative evaluation, we had a pre-evaluation (see D4.4) study of the MoodMap App v2 (version 2) in order to capture and extract requirements for preparing a successful summative evaluation.

The MoodMap App was used by BT employees having all three roles i.e. advisors, managers and coaches. The advisors were asked to fill in their individual moods after a call, after a coaching session, after a break or whenever they thought their mood was relevant for them or has changed in a significant way. Additionally they were compulsory asked to insert a note to each of the captured moods. The coaches used the inserted moods to better support the advisors during their work and to improve and/or adapt the coaching sessions to the advisors' needs. The managers also followed the mood development of each single advisor to directly contact them if necessary. On the other hand the managers also kept an eye on the mood development of the whole team in order to arrange *one-to-one meetings* (i.e. a meeting between a manager and a single advisor) *or huddles* (whole team meetings) to discuss about any issue and improve the whole team spirit. Additionally managers and coaches were also

asked to capture their moods and use the MoodMap App for themselves with the goal to reflect on their own mood development and how they were influenced by their emotions.

The summative evaluation of the MoodMap App started on the 20th of November and lasted until the 20th of December. The MoodMap App was introduced by the responsible project manager of BT to six different teams (encompassing 103 participants) and four of these teams were selected to pursue with the evaluation. During this introduction phase the team members were asked to fill in a pre-questionnaire including the MIRROR consent form. Regarding the application usage, all types of participants (advisors, coaches and managers) of the trial were asked to insert their moods during all days during the evaluation period. They were also asked to reflect about their inserted moods and notes individually. The coaches and managers were additionally instructed to use the team visualisations in order to reflect about the mood development of their teams and take actions if necessary. At the end of the trial, they were requested to fill in a post-questionnaire.

Participants

Altogether 67 employees participated in this evaluation, belonging to four different teams of two call centres namely Team AMc (Dundee), Team AMo (Dundee), Team GMa (Alness), Team STh (Alness). The demographic questionnaires were filled by 43 participants, thus the information of the demographic data is referring only to this 64% of the participants. These 43 participants consisted of 26 men and 17 women. Participants in the evaluation were between 20 and 59 years old, with 56% of them aged between 20 and 29, 26% between 30 and 39, 14% between 40 and 49 and 5% between 50 and 59. The participants had on average *M* = 3.47 (*SD* = 3.66) years in their current position, ranging from participants with less than one year in that position and to a maximum of 15 years. The average number of years in their current team was *M* = 1.76 (*SD* = 2.28) and the average years in a similar position were *M* = 5.21 (*SD* = 4.32).

Due to the fact that not all participants have filled in both questionnaires, Table 5.3.1 gives a summary of how many answers we received from each team. In the sections where it is referred to the evaluation of both questionnaires, only the answers of those users who have filled in the pre- as well as the post-questionnaire will be considered (*N* = 26).

Table 5.3.1. Summary of the filled in questionnaires

	Pre-questionnaire	Post-questionnaire	Both questionnaires
AMc (*N* = 12)	*12*	*8*	*7*
AMo (*N* = 13)	*13*	*13*	*12*
GMa (*N* = 18)	*5*	*13*	*4*
STh (*N* = 15)	*13*	*4*	*3*
Total	*43*	*38*	*26*

Summative evaluation methods used

Besides capturing log-data, two questionnaires were used: A pre-questionnaire, which gathered demographic information (as outlined in the toolbox) and contained the short reflection scale (SRS) and questions about participants' expectations with regard to the use of

the MoodMap App. The post-questionnaire contained all questions concerning the evaluation levels 2 through 4.

Level 1: Reaction: all usage data, including for example the number of clicks per visualisation, can be obtained from the logs of the MoodMap App. The database of the application itself stored the information about the data captured, i.e. mood values, notes and context data. In the post-questionnaire, questions regarding usage (USE) and user satisfaction (SAT) were added.

Level 2: Learning: The short reflection scale (CR) was included in the pre- and post-questionnaires. Additionally, the post-questionnaire contained 8 app specific reflection questions (CA) which fitted best to our approach as well as the two Learning Outcome questions (CL).

Level 3: Behaviour: We used the core behaviour question (CB) with regard to work performance. Additional questions regarding behavioural intentions (BI) were used, depending on the role of the participant, including questions about work improvement in general, employee satisfaction, and improvement of work performance or customer satisfaction (WP).

Level 4: Data on Key Performance Indicators (KPI) was provided by BT at individual and team level. At individual level volume and average rating were available. The KPIs at team level were Net Promoter Indicator, Advisor Satisfaction and Recap.

The post questionnaire consisted of further questions with regard general application effects (GAE), long-term usage (LT), benefits and insights as well as the impact of reflection in their daily work.

After the trial, two managers and one advisor took part in an interview over the phone. The posed questions covered the following topics: feedback to the overall experience, subjective feeling of the application's acceptance within the team, capturing mood, benefits and insights as well as general comments.

5.3.4 Results

5.3.4.1 Level 1: Reaction (Usage)

The participants used the MoodMap App on 31 consecutive days. Except for 5 days in which the app was not used at all (corresponding to days off), the app was used by all users for 8 hours and 42 minutes (*SD* = 0.09) on average every day. In that time users where entering the MoodMap App repeatedly and using the different available features.

In total, 991 moods were captured during the whole evaluation period. On average, users captured 17.39 moods (*SD* = 24.50) during the whole evaluation period, with a range of 1 to 136 moods per user. Considering the moods captured in each team, on average 23.72 (*SD* = 34.45) moods were captured per person in GMa Team, 21.00 (*SD* = 24.52) in STh Team, 10.62 (*SD* = 11.91) in AMc Team and 9.25 (*SD* = 10.95) in AMo Team. These results show that users in teams GMa and STh captured on average twice as many moods than teams AMc and AMo.

The analysis of the moods captured in each team revealed that all teams followed a very similar distribution. Each team had a percentage of its members (approximately 25%) who actively captured a higher number of moods (from 20 to 136), as well as another percentage of people (also approximately 25%) who captured fewer moods (from 1 to 5) but were rather checking the visualizations. This fact is also mirrored by the high standard deviations shown above.

Regarding the notes attached to the moods, a total of 946 non-empty notes were captured by the users and served as annotation for their affective states.

Users also had to enter a context to each captured mood i.e. in which situation the mood was captured. Figure 5.3.1 shows the distribution of all captured contexts (*N* = 991) among the four available categories i.e. after a call, after a break, after a coaching session and other. In order to get more insights about this 62% of "other" situations, the notes of those moods were analysed. This analysis revealed the following context as the most common ones: start or end of shift, before break/lunch, back from lunch or a certain event, problem or issue (crash, waiting for other departments), feeling better after dealing with a problem, successful events, feeling tired or having finished a certain task.

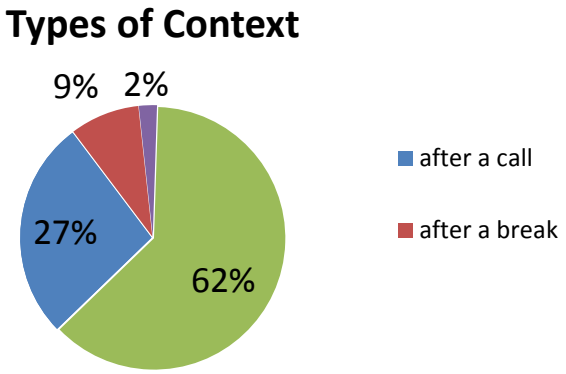

Figure 5.3.1. Distribution of context among all users

Figure 5.3.2 shows the distribution (absolute numbers) of captured moods and context as well as notes for each team that participated in the evaluation.

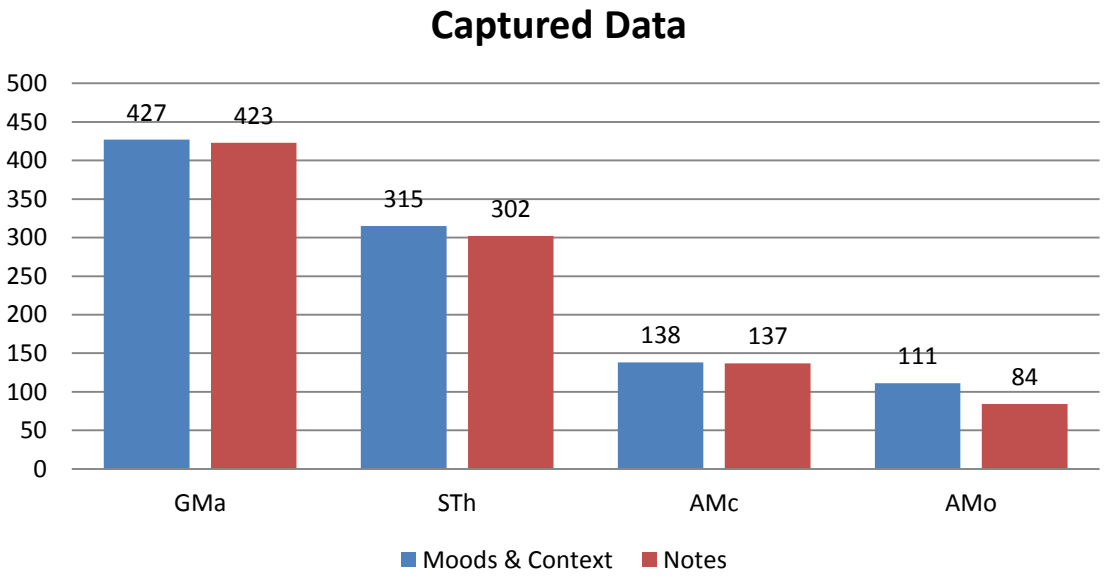

Figure 5.3.2. Absolute numbers of moods and context and notes captured by each team.

From all the participants that were registered in the MoodMap App, 53 used the available visualizations during their working shift and a total of 1914 interactions were logged (see Figure 5.3.3). Each of these users had on average 36.11 interactions with the application (*SD* = 60.63)

during the whole evaluation period. Such a high standard deviation also shows that the usage of the app was very polarized, having users who were very active, whereas others almost did not use the app at all.

As shown in Figure 5.3.3, the three features that were most used were the Capture Mood, Timeline and Compare Me visualizations. This was also confirmed by the interviews and the questionnaires. From the available visualizations, users were mostly using the timeline of the day, were they could review their mood development during the present day. The preferred visualization for the users was the Compare Me visualization, where employees could compare their own mood with the mood of their colleagues. This confirms our hypothesis that comparing themselves to the team in a quick and intuitive way may be supportive and useful, both in terms of creating curiosity in the users as well as allowing users to detect discrepancies that can initiate a reflective process.

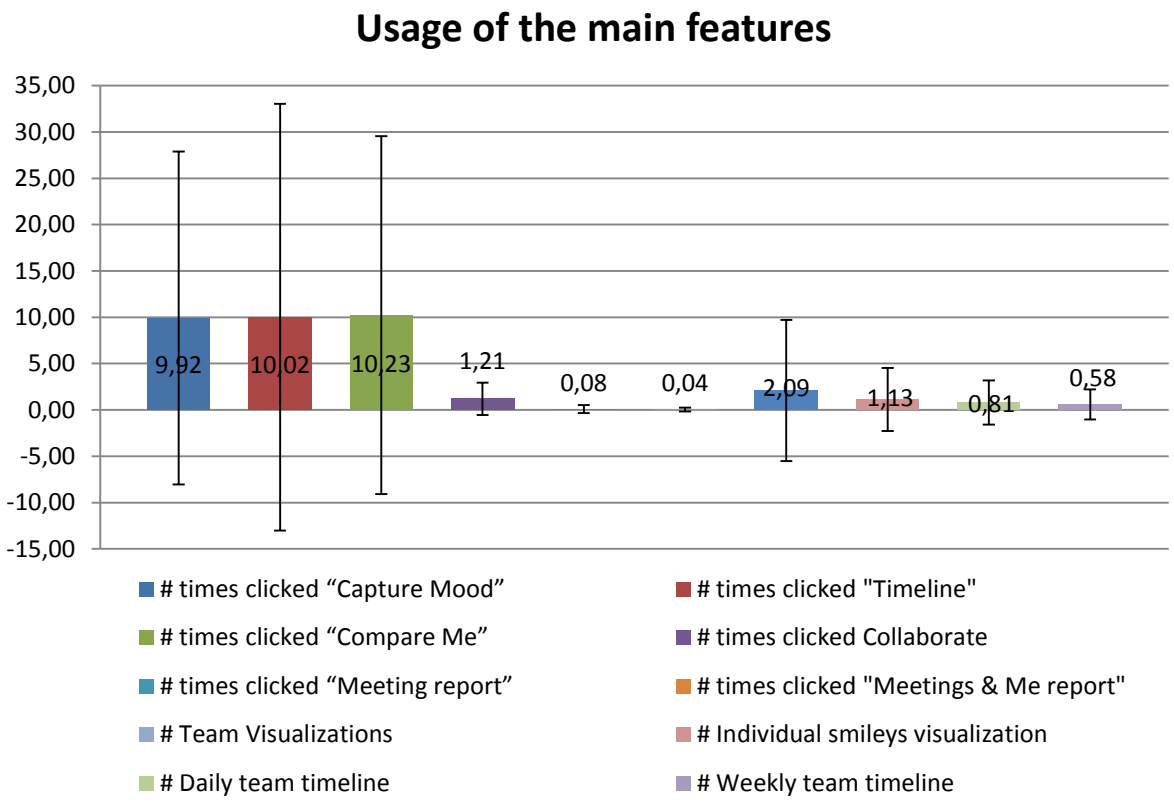

Figure 5.3.3. Usage of the main features of the MoodMap App. For each feature, average and standard deviation are depicted

Overall impression

The answers (*N* = 38) to the open questions of the post-questionnaire that referred to the usage of the MoodMap App and participants' reaction to it showed that the users' opinions were very ambiguous – from very positive over neutral to negative. Regarding the situations in which it was useful to capture a mood, the responses range from "*It was interesting to capture moods at the start and end of shifts, dependent on how much work was left to do etc."* to "*difficult to use an app when calculating mood because I FEEL. I do not need a technology to validate it".* If the users would or would not share their moods with managers was also very

diverse. Some would especially share their mood about an awful call or a bad situation others would not share such a mood at all.

The final comments regarding the MoodMap App of the participants in general showed up again their different attitudes towards the application. Four participants provided positive feedback like "*The mood map was fun to use and did lift team spirit for those that used it but I wouldn't say it has impacted on my day to day work / performance" or "I really enjoyed it. To make it more fun, we would use a # in the same way as twitter. Helped improve mood further."* Five of the participants provided neutral statements and four of the participants' opinions were rather negative e.g. "*I think BT should spend their money more wisely.*"

The following items were all rated on a 5-point Likert Scale (strongly disagree – strongly agree). Figure 5.3.4 (blue bar) shows that the participants (*N* = 38) stated that they have slightly agreed to be satisfied with the MoodMap App. If we consider this score for each individual team, it shows that the teams, who have used the MoodMap App at most Team GMa, Team STh agreed to be more satisfied with the MoodMap App, than those who have not so intensively used the application Team AMc, Team AMo.

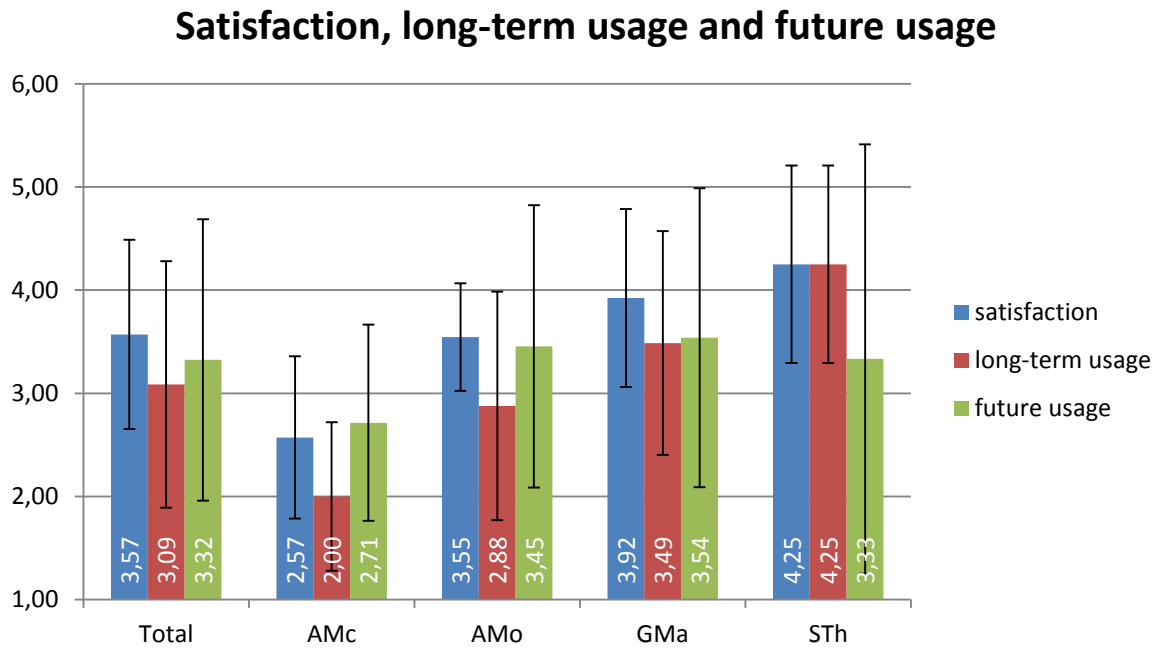

Figure 5.3.4. Mean scores of satisfaction, long-term usage and future usage in total and per team

The overall score regarding the long-term usage of the MoodMap App during work (Figure 5.3.4, red bar) is rated rather neutral. Again here the two teams Team GMa and Team STh, who have intensively used the application agreed that they would like to continue using the MoodMap App. In contrast, the scores of the two teams with the lower usage namely Team and Team AMo rated the long-term usage neutral or slightly disagreed to it.

The future usage was rated neutral or slightly positive. In this case the mean score represents also three of the individual teams except Team AMc.

A Pearson product-moment correlation coefficient was computed to assess the relationship between several variables from Level 1. Concretely, the number of captured moods, interactions with the application, subjective usage, self-expressiveness of feelings and social media attitude were investigated. Due to the difference between the data available for each

participant, it has to be taken into account that only a selection of the users could be considered in this analysis i.e. the users who had data available for each of the variables involved and mentioned below.

There was a very strong positive correlation between number of captured moods and number of interactions in the MoodMap App, $r = .870$, $p = .002$, for $N = 64$. That means that users who captured more moods were also using the visualizations more often. Also a significant moderate positive correlation between number of interactions and subjective usage ($r = .496$, $p < .001$, $N = 35$) was found. Finally, a strong positive correlation between captured moods and subjective usage was determined, too ($r = .626$, $p < .001$, $N = 35$). This result indicates that all three variables are correlated. This outcome is the expected result, as a higher usage of the app produces the capturing of more moods. It also shows that users could properly assess how often they used the app during their work and were aware of the introduction of the app in their daily practices. We also found that comfortableness of users with expressing their feelings at work is correlated with captured moods and app interactions. There is a moderate positive correlation between interactions and expression of feelings ($r = .411$, $p < .014$, $N = 35$) and a weak positive correlation between captured moods and expression of feelings ($r = .384$, $p < .023$, $N = 35$).

We found no correlation among the other Level 1 variables mentioned above e.g. social media attitude, which we expected to correlate with the usage of the app. This was also confirmed by an interview, where a highly active user in the MoodMap App stated not having any account in social media platforms.

Barriers

Figure 5.3.5 shows the scores of the possible barriers of using the MMA including general barriers such not having time, not having physical space, not having seen any advantage or not having motivation to use the app. Users rated these barriers with neutral to slightly disagreement, which also corresponds to each single team. The social media attitude (*Figure 5.3.5*, social attitude) i.e. how likely it is that they use social networking platforms (e.g. Facebook, Twitter, LinkedIn, Google+, MySpace…) was rated slightly positive, which means that the social media attitude is no barrier for using the MoodMap App. The social privacy concerns (Figure 5.3.5, social privacy concerns, rated with a 4-point Likert scale: not at all – high) shows that the participants are only little to somewhat concerned about their social privacy. Regarding their comfort with the self-expression of feelings (*Figure 5.3.5*, self-expression) in general and during work, the participants rated it neutral or slightly agreed. Having a look at this aspect at team level, only one team slightly disagreed with being comfortable in expressing their mood (Team AMc: $M = 2.50$, $SD = 0.80$), all other teams stated their self-expression rather positive. Sharing of emotions (Figure 5.3.5, sharing) with managers and coaches was not seen as a significant problem.

From the open questions of the post-questionnaire we received many different responses regarding the barriers of the MoodMap App usage. Responses from 13 participants, who have answered this question show, that they either do not see any direct barriers for using the MoodMap App, that they perceived the MoodMap App as easy-to-use, that the app could become part of their normal working day, or that the app has only a clear benefit for the coaches and managers but not the advisors, or that they do not think that the MoodMap App is useful for them.

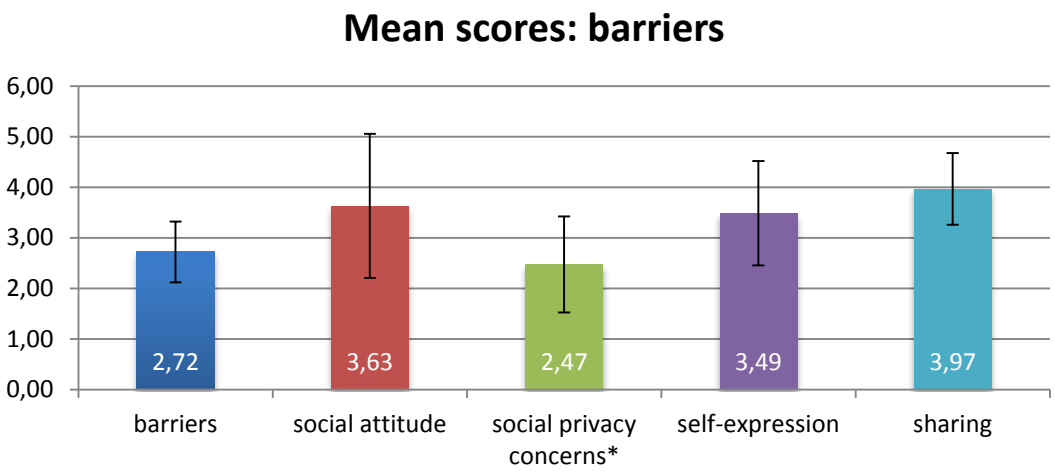

*Figure 5.3.5. Mean scores for possible barriers including general barriers, social attitude, social privacy concerns (*rated on a 4-point Likert scale), self-expression and sharing*

5.3.4.2 Level 2: Learning

The goal of the MoodMap App is to provide users' an easy and integrated tool to track their mood during work. Thereby the participants should become aware of their mood and how their mood influences their daily working tasks. Reflecting on their own mood development as well as on the development of their team's mood during a working day and establishing a relationship with their work, can give them explicit awareness and new individual insights about their attitude to work, confidence as well as skills. Not only the individual mood can help them acquire new insights, but also the mood of the team colleagues can contribute to detect discrepancies and consequently acquire new knowledge.

Learning Process

App-specific reflection questions

Altogether eight app-specific questions of the evaluation toolbox were included in the post-questionnaire, seven regular questions and one control question. Figure 5.3.6 presents the mean ratings (from 5-point Likert scales) of the app-specific reflection questions and the corresponding control question of the whole evaluation and per team. The overall mean shows that the participants slightly agreed that the application has potential to initiate reflection by capturing data relevant for reflection and visualizing data to reconstruct working experiences as well as capturing learning outcomes. Very interesting is the rating of the single teams. While Team AMc (*M* = 2.09 (*SD* = 0.99)) and Team AMo (M=2.70 (*SD* = *1.48*)) rated the app specific reflection questions rather neutral or slightly disagreed, the other two teams Team GMa (*M* = 3.44, *SD* = 0.69) and Team STh (*M* = 4.25, *SD* = 0.96) rated it very positive. Considering the app-specific reflection ratings of the most active teams showed that the teams with higher MoodMap App usage had higher ratings concerning the app's potential to support reflection.

The "app specific reflection control" question asked if the application is able to show how many calls a user had had during a day. With the mean ratings of the summarized value *M* = 2.74 (*SD* = 1.15) the participants clearly disagreed. Only Team STh with *M* = 4.11 (*SD* = 0.85) agreed, although there is no direct technical connection of the MoodMap App and the call-centre's calling system.

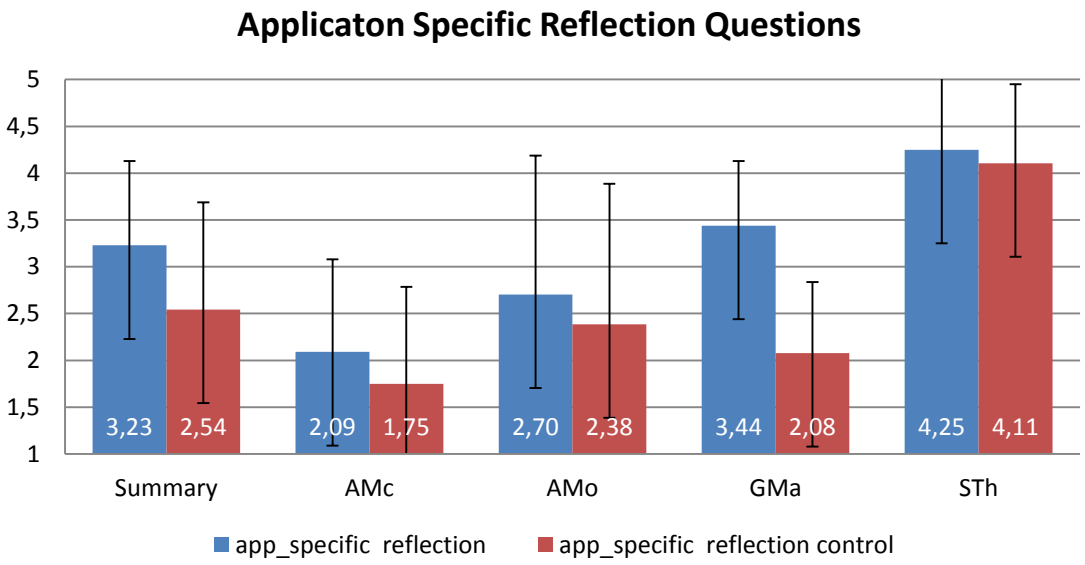

Figure 5.3.6. Mean rating for 7 application specific reflection questions per team

Analysis of notes

All together the participants have captured 991 moods and attached 983 non empty notes. After analysing the notes against the reflection-coding schema (see section 2.2), 239 notes could be identified as individual reflective items, the remaining notes were not work related and could therefore not be addressed to any of the categories. Three different researchers conducted the first rating of the notes independently. The inter-coder reliability exceeded an average compliance with regard to the following categories: Category 1: Experience or issue/problem report: 87%, Category 2a: own emotions: 91%, Category 2b: emotions of customers: 93% and Category 3: interpretation and justification: 92%. Category 4 and 7b were not mentioned here because only very view notes were assigned to these categories. After the first categorisation round, the researchers discussed all the notes where their ratings differed in a second round, which resulted in 100% accordance. Each note could be assigned to more than one category. Most of the notes were rated with Category 2a, which describe own emotions of the participant. Figure 5.3.7 shows a summary of the notes to the corresponding categories.

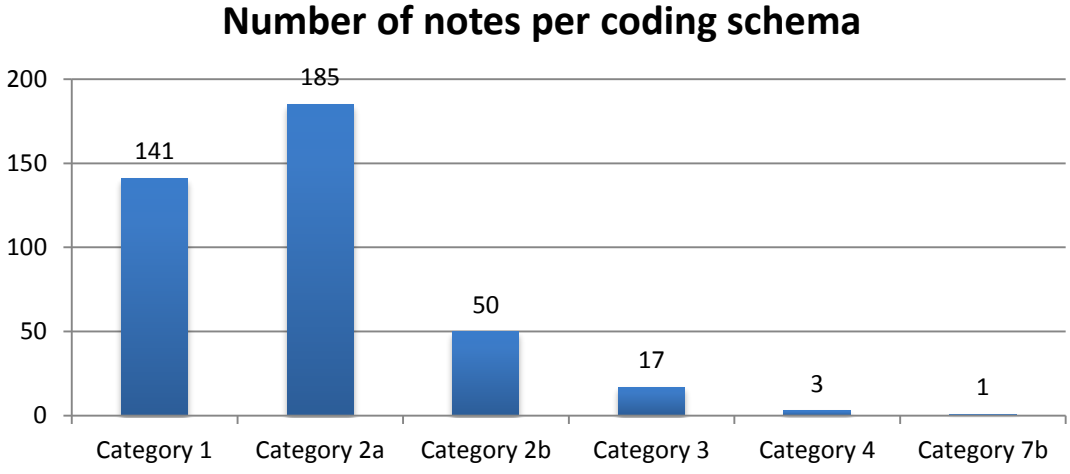

Figure 5.3.7. Number of notes per category

Table 5.3.2. Examples of categories and corresponding notes

Category	Example of notes
1: experience or issue	"Coach is now dealing with the horrible case and its Friday! :)"
2a: own emotions	"Talk with manager, feeling a bit more positive"
1, 2b: customer emotions	"Got customer information and he is happy"
1, 3: interpretation of actions	"Back and forth we go, another day of getting nowhere with our control desks. Honestly not sure why the customer wants to stay with BT at this stage, no one."
1, 4: linking of experiences	*"Oneview* crashed during my 1st call. Second morning in a row. Emailed screen shot to my manager"
1, 7b: solution suggestion	*"KCI 2* for important job and customer did not have much of a clue and had unrealistic expectations. Will have to refer to sales to move"

Short Reflection Scale (pre- and post-)

Figure 5.3.8 shows the mean ratings obtained for the Short Reflection Scale (SRS) as well as the two subscales concerning individual and team reflection. Comparing the scores on all three levels, we found no significant differences between the SRS of the ratings of pre-questionnaire to the ratings of the post-questionnaire. For the overall comparison (N = 26, here we used only those answers of the participants, who have filled in both questionnaires) of the SRS the mean in the pre-questionnaire was M = 3.98 (SD = 0.49) and in the post-questionnaire M = 3.89 (SD = 0.53).

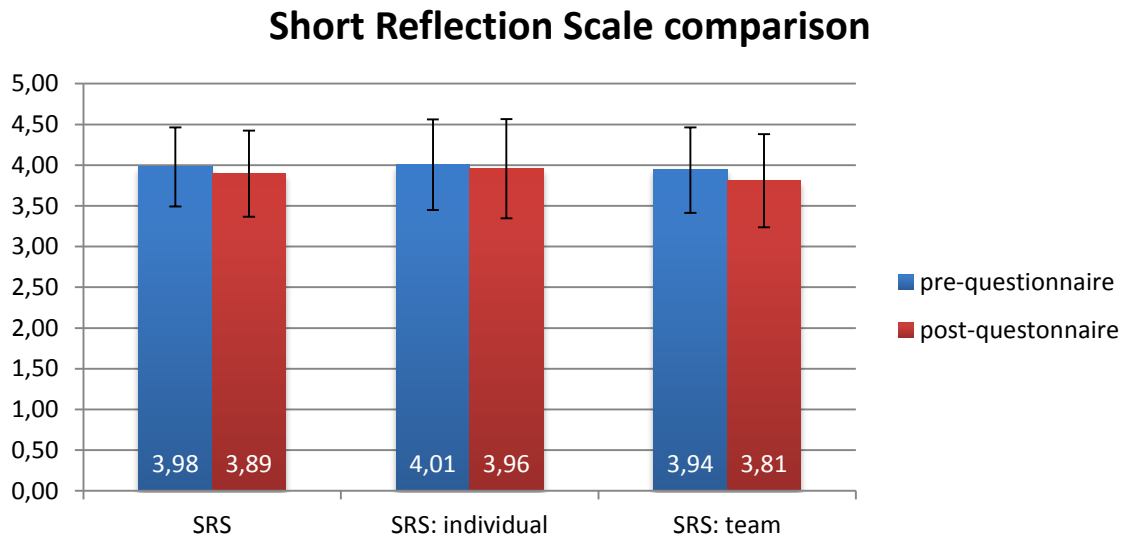

Figure 5.3.8. Short Reflection Scale before and after the usage of the MoodMap App

Furthermore, we conducted a Kruskal-Wallis-Test in order to compare the four different teams to each other, but no difference was recognizable (all p > .05), neither within the pre- nor within the post-questionnaire. Additionally, we also conducted a related t-test across all teams (N =

26), which also shows no significant neither between the overall scores nor the individual or collaborative scores (see Table 5.3.3).

Table 5.3.3. t-test values for the analysis of SRS scores in pre- and post-questionnaire

	t-value	degrees of freedom	p-value
SRS_Pre - SRS_Post	.998	25	.328
SRS_Ind_Pre - SRS_Ind_Post	.528	25	.602
SRS_team_Pre - SRS_team_Post	1.145	25	.263
SRS_Ind_Pre – SRS_Team_Pre	.689	25	.497
SRS_Ind_Post – SRS_Team_Post	1.402	25	.173

Learning Outcomes

The learning outcomes stated within the post-questionnaire (N = 38) were seen as nearly neutral M = 2.87 (SD = 0.99), i.e. participants could not decide whether they gained a deeper understanding of their work life and what to change about their work behaviour with or without regard to the coaching session. Having a closer look on each team, the mean score mentioned above also represents the scores of Team AMc, Team AMo and Team GMa. Only Team STh with M = 4.00 (SD = 0.72) agreed to have gained a deeper understanding with regard to their work-life and what to change about their work behaviour. This polarized result may also be influenced by the fact that only four participants of this team have filled in the post-questionnaire.

The analysis of the open questions from the post-questionnaire provided more details about their understanding of their work life and their conscious decisions of what to change in the future. Only five participants agreed to have made a conscious decision on how to behave in the future. On the one hand they would try to be more positive regarding distressed customers and on the other hand they mentioned the decision to let things at work affect them less. Five participants agreed to have gained a deeper understanding of their work lives. It was mentioned that a larger work load causes more stress at different times of the day. Another statement said that they would be more prone to remember the things that went wrong rather than what went right. The MoodMap App was unfortunately not used in the coaching sessions; therefore no change in behaviour can be mentioned with regard to the coaching sessions. From the coaches and the managers we received only four ratings M = 2.5 (SD = 1.73) with regard to making a conscious decision about how to behave in future in coaching sessions or team meetings.

5.3.4.3 Level 3: Behaviour

Behavioural change (measured by question CB1) was rated by the participants with M = 2.77 *(SD = 1.27)*, which implies that the participants tend to state neutral or slightly disagreed that the MoodMap App helped them to improve their work at the call centre (e.g. during the shift, after customer's calls, during and after coaching sessions, during team meetings or in other situations). 35% of the participants agreed with the MoodMap App helping them to improve their work performance and mentioned that their work is improved by becoming aware through the app. They stated that that if their mood was low, they tried to lift it, that they would try to

have a more positive attitude towards negative situations. A participant also mentioned that he/she would use the MoodMap App to state how much work they had completed.

We also asked the participants if they had recognised any improvements in the attitude of their managers and coaches. Most of the answers stated that the attitude of the four managers and their coaches was already great, being approachable, supportive and easy to interact with, so they had not noticed any further improvements. Nonetheless, in Team AMc a participant recognized that their manager had shown little more concern about their current moods and in Team AMo another participant confirmed having recognized improvements, but unfortunately did not give any further details.

Further questions on work improvement were divided according to the participant's role. Coaches and managers were asked if they had noticed any improvements in the work performance of their advisors since they started using the application. Answers to this question were available from teams AMo and GMa, both from the manager and a coach respectively. Although the opinions differed (one coach and one manager agreed by rating 4, whereas the other coach and manager disagreed by rating 1), we could appreciate a direct relationship with the question on making a conscious decision about how to behave in future mentioned above (Learning Outcomes). The manager or coach that agreed with having made a conscious decision and behaving differently also noticed improvements in the work of his/her advisors, and vice versa.

Table 5.3.4 shows the mean and standard deviation for several questions related to participants' behaviour. For the comparison of the pre-test and post-test situation, only the answers of the 26 participants whose data for both questionnaires is available were taken into account. As the results show (see coaching_sessions in Table 5.3.4) the perception of the participants was slightly improved. Average rating to the question whether being aware of own emotions during customer calls helps to reduce the advisor's customer repeats (number of times a customer has to call again to solve an issue, which should be always solved "Right First Time"; see customer_repeats in Table 5.3.4) was also slightly reduced. Participants were also asked whether reflecting on their work and on their emotions helps them to improve their customer satisfaction. The comparison between the answers before and after the evaluation (see work_customer and emotions_customer in Table 5.3.4) show that participants were more confident after the evaluation regarding reflection on, but they were less confident with reflection on their moods getting to affect customer.

Table 5.3.4. Results to questions regarding participants' behaviour with a 5-point likert scale (N = 26)

	Pre-questionnaire	Post-questionnaire
coaching_sessions	*M = 3.87, SD = 1.45*	*M = 4.00, SD = 1.26*
cover_coaching	*M = 3.87, SD = 1.45*	*M = 3.87, SD = 1.45*
customer_repeats	*M = 3.80, SD = 1.16*	*M = 3.43, SD = 1.43*
feedback_processes	*M = 3.68, SD = 1.03*	*M = 3.68, SD = 1.03*
reflect_feel	*M = 3.40, SD = 1.34*	*M = 3.40, SD = 1.34*
work_customer	*M = 3.32, SD = 1.33*	*M = 3.80, SD = 1.16*
emotions_customer	*M = 3.74, SD = 1.54*	*M = 3.32, SD = 1.33*

Accordingly, we asked coaches and managers at the beginning and at the end of the evaluation if they are aware of the coaching needs of their team. From the answers of 2 coaches whose data for both questionnaires is available, no changes were produced in that direction, having both a pre- and post-value of $M = 4.00$ ($SD = 1.09$).

Additionally, the conducted Pearson correlation showed that there is a positive correlation between the customer's average rating of the participants after the usage period of the MoodMap App and the user's comfortableness with expressing feelings in general ($r = .635$, $p < 0.026$ for $N = 12$). This shows that the usage of the app may have a positive impact on better communication with emotions and advisor's empathy, resulting in a better rating of the services by the customers.

Regarding the Short Reflection Scale score of the participants and the customer's average rating of the participants, it could be also proved through a Pearson test, that they positively correlate with a score of 0.586 ($p < 0.035$) for $N = 13$. This implies that, for these 13 participants the ratings for this KPI is higher when the user has a higher reflection score. That means that the usage of the app has a positive impact in both parameters, although this interpretation can only be confirmed for those 13 participants whose data was available.

Additionally to the feedback gathered with the questionnaires, gained insights and behavioural changes were collected through several interviews. Interviews with two managers and one advisor were conducted. Feedback from both managers was very positive, especially regarding the insights they gained about their teams and how this positively affected the teams' work. Also the advisor from Team Sth perceived the experience with the MoodMap App as very positive and he learned some insights when comparing himself to the team.

Possible changes a manager was considering thanks to the MoodMap App were modifying the time or maybe even the organization of the huddles themselves if the advisors come in a low mood or also defining some new tasks or activities for certain points in a day where the mood dropped. Also insights regarding their individual mood were mentioned by a manager *"As the moods were easily visible, this allows the managers to incorporate their coaching style to reflect the current mood*" and these insights made him also change his behaviour at work: *"when my energy level is low, I leave the desk and do something else until my level is up again. Then I return to my desk"*.

Insights for uptake in other teams were also discussed: *"I think it definitely improved my insight into my team. And I think it would also be useful for managers that have like a team that work away from the offices, that work at home so you can't always see them, but you can see of what kind of moods they are in - and how they're feeling and stuff. You know, if I wasn't in the office, I can still see how my team is feeling. […] It is definitely an improvement for me".*

The interviewed advisor commented on insights he had gained by reflecting on his mood and highlighted the fact that he could see the mood of his colleagues and compare himself to them: *"Yeah, I like the way you can see the team members as well, you can see where they are, or you sort of wonder yourself why are they there, or why are they are up there and I am down here or vice versa. So you sort of wonder and I ask these things to myself if they have to had a really bad day or have to had a bad call or just generally feeling unavailable […] it is quite a good thing to look at, and I always compare myself to others"*. He also mentioned that reflecting on a certain call helped him move on to the next customer feeling better and not being affected by past negative experiences.

The advisor also admitted not having used it in his coaching sessions, but being willing to use it: "*I think we should use them more in coaching, I think we should incorporate it in peer-*

coaching, but make it positive, don't reflect on a negative thing, […] So everybody likes positive feedback, nobody likes negative feedback, so I think that by using it in coaching you should always make it positive…".

Further details about the interviews can be found in section 12.1 Interviews with participants.

5.3.4.4 Level 4: Results

A paired-samples t-test was conducted to compare the individual KPI metrics before the usage of the MoodMap App and after it. Concretely, the metrics were provided for three reporting periods:

- Before using the app: 1st August – 21st October (81 days)
- While using the app: 21st October – 19th December (59 days)
- After having ceased using the app: 6th January – 7th February (32 days)

The provided KPIs were *volume*, number of customers that delivered a rating to the advisor after the call, and *average rating*, which indicates the customer satisfaction rating (0-100). Data was available for the Teams GMa and STh.

A comparison of the time period before the app was used with the period, in which the MoodMap App was used (during: from 21st October until 19th December) was conducted with paired-samples t-tests for the Teams GMa and STh. The results are shown in Table 5.3.5 below.

Table 5.3.5. Descriptive Statistics and t-test Results for Volume and Ratings for Teams GMa and STh before and during the usage period

	before		during					
Outcome	M	SD	M	SD	n	t	df	p
GMa Volume	10.68	4.08	10.05	3.95	19	.55	18	0.59
GMa Avg. Rating	82.79	7.98	89.58	5.81	19	-3.39	18	0.003
STh Volume	27.65	17.00	24.41	17.91	17	1.00	16	0.33
STh Avg. Rating	82.82	8.24	83.00	11.86	17	-0.06	16	0.95

As displayed in Table 5.3.5, there are statistically significant differences, at the .01 significance level, between the periods before and during app usage scores for the average ratings in Team GMa, but not for the volume. The results show that the *average rating* delivered by the customers increased on average 8.20% during the period where the MoodMap App was used, which is a very positive result. However, the impact of the application in their work environment cannot be totally isolated and other factors may affect the development of the KPIs. As some interviews with the managers of the call centres revealed, other changes in the company due to measures they are taking simultaneously may also affect the KPIs.

Regarding Team STh, *volume* was slightly reduced in that period whereas the *average rating* slightly increased. However these results are statistically not significant.

The data of the period of app usage was also compared to data of the period after the usage of the MMa was ceased (19th December to 7th February). The intention of these measurements was to follow up if the identified improvement in the KPIs could be maintained for longer period of time after the usage of the app.

Table 5.3.6. Descriptive Statistics and t-test Results for Volume and Ratings for Teams GMa and STh during and after the usage period

	during		after					
Outcome	M	SD	M	SD	n	t	df	p
GMa Volume	10.05	3.95	3.47	2.01	19	6.30	18	<.001
GMa Avg. Rating	89.58	5.81	89.33	13.46	19	0.013	18	.99
STh Volume	27.00	17.46	10.27	5.81	15	4.06	14	<.001
STh Avg. Rating	84.27	8.20	77.00	21.69	15	1.27	14	.23

As displayed in Table 5.3.6, there are statistically significant differences, at the .01 significance level, in scores for the time in which the MMA was used compared to follow-up scores for *volume* in both teams. According to the managers at BT, the period in January was affected also by extremely bad weather, which made leak times increase and therefore there were less resolutions of cases. This also causes more stress in the advisors and less satisfaction in the customers. The significant reduction in the volume metric was on average 65.45% for GMa Team and 61.98% for STh team. With regard to the *average rating* of the advisors during the period after the app usage, it was slightly reduced in both teams, but the reduction was not significant according to the t-test (Team GMa -0.06% and Team STh -8.62%).

For the teams GMa and STh data regarding Key Performance Indicators at team level were provided (see Table 5.3.6) at the same three reporting periods mentioned above. These metrics are provided by the customers through SMS message once they have spoken to the advisors. The available KPIs are the following:

- *Net Promoter Indicator (NPI)*: it is based on customer advocacy and reflects the answers to the question 'How likely are you to recommend our services to others based on your recent experience with us'. In terms of percentage can range from -100 to +100.
- *Advisor Satisfaction (Advisor Sat)*: indicates the customer overall satisfaction with the call. Customers answer in a scale 1-10 and the percentage is calculated dependent on how many customers score the advisor and what the score is.
- *Recap*: indicates whether the advisors proactively summarized the call to the customer, in order to help assist with lowering the amount of repeat calls they receive as a business. The question that the customer answers is 'Did the last advisor recap what had been agreed?'

Table 5.3.7. Results for change in percentage of team KPIs during and after the app usage period, for teams GMa and STh.

Team	NPI		Advisor Sat		Recap	
	during usage	after usage	during usage	after usage	during usage	after usage
GMa	40.00	17.14	7.14	-1.11	4.71	-3.37
STh	16.67	-34.29	1.19	-3.53	3.80	-1.22

The metrics from Team GMa show that *NPI* increased 40.00% (from a score of 25 to 35 points) during the app usage period, whereas its improvement afterwards was only 17.14% (from 35 to 41). The metrics from Team STh show a similar behaviour. However, the difference between the usage period and the period after usage regarding the NPI is major, having an improvement of 16.67% with the MoodMap App and a decrease of 34.29% during the period after the cessation.

Regarding *Advisor Sat* and *Recap*, both metrics were slightly improved during the MoodMap App period, but they decreased minimally after the usage cessation. This same behaviour was detected in both teams (see Table 5.3.7 for details).

The question related to the loyalty metric was only answered by a low number of participants (*N* = 38) and these were users who were not so active in the app, while we lack the answers from the most active users. Due to this fact, the loyalty metric could not be taken into account for the analysis.

After the evaluation of the MoodMap App at BT call centers, also a summative evaluation with the IAA and IMA was conducted at the same testbed (see Chapter 5.7). In the interviews performed by DFKI and the BT management, questions about the MoodMap App and the interest to continue using it were included. One possible approach suggested to the interviewees was to integrate the MoodMap App and the IAA/IAM in their own system on place. This approach was positively received and in the concrete case of the MoodMap App 60% of the interviewees approved its integration. Therefore, discussions are currently ongoing to advice BT during the development of the business case as well as the internal development.

5.3.5 Conclusion & Discussion

In general we can state that the introduction of the MoodMap App v3 in a call-centre setting has a high potential to improve the coaching culture with the support of emotional awareness and reflective learning. However, the evaluation did not reveal the expected results regarding coaching support. The approach was improving their coaching session, by using the MoodMap App and the individual mood development of the single advisors as additional coaching input. During the evaluation, the coaches were not active and the MoodMap App was not used in the coaching sessions. Though this was also a goal for the management of the call centres, they admitted not having invested enough efforts to communicate this. The lessons learned from the organizing manager at BT was that he regrets not having done another call-out to coaches and advisors for better participation and for using the MoodMap App directly within the coaching session. He also mentioned that, if he does another study, he will ensure that managers and coaches are more active during such an evaluation.

One of the main challenges was to achieve a complete dataset for the questionnaires, as well as a control group. Such methodologies are not usual in their work environment and there are many barriers to achieve this. Although the organizing manager at BT tried many times to get back the questionnaires (through the respective managers) and a control group, his efforts were unfortunately not successful. This was also the case for the KPIs regarding teams AMc and AMo, which could not be made available.

As a result, the interpretation of the gathered data within the application itself, the usage of the MoodMap App derived from the log files as well as the pre- and post-questionnaires was rather difficult. The prime reason is that the users who used the application most and the users who filled in the pre- and post-questionnaires were partially not the same. Therefore we had to decide, which available combination of data can be meaningfully used for which type of evaluation result and interpretation purposes. Due to this fact, only 26 participants delivered answers for both questionnaires and therefore analysis of pre- and post- situation was rather limited.

From Team STh only 4 out of 18 participants have filled in the post-questionnaire, although this team was one of the two most active ones. Due to the low number of participants the obtained results might not completely mirror the whole team. On the other hand we had conducted two very positive interviews with the manager and one advisor of the team. The positive attitudes of these two participants as well as the low numbers of filled in questionnaires let us suspect, that significant positive results regarding the benefits of the MoodMap App were lost. The fact that teams GMa and STh captured twice as many moods as the other two teams is also remarkable.

In general we perceived ambiguous attitudes towards the usage of the MoodMap App. We have statements from finding the MoodMap App very useful and that it could be easily integrated into BT's daily working routines to comments that BT should spend their money more wisely. This shows that some of the participants really liked and accepted the application, while others could not do anything with it. This disagreement is also mirrored by the very high standard deviation values shown at reaction level analysis.

When inserting a mood in the MoodMap App, the users were compulsory asked to insert a context and a note. Therefore we predefined three types of context relevant for a call-centre namely "after a call", "after a break" and "after a coaching session", and one neutral in form of "other". Our hypothesis was that especially the first one might be chosen the most. In contrast, the results showed that the moods were not captured in the expected situations, as 62% of the contexts were classified as "other" (see *Figure 1)*. Secondly this fact also showed that it is important to leave users space for freedom in order to find out in which situations the moods were really captured during work.

Additionally we analysed the notes according to the reflection coding schema. We received a high number of moods describing a working experience, moods with emotions of the participant him/herself and moods of the customers. There were also some notes regarding interpretation and justification of taken actions. Participants did not introduce notes regarding linking an experience to different pieces of knowledge, responding to interpretation of action, working on a solution (only 1 note was rated to this category), insights/learning from reflection, or drawing conclusions and implications from reflection. This can be explained with the fact that the compulsory insertion of notes was directly related to the currently inserted mood of the participant. No further questions with regard to reflective learning at all were posed anywhere in the MoodMap App, which leaves room for improvement to this regard.

The questions of the Short Reflection Scale (SRS) were posed in the pre- as well as in the post- questionnaire with the goal to find out, if the usage of the application contributed to improve the reflective practices of the participants ($N = 26$). The values of the pre-questionnaire show that the willingness to reflect was rather high right from the beginning and did not show any significant improvement after the evaluation. An explanation of this phenomenon is that most of the participants see themselves as reflective persons, and for the initial value was rather high it is more challenging to show an improvement. Another possible explanation for this would be the relative short evaluation period, as we have no evidence how likely it is that the answers to these questions significantly change in 4 weeks.

The result with regard to reflective learning showed, that some of the participants could recognize a clear benefit for themselves but especially for the coaches and managers, if they can see how each individual advisor of the team is feeling and react to it. Some of the advisors, coaches and managers mentioned that they have gained a deeper understanding with regard to their work-life and what to change about their work behaviour. The advisors stated that they try to be more positive regarding distressed customers and to the decision to let things at work affect them less. Two of the managers and coaches also confirmed having taken conscious decisions to change their work behaviour, but unfortunately they did not report the actions taken.

With regard to the Key Performance Indicators that allow us to measure the targeted outcomes, we could show a significant improvement both at individual and team level during the period where the MoodMap App was used. In the case of Team GMa, these results were even statistically significant. Although we are aware of the other factors affecting the measured Key Performance Indicators, these were very positive results for the evaluation.

During the preparation phase of this evaluation as well as during the evaluation itself we could derive several lessons learned. First, there has to be a clear goal or benefit for all parties involved. Second it is important to have the full support at the organisational level as well as from the manager's level in order to conduct such a big evaluation successfully. It is also important that the stated goal will be implemented and established by the management or the organisation. For example, within this evaluation it was not possible to get a control group, whose task would have only been to fill in the pre- and the post-questionnaires. Third, each single participant needs to see a clear goal or benefit for him/herself when using such an application and have the feeling that there will be actions taken towards the goal from the side of the management. The actions which could be taken, present the following success story reported by one manager:

"I sit in a little corner of the office so I don't actually get a chance to interrupt with all of my team all the time. So I find the MoodMap App very useful to see how everyone was feeling cause not everyone obviously comes to tell you how they are feeling and I had one guy, he sits quite far away from me and he was on a really hard time with a very difficult customer. Off site and all by the way he made a comment on the MoodMap App of having a really hard time and that he was not feeling like that he was getting any help. So straight away I went over to him and asked what I could do to help him and an hour later his mood had gone from like really low to really high because I had gone over to helped. […] I would have never known about that and he would have probably struggled on, so there sitting without me knowing anything."

The successful evaluation of the MoodMap App as well as the described success story at BT also contributed to the organisation of another evaluation at another MIRROR testbed, namely REGOLA. The good practices and examples that emerged from this evaluation also facilitated the initialization of the evaluation as well as increased the interest of the organisation itself.

Concluding, we see that the MoodMap App has the potential to trigger reflective learning and especially to facilitate the improvement of coaching sessions in a call-centre setting like the one described above, although further efforts at organisational level will be needed to this respect. Using the MoodMap App in such a setting, with a predefined and clear goal for all participants might lead to a very good exploitable for the whole MIRROR project with regard to supporting coaching sessions in business environments.

5.4 The MoodMap App evaluation at Regola

Please find a short description of the app as well as some theoretical assumptions which are identical to the BT evaluation of the MoodMap App in section 5.3.

5.4.1 Organisational context

Test bed organisation and the organisational unit

Please find a general description of Regola in section 3.6. The evaluation was set up for all 5 departments, that is all employees of the company participated in it. Each of the departments is led by a manager, one of the managers is responsible for two departments. Each department consists of 4 to 16 staff members.

Test users and their job roles

Regola has different departments, from development, service desk and support to sales as well as management. Therefore, there are several teams and profiles, according to their competences i.e. system technicians, developers, project managers, call takers from service desks, sales consultants and marketing staff. In the following we describe all these departments.

Department Am consists of four staff members. Their daily main tasks include controlling, finance, accounting, secretary activities, personnel management, banks relationship and registration trademarks.

Department Co consists of one sales director and three staff members. Their daily working tasks encompass scouting, tender/conferences/exhibitions/fairs attendance, preparation of events, social management, project and opportunities evaluation, assisting clients and partners by clarifying and documenting objectives, account management and contact (B2B) as well as the preparation of demos.

Department Qu consists of one operational director and six staff members. The employees consist of software skilled operators, application specialists, system technicians and network security specialists. Their main tasks are to provide first, second and third level support and assistance to clients. These activities are mainly performed via mobiles, web instructions, remote connections and eventually on-site actions to investigate and solve critical issues, according to predefined Service Level Agreements (SLA).

Department Pm consists of one chief technical officer (CTO) and five project managers. The major tasks of the CTO are to plan the global activities, design global projects, choose the right technologies and organise coordination meetings. The CTO receives a huge amount of input from each of the projects. Additionally he keeps control on the global design of the projects in terms of integration, choosing and deepening new technologies and he is also interacting and handling requests from sales department. The project managers deal with team leading and organize the corresponding meetings. Furthermore they are responsible for the requirements engineering processes.

Department Sv consists of the same CTO as Department Pm and of 16 developers. Their main tasks are new developments, bug fixing and testing of the produced code.

Identified need and potential for reflective learning

The use of the MoodMap App was initially identified by the human resources manager of the company with the goal to provide a possibility for self-reflection, self-development and stress detection during work.

Possible stressful situations were identified per department:

Department Am: the rectification of errors, the respect of deadlines, collecting all the necessary information from other areas of the company, difficult customer contact (who persist in not paying the outstanding), obtaining financial resources.

Department Co: simultaneous requests and activities, imminent deadlines of tender or commercial opportunities, complaints or difficult relationships with clients or competitors. Furthermore the introduction of several different projects to potential clients, packaging and communicating to recipients all the latest innovations from the production area as well as handling periodic or yearly contracts.

Department Qu: the crash of an emergency centre or a critical problem support in the global infrastructure, tasks registered with high priority, e.g. typically blocking issues occurred in our software applications, providing support by voice especially in another language than the native language, having a 24h turnover to support fix critical issues (e.g. in an emergency centre).

Department Pm: hard deadlines, deepening and studying new technologies, doing requirements analysis documentation, checking the developers' progress and debating issues with them. Additionally they have also to discuss with final customers.

Department Sv: meet the project deadlines, provide urgent bug fixes if necessary and use new technology applied to projects.

During their daily work stressful and emotional situations occur very often. The evaluation of the MoodMap App in this setting has a high potential to help employees become aware of the emotional activities in each department and to reflect about them.

Potential organizational impact

In Italy there is a law that certain companies and in this case REGOLA have to verify the level of stress of their employees. For the management of Regola the MoodMap App was seen as a promising tool to support them with lowering stress levels of each single employee but also to help the whole organisation.

The overall goal from the organisational point of view was that this evaluation should

iv. create awareness that the organisation is caring about their staff

v. create awareness in managers that all individual employees and their work is important for a company

vi. create self-awareness about the employees own mood development and stress-levels.

As a result they would like to understand how they can improve the mood of their employees or how to change particular things, which came up during the usage of the MoodMap App.

5.4.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

Within the MoodMap App the user has the possibility to capture her individual mood states, which she has during a working day or during a meeting. The mood is composed of two levels namely the valence level (feeling good – feeling bad) and the arousal level (high energy – low energy). Additionally, the user is motivated to think about her current mood with the help of so-called reflection interventions and reflection amplifiers. These interventions ask to actively use the MoodMap App, while the reflection amplifiers make aware of significant mood changes on an individual or collaborative level. These pop-ups should trigger the users to think about their mood and the relationship with their work during the usage of the application.

When presenting the individual mood development on the timeline visualisation, the context information as well as the notes (if available) are visualized in the same graph to provide the user information about her mood development. When having a look at this presentation of the mood development during the day, it enables the user to recall past working experiences more easily. During a working day the user also has the possibility to compare her mood with the mood of the associated colleagues of the same department. One visualisation called CompareMe shows the valence and arousal level split up into two bars and shows the user's level in comparison to the average department levels. The second visualisation call Collaborate shows the average department member mood in form of a red cross directly on the mood map. After clicking on the cross, it shows all individual mood points of the department members in an anonymous way.

When a working day is over, two different reports are available which cover a summary of the day as well as the average mood development of the department members during that day. Additional statistical information is also available covering the key indicators about the shift like number of participants in the same shift, number of moods entered etc. The user has also the possibility to insert insights or learning outcomes in the reflection journal. These entries can be used later on for further learning when revisiting the data of the meeting again.

Additionally, there exist special views regarding each single department, which are only accessible by managers. These views give the managers more insights about the moods of each single employee of his/her department. The day visualisation presents the current mood status of each single employee in form of a smiley. By clicking on the smiley, the corresponding mood timeline of the employee is visualised. The daily timeline visualisation shows in two timelines the average valence and arousal development of the department during the day. Additionally each single mood points are added here as well. The third visualization, the weekly timeline, presents the mood development of the whole department members during a week. Altogether these views serve for managers as triggers to reflect upon the mood development of their single employees as well as the mood development of the whole department. This reflection should lead to outcomes that make them more proactive with regard to their employees.

In order to use the application while working, it was directly embedded in their working processes and their daily routines. In this setting, the MoodMap App should give the individual the possibility to reflect about her own mood as well as the mood of the whole department. With the help of some guidance offered by the app, the participants should be activated to re-evaluate their mood or mood changes with the goal to learn from it. This includes learning from critical reflective thinking about own working behaviour during a certain meeting, a customer call or any other challenging working situation.

In this first trial, the MoodMap App was not integrated in Regola's own system, but it was made accessible directly from the employees' desktop (through an icon). The manager responsible for this evaluation from Regola's side made the participants aware during the introduction of the MoodMap App, to use the application in different situations and significant work-related activities e.g. if the service desk has to deal with a particular request they should click on the mood. Very short meetings with the corresponding managers were planned in order to find out how the mood or particular things according to the results of the MoodMap App can be improved. Additionally the manager responsible for this evaluation also wanted to study where the best places for the MoodMap App will be in Regola's daily activities.

Within this evaluation the MoodMap App v3 (version 3) was used. The MoodMap App v3 is described in detail in Chapter 7 of D9.5 Whenever the term MoodMap App is used in this document, we refer to MoodMap App v3. Only when referring to another version of the MoodMap App we explicitly mention the version number.

Relation to MIRROR CSRL Model

As the relation to the CSRL model is almost the same as for the MoodMap App evaluation at BT we only highlight here the differences between the two app versions.

Plan and do work

This stage is supported in the same way for both evaluations. The only difference is that users do not have to enter compulsory a context and a personal note to a mood entry.

Initiate reflection

In order to initiate reflection, the application follows two approaches.

First, the MoodMap App has a reflection guidance mechanism, which motivates users to reflect about the individual mood or the department mood during the usage of the app. During the usage of the application reflection interventions and reflection amplifiers in form of pop-ups appear. The first ones by motivate the users to actively use the MoodMap App. The second ones make aware of significant mood changes on an individual as well as collaborative level. By thinking about and answering the posed questions in the pop-ups, the users start to reflect or at least think about the currently inserted mood.

Secondly, reflection can also be initialized by exploring and analysing the different visualizations and data reports, which is the same in both MoodMap App versions.

Conduct a reflection session

See BT evaluation.

Apply reflection outcome

See BT evaluation.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

See BT evaluation. Again, the only difference is the voluntary entry of context and notes and the providing of reflection interventions and reflection amplifiers in form of pop-ups with reflective questions, which can trigger reflection.

5.4.3 Research approach

Design and procedure

The summative evaluation of the MoodMap App at Regola started on the 12th of February and lasted until the 31st of March 2014. The MoodMap App was introduced by a responsible project manager of Regola to all five departments. During this introduction the project manager described on the one hand the MoodMap App itself and what it should be used for and on the other hand he presented a success story which emerged during the MoodMap App usage at BT (section 5.3) in order to show them a meaningful insight and a clear benefit. During this introduction phase the department members were asked to fill in a pre-questionnaire including the MIRROR consent form. During the application usage period, all participants of the trial were asked to insert their moods during their working shifts and to reflect about their inserted moods and notes individually. The managers were additionally instructed to use the team visualisations in order to reflect about the mood development of their departments and take actions if necessary.

At the end of the trial, all participants were requested to fill in a post-questionnaire. Additionally, seven participants of the evaluation took part in an interview with the researchers with the aim of gaining more insights about the evaluation and the usage of the MoodMap App at Regola.

Participants

Out of 38 employees of Regola, 35 have participated in the MoodMap App evaluation. The reasons for not participating were not related to the evaluation itself, but it was due to personal or professional reasons. The 35 participants were split over the five following departments: Department Am, Department Co, Department Qu, Department Pm and Department Sv. The 35 participants consisted of 27 male and 8 women. Participants in the evaluation were between 20 and 59 years old, with 9% of them aged between 20 and 29, 77% between 30 and 39, 11% between 40 and 49 and 3% between 50 and 59. They had worked on average $M = 6.77$ ($SD = 1.52$) years in their current position, ranging from participants with less than one year in that position and to a maximum of 15 years. The average years in a similar position were $M = 4.14$ ($SD = 2.96$). 80% of the participants are working full-time, 17% are working part-time and one participant has not stated his job scope.

All participants have filled in the pre-questionnaire and the post-questionnaire. Two of the participants have filled in the pre-questionnaire after the trial, so only demographic data was used for analysis (questions regarding reflection practices and expectations were not considered). One participant has filled in the post-questionnaire with invalid data; therefore these answers were also removed from the dataset. Altogether 32 participants have filled in both questionnaires and their answers constitute the dataset used for comparisons of pre- and the post-questionnaire variables.

Summative evaluation methods used

Besides capturing log-data, two questionnaires were used: A pre-questionnaire, which gathered demographic information (as outlined in the toolbox), contained the short reflection scale (SRS) and questions about participants' expectations with regard to the use of the MoodMap App. The post-questionnaire consisted of all questions concerning the evaluation levels 2 through 4.

Level 1: Reaction: all usage data, including for example the number of clicks per visualisation, can be obtained from the log files of the MoodMap App. The database of the application itself store the information about the data captured, i.e. mood values, notes and context data. In the post-questionnaire, questions regarding usage (USE) and user satisfaction (SAT) were added.

Level 2: Learning: The short reflection scale (CR) was included in the pre- and post-questionnaires. Additionally, the post-questionnaire contained 8 app specific reflection questions (CA) which fitted best to our approach as well as two Learning Outcome questions (CL).

Level 3: Behaviour: We used the core behaviour question with regard to work performance (CB).

Level 4: As KPIs could not be provided by Regola, concrete questions about their KPIs were asked to the different departments before and after the evaluation (e.g. number of bugs per week or number of hours dedicated to a certain task). The post-questionnaire included additional questions about employee satisfaction as well as improvement of work performance at individual and collaborative level (WP).

Additionally, the pre-questionnaire contained questions about expectations (EXP) the participants might have with regard to the use of the MoodMap App.

The post questionnaire consisted of further questions referring to usage (USE) and user satisfaction (SAT), general application effects (GAE), long-term usage (LT), benefits and insights as well as the impact of reflection in their daily work.

After the trial, two managers and five employees of different departments took part in an interview over Skype. The posed questions covered the following topics: feedback to the overall experience, subjective feeling of the application's acceptance within the team, capturing mood, benefits and insights as well as general comments.

Several concrete hypotheses were investigated during this evaluation at different evaluation levels:

- Reaction (Usage): The long-term usage of the App is influenced by the willingness to share moods, the motivation to reflect about moods, the dealing with emotions in general and the loyalty metric and vice versa.
- Learning: Comparing one own mood to the team mood creates curiosity in the users and allows users to detect discrepancies that can initiate a reflective process.
- Behaviour and Results: Participants, who learned due to the usage of the MoodMap app and also changed their behaviour, have rated the KPIs (reflect on feelings to improve individual and collaborative work performance) with a higher value.

If and how the hypotheses were met is described in the corresponding section below.

5.4.4 Results

5.4.4.1 Level 1: Reaction (Usage)

The participants used the MoodMap App on 48 consecutive days. Except for 12 days in which the app was not used at all (corresponding to days off), on average 10 hours and 32 minutes (*SD* = 0.129) lie between the first and the last logged event of every day. In that time users were entering the MoodMap App repeatedly and using the different available features.

In total, 2250 moods were captured by the 35 participants during the whole evaluation period. On average, users captured 64.29 moods (*SD* = 33.27) during the whole evaluation period, ranging between 12 and 143 moods per user. Considering the moods captured in each department, on average 41.50 (*SD* = 39.71) moods were captured in Department Am (*N* = 4), 38.75 (*SD* = 21.20) in Department Co (*N* = 4), 59.17 (*SD* = 29.17) in Department Qu (*N* = 6), 79.20 (*SD* = 38.62) in Department Pm (*N* = 5), and 73.63 (*SD* = 30.54) in Department Sv (*N* = 16). These results show that participants of the departments Pm and Sv captured on average approximately twice as many moods per member than the other three departments.

Regarding the notes attached to the moods, a total of 226 non-empty notes were captured by the users and served as annotation for their affective states. The contextualization of moods was not used regularly, as only 31 contexts were captured in total by all participants.

Figure 5.4.1 shows the distribution (absolute numbers) of captured moods, notes and context for each department that participated in the evaluation.

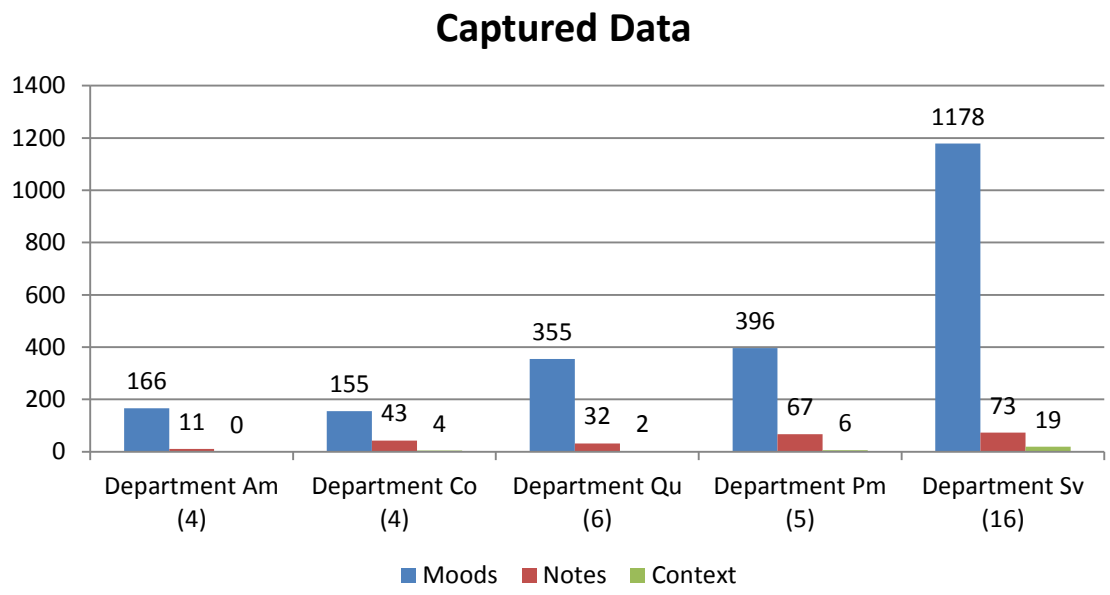

Figure 5.4.1. Absolute numbers of moods, notes and context captured by each department (with number of members of each department)

All 35 participants, who were registered in the MoodMap App, used the app during their working shift and used the main features of the app a total of 1767 times. Each of these users had on average 50.49 interactions with the application (*SD* = 101.56) during the whole evaluation period. Such a high standard deviation also shows that the usage of the app was very polarized, having users who were very active, whereas others almost did not use the app. Figure 5.4.2 shows the average usage of each feature of the MoodMap App. The three features that were most used were the Capture Mood, Timeline, and Compare Me visualizations. This was also confirmed by the interviews and the questionnaires. The preferred visualization for the users was the Compare Me visualization, where employees could compare their own mood with the mood of their colleagues. This confirms our Hypothesis 1, that comparing themselves to the team in a quick and intuitive way may be supportive and useful, both in terms of creating curiosity in the users as well as allowing users to detect discrepancies that can initiate a reflective process.

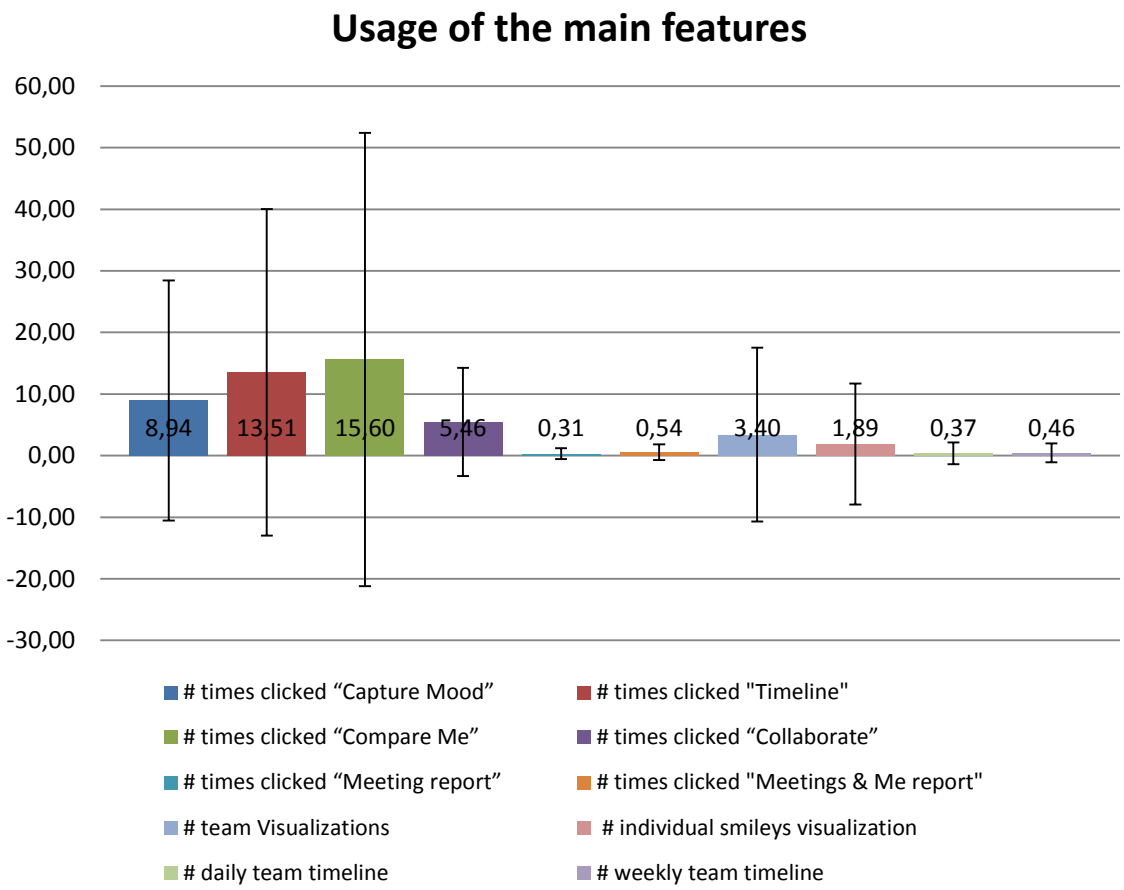

Figure 5.4.2. Usage of the main features of the MoodMap App. For each feature, average and standard deviation are depicted

Overall impression

Summarizing the answers (*N* = 34) of the open question of the post-questionnaire about the usage of the MoodMap App showed that the users' opinions were very ambiguous – from very positive over neutral to negative.

The final statements of the participants revealed again their different attitudes towards the MoodMap App. On the one hand, we have several positive statements which refer mainly to the managers' perspective and that they could use the MoodMap App to better support their team e.g. "*To be useful it should be used by managers to check the mood of its staff and possibly implement appropriate corrective actions.*" Another positive statement was "*It's a great project but maybe the effectiveness was a bit limited by the lack of time to make meetings on the subject*." Another very interesting statement was that "*MMA and the concepts in which it is based are intelligent. But the Italian culture and mentality, especially in working environments, do not give any importance to feelings / energy / health / stress of people so the use of MMA has showed little profit and 'was seen as a loss of time'*." This statement was very interesting because it is Italian law that dictates that companies have to verify the level of stress of their employees, but it seems that this has not been included in their organisational culture yet. Additionally we also received some ideas for improvement e.g. "*The app should be made much more streamlined and straight forward [...] use a faster realization, maybe a widget on your desktop"*. In contrast, there were also some critical statements regarding the usefulness of the MoodMap App, especially one from a manager: "*The objectives of the MoodMap are*

unrealistic. The idea of thinking about the own mood is good, but I just consider it inapplicable […]". He also stated several reasons why users would never state their real mood into such a tool (e.g. ethical reasons, non-work-related personal bad feeling or depressions) and questioned what one should do if mood is not optimal "*What happens if an employee is going through a bad time?*". This fact suggests that this participant may have not understood the purpose of supporting reflection with the MoodMap App and the goal of the whole evaluation, where such findings should help people improve their work and collaboration.

Satisfaction, Long Term usage and future usage

The following mean scores were all rated on a 5-point Likert Scale (strongly disagree – strongly agree). In *Figure 5.4.3* (blue bars, satisfaction) the participants (*N* = 34) stated that they are neutral or have slightly agreed to be satisfied with the MoodMap App *M* = 3.24 (*SD* = 0.85). Four departments mirror the overall mean scores. Only Department Pm *was* not satisfied with the application, although they have captured most of the moods per participant during this trial. This could be explained with the fact that they have neither used the available reports in the application nor that they have conducted any meetings to reflect about their moods within their team to gain explicitly any insights or benefits. Furthermore many of them did a lot of travelling during the evaluation period and seemed to be the most stressed ones, therefore they just captured the mood and kept working, but could not perceive any benefit for themselves.

The overall score regarding the long-term usage of the MoodMap App during work (*Figure 5.4.3*, red bar, long-term usage) is rated rather neutral *M* = 2.94 (*SD* = 1.00). While Department Co agreed to continue using the MoodMap App, the other four departments rated the long-term usage neutral or slightly disagreed to it.

The future usage (Figure 5.4.3, green bar, future usage) of the MoodMap App was rated with an average value of *M* = 2.76 (*SD* = 1.35). Especially Department Co agreed to use the MoodMap App also in the future. The other departments Department Am, Department Pm and Department Sv slightly disagreed to use the application in future. For all values, the standard deviations are rather high; this shows that the participants diverge in their opinion regarding the future usage of the app.

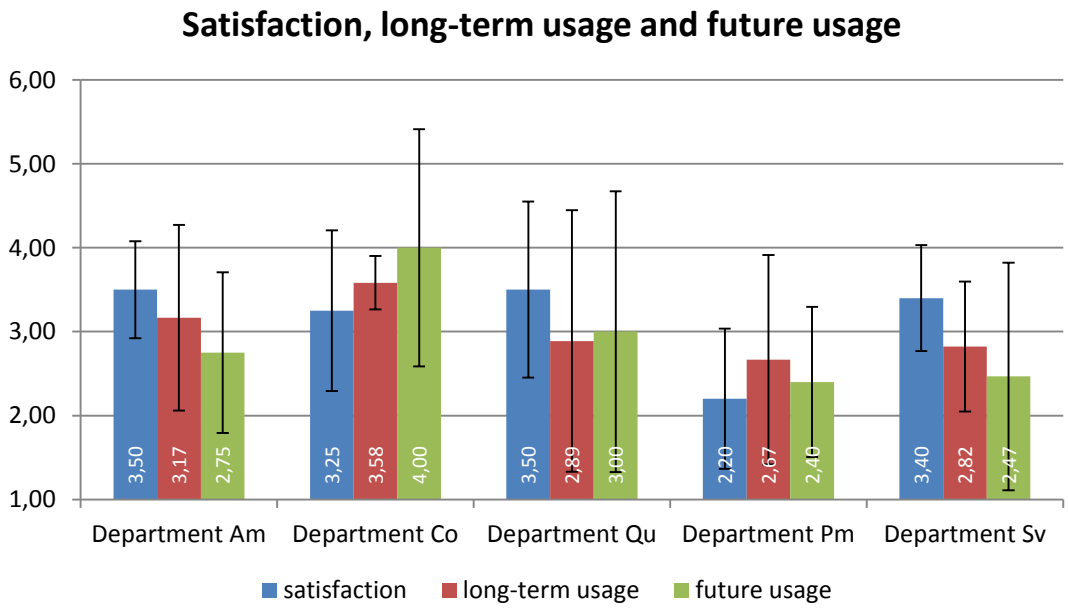

Figure 5.4.3. Mean scores of satisfaction, long-term usage and future usage

A Pearson product-moment correlation coefficient was computed to assess the relationship between several variables from Level 1. Concretely, the number of captured moods (notes and context), interactions with the application, subjective usage, self-expressiveness of feelings and social media attitude as well as sharing, satisfaction, long term usage, the motivation to reflect alone or in the team, learning outcomes, change in behaviour and KPI's were investigated.

There was a positive correlation between number of captured moods and number of interactions with the visualizations in the MoodMap App, $r = .489$, $p = .003$, for $N = 34$. This analysis reveals that most active users (e.g. the ones who visited the visualizations) were also the ones who captured more moods. Self-expression of feelings in general and during work clearly emerged, $r = .710$, $p < .001$, $N = 34$. This reflects, that people who are usually comfortable to express themselves, don't encounter additional barriers to do it at work, and vice versa.

Additionally we see a strong positive correlation between the motivation to reflect on work in general and the opinion that the app helped them to deal with their emotions ($r = .720$, $p < .001$, $N = 34$). This result indicates that the MoodMap App could help them to deal with emotions, especially for people who are more motivated to reflect on work.

A strong positive correlation exists between the number of captured moods and the subjective application usage ($r = .489$, $p < .001$, $N = 34$). This suggests that users were aware of their activities done in the MoodMap App regarding the capturing of moods. Participants who were more comfortable with sharing their moods with managers also inserted more notes in the MoodMap App, as it is shown by the positive correlation between these two variables ($r = .488$, $p = .006$, $N = 30$).

Finally, we also found a positive correlation between moods and notes ($r = .395$, $p = .019$, $N = 35$) and notes and context ($r = .377$, $p = .025$, $N = 35$).

Barriers

Figure 5.4.4 shows the scores of the possible barriers (Figure 5.4.4, blue bars) in relation to not having time, not having physical space, not having seen any advantage or not having motivation to use the app, which were rated with neutral to slightly disagreement M = 2.66 (SD = 0.58) by all departments. The social media attitude (*Figure 5.4.4*Figure 5.4.4, red bars) i.e. how likely it is that they use social networking platforms (e.g. Facebook, Twitter, LinkedIn, Google+, MySpace) was rated on average with *M* = 3.50 (*SD* = 1.08). This means that they slightly agreed to use such platforms, and that the social media attitude is mostly not a barrier to use the MoodMap App. While four of the departments mirror the average score, Department Co rated this item with *M* = 4.50 (*SD* = 0.58). This shows that the participants of Department Co strongly agree to use social networking platforms. In contrast, the social privacy concerns (Figure 5.4.4, green bars, rated with a 4-point Likert scale: not at all – high) are rated with *M* = 2.68 (*SD* = 1.17). This shows that the participants are only from *a little bit* to *somewhat* concerned about their privacy on social networking sites in general, which is therefore not seen as a major barrier to use the MoodMap App. Department Qu and Department Sv are more concerned about their privacy than the other three departments. Department Cu is least concerned regarding privacy which again matches with their overall social media attitude. Regarding the self-expression of feelings (Figure 5.4.4, purple bars) in general and during work *M* = 3.41 (*SD* = 0.84), the participants rated it neutral or slightly agreed, that they feel comfortable with this fact. While Department Co rated the self-expression very high, Department Pm rather low. The other three departments stated their self-expression rather neutral. Sharing of emotions (Figure 5.4.4, cyan bars) with managers or department members was rated with *M* = 3.53 (*SD* = 0.86), which means that sharing of individual moods was not seen as a significant problem. In this case the average value represents the scores for all five departments.

Additionally, the participants mentioned several situations when they would share their moods with their managers, for example "*Upon the occurrence of important working events that make my mood change considerably*" or "*In the situations where a manager is able to listen constructively the issues/problems identified in the work processes*". One manager refused to share his moods at all "*Never. The moods are extremely personal and depend on thousands of factors. I would not want to share it to work*". Situations where the participants would not share their moods with their managers encompass mostly non-work related private issues or situations where users are discontent or dissatisfied with their work. Again, from a manager we have the following sentence "*Never. It is not ethical to ask to share the mood in the workplace.*"

From the open questions of the post-questionnaire (*N* = 34) we received many different responses regarding the barriers of the MoodMap App usage. 26 participants did not answer this question. Three of the participants did not see any barriers to use the MoodMap App at all. Four of the participants mentioned that they have no time to use the MoodMap App because of urgent and tight timelines. One participant stated that he would like to have a faster approach to insert his moods e.g. in form of a widget. And one participant stated that that "*The graphical interface is not usable*".

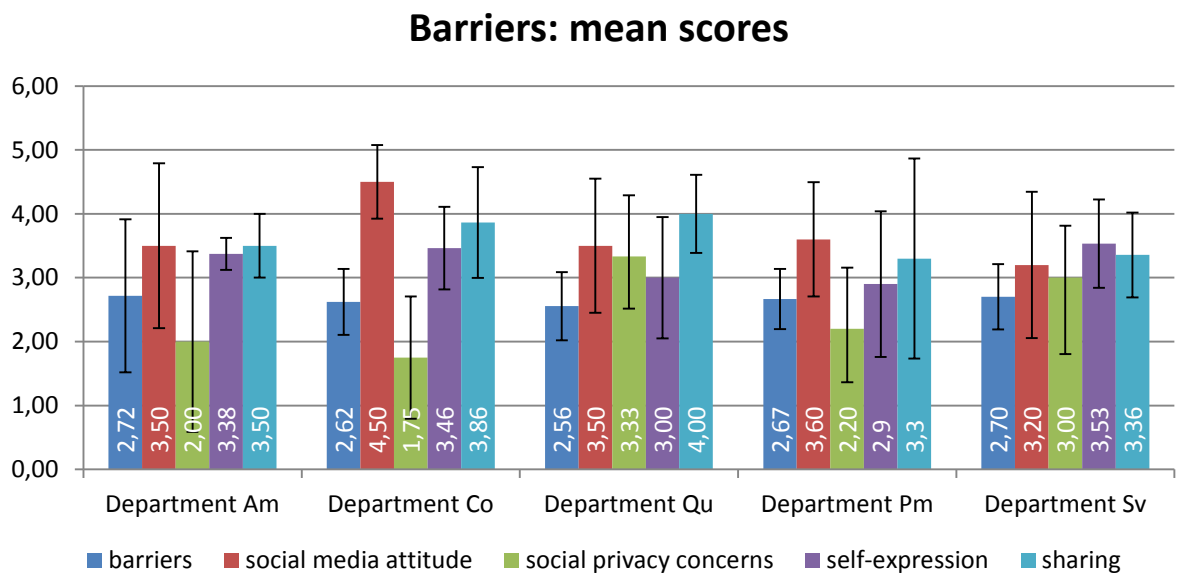

*Figure 5.4.4. Mean scores for possible barriers including general barriers, social attitude, social privacy concerns (*rated on a 4-point Likert scale), self-expression and sharing*

5.4.4.2 Level 2: Learning

Learning Process

App specific reflection questions

Altogether nine app-specific questions of the evaluation toolbox were included in the post-questionnaire, eight regular questions and one control question ($N = 34$). Figure 5.4.5 presents the mean ratings (from 5-point Likert scales) of the app-specific reflection questions and the corresponding control question of the whole evaluation and per department. The overall mean $M = 2.86$ ($SD = 0.80$) shows that the participants were neutral, that the application has potential to initiate reflection by capturing data relevant for reflection and visualizing data to reconstruct working experiences as well as capturing learning outcomes. While Department Pm disagreed that the application has potential to initiate reflection, the other four departments answered this question neutrally. These ratings are also very interesting with regard to the application usage and how many moods per participant per department were captured during the whole evaluation period. The results show that departments having captured less moods slightly agreed that the application has potential to trigger reflective learning, departments with higher app usage stated the opposite These results are also aligned with the insights gained from the interviews: especially the managers did not use the application for other purposes than for capturing moods. Therefore they could not exploit the capacity of the different visualizations regarding reflective learning about their own moods nor on the moods of their department. Furthermore, they did not include the application in their meetings, so in those situations the potential to trigger reflection was also missed.

The “app specific reflection control” question asked if the application can help to simulate their work process. With the summarized mean ratings $M = 2.82$ ($SD = 0.97$) the participants answered this question rather neutral. This rating is rather too high, because the application does no simulation of any work processes at all.

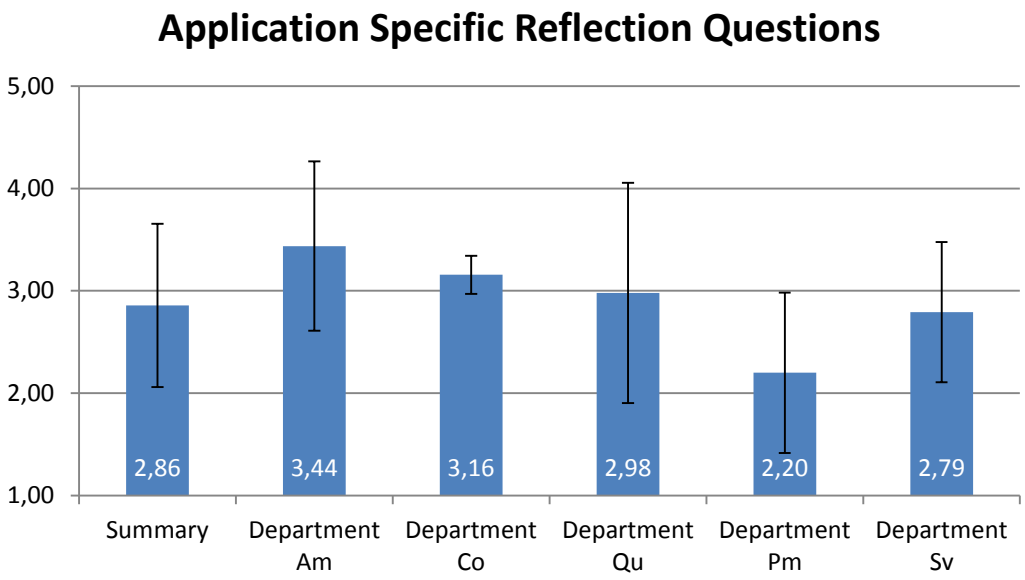

Figure 5.4.5. Application specific reflection questions for the whole evaluation and per department

Application specific questions

Figure 5.4.6 presents the mean scores of the applications specific questions (*N* = 34) contained in the post-questionnaire. We will only mention the values of the single departments, when they clearly differ from the average scores.

All eight items were on average rated neutrally (see *Figure 5.4.6*) covering mean values between 2.71 and 3.29. Department Co *M* = 4.50 (*SD* = 1.67) strongly agreed to become aware of their own mood (Figure 5.4.6, own mood awareness) and also stated to become more aware of their colleagues mood *M* = 4.00 (*SD* = 1.17) (Figure 5.4.6, colleagues mood awareness). Regarding the identification (Figure 5.4.6, identification) of significant situations, significant mood changes or topics worth reflecting upon was positively rated by Department Am *M* =.47 (*SD* = 0.42) and Department Co *M* = 3.59 (*SD* = 0.89) and negatively rated by Department Pm. In addition, only Department Am *M* = 3.50 (*SD* = 0.58) agreed that the MoodMap App guided them to reserve space for reflection (Figure 5.4.6, space for reflection).

Regarding the question if the MoodMap App helped the participants to gain a better understanding of the department and its members (Figure 5.4.6, department understanding) was agreed by Department Am *M* = 3.75 (*SD* = 0.50) but disagreed by Department Pm *M* = 2.00 (*SD* = 0.71). Only Department Am *M* = 3.58 (*SD* = 0.79) agreed that the app helped them to better understand their emotions, how their emotions affect their work or to deal better with their emotions (Figure 5.4.6, emotion understanding).That the capturing of moods during their daily work is useful to reflect and think about their work performance (Figure 5.4.6, reflect on work performance) was confirmed by Department Am *M* = 3.75 (*SD* = 0.50) and Department Co *M* = 3.75 (*SD* = 1.41).

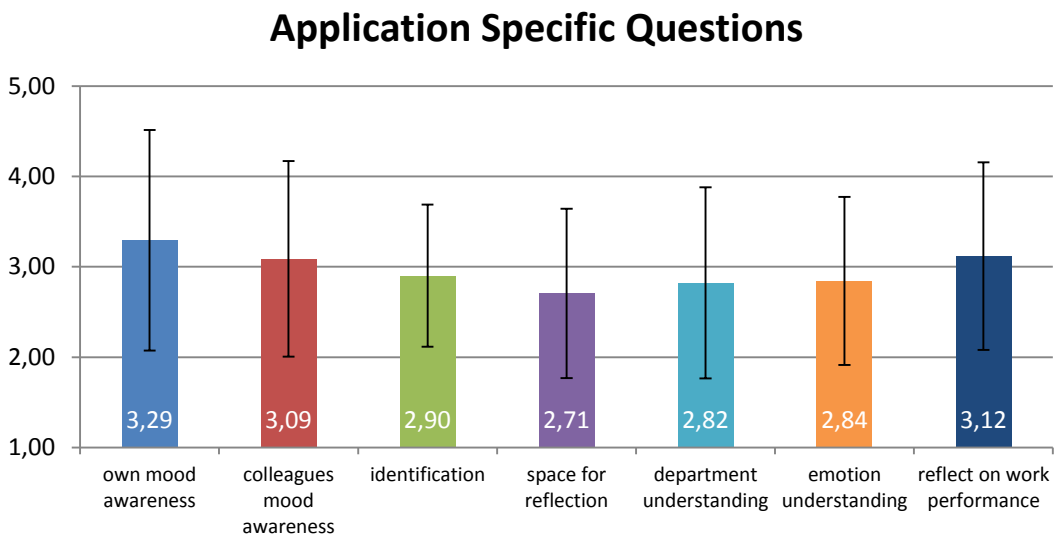

Figure 5.4.6. Mean scores of the application specific questions.

The participants were asked to capture their mood during the day. The usefulness of capturing the mood at the beginning of the day was rated with *M* = 3.50 (*SD* = 1.05), in the middle of the day with *M* = 3.62 (SD = 1.10) and at the end of the day with *M* = 3.59 (*SD* = 1.10). Regarding further situations in which it was useful to capture a mood, the responses range from during or after meetings until stressful situations, including "*moments of tension*" or "*during / after the execution of tasks that require concentration and accuracy*" and in phases where the user was feeling particular fatigue or tiredness.

Analysis of notes

The participants have captured all together 2250 moods and 225 non empty notes. After analysing the notes with the reflection-coding schema (section 2.2), 207 notes could be identified as individual reflective items. In the first round, two researchers categorized the notes independently. In a second round they analysed the notes together in order to get full accordance. Beside the categories defined in the reflection coding schema, a new category "Category 2ap" was added to the schema. It is a subcategory of "Category 2a" (own emotions) and represents the physical condition (e.g. pain, illness) related to own emotions. Table 5.4.1 shows examples of notes and the corresponding categories.

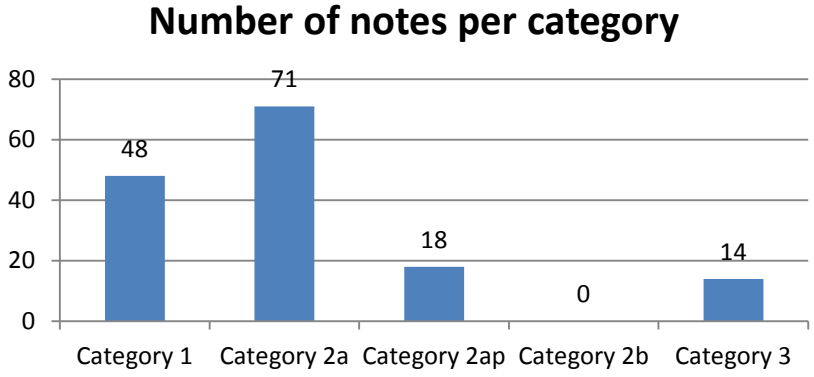

Figure 5.4.7. Number of notes per category according to the Reflection Coding Scheme

Table 5.4.1. Examples of categories and corresponding notes

Category	Example of notes
1: experience or issue	"Positive call with a customer"
2a, 2ap: own emotions	"I have a little bit of stomach ache, but I am feeling positive :-)"
1, 2a: experience and own emotions	"I feel better, but I just finished a call about analysis with the customer ..."
1, 3: interpretation of actions	"Emergencies that have distorted the work plan, overload of tasks and little fluidity (lack of feedback etc.)."

Short Reflection Scale (pre- and post)

Figure 5.4.8 represents the mean ratings for the Short Reflection Scale (SRS) as well as the two subscales concerning individual and team reflection. Comparing the scores on all three levels, we found no significant differences between the SRS of the ratings of pre-questionnaire to the ratings of the post-questionnaire. For the overall comparison we used the answers of those participants who have filled in both questionnaires on time and with meaningful content (*N* = 32). Figure 5.4.8 shows means and standard deviations for the SRS and the subscales individual reflection and team reflection in the pre- and post-questionnaire.

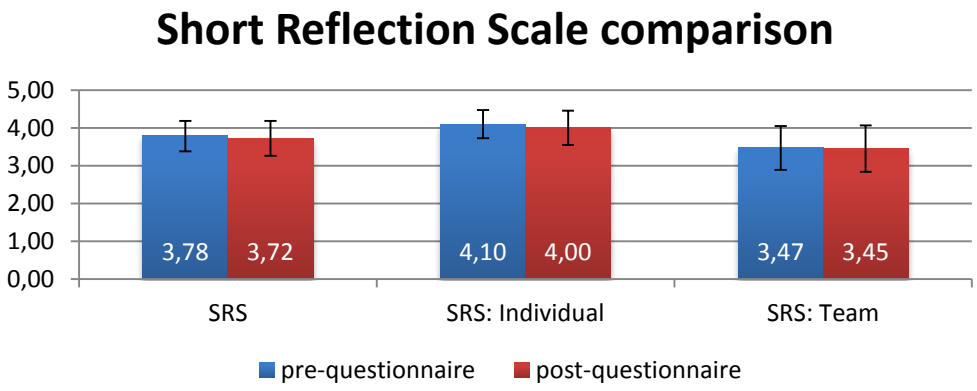

Figure 5.4.8. Short Reflection Scale before and after the usage of the MoodMap App

A two-way ANOVA with the factors time (pre- vs. post-questionnaire) and type of reflection (individual vs. team) revealed no effect of time but a significant effect of reflection type ($df = 1 (31)$, $F = 60.52$, $p < .001$). Post-hoc related t-tests show that individual reflection scores are significantly higher than team reflection scores at both times (pre/post-qu: $t(31)=6.2/5.7$, both $p < .001$).

Long term usage

Several variables from Level 1 and Level 2 correlate regarding the long-term usage in general. The long-term usage covers questions with regard to the long-term advantage of using the application (long-term 1), the wish to continue the usage of the app as part of the work-life (long-term 2) and to being practical to continue using the application during work-life (long-term 3). Sharing of moods with managers positively correlates with all long term variables which indicate that following moods over a longer period of time might positively influence the

collaboration between managers - and employees. Also the motivation to reflect on work in general as well as to deal better with emotions are increased by a longer usage of the App. Additionally recommending the App to others would increase if the App is used longer and more clear insights and benefits are gained on an individual level. These correlations confirm our Hypothesis 2 that these 4 factors influence the long-term usage of the MoodMap App. Table 5.4.2 presents the detailed correlations.

Table 5.4.2: Correlation between long-term usage and sharing, motivate to reflect, deal with emotions and loyalty metric (N = 34)

	long-term advantage (LT1)	**continuation of usage (LT2)**	**practicality during work-life (LT3)**
sharing	$r = .605$, $p < .001$	$r = .702$, $p < .001$	$r = .590$, $p = .001$
motivation to reflect	$r = .464$, $p = .006$	$r = .507$, $p = .002$	$r = .467$, $p = .005$
deal with emotions	$r = .647$, $p < .001$	$r = .566$, $p < .001$	$r = .516$, $p = .002$
loyalty metric	$r = .711$, $p < .001$	$r = .645$, $p < .001$	$r = .488$, $p = .003$

A moderate positive correlation between the user's satisfaction with the application and seeing the long-term advantage of using the application was also found, $r = .494$, $p = .003$, $N = 34$. With regard to the user's satisfaction with the application we also saw a strong positive correlation to the motivation to reflect on work in general ($r = .729$, $p < .001$, $N = 34$), to deal with emotions ($r = .592$, $p < .001$, $N = 34$) and the recommendation of the application to others ($r = .704$, $p < .001$, $N = 34$).

Learning Outcomes

The learning outcomes stated within the post-questionnaire ($N = 34$) were seen as nearly neutral to slightly disagree $M = 2.66$ ($SD = 0.95$), i.e. participants were not in agreement whether they gained a deeper understanding of their work life and what to change about their work behaviour with or without regard to their work within their department. While four of the departments mirror the average values Department Pm M = 2.10 (SD = 0.89) clearly stated that they did not perceive any deeper understanding of their work life or did not consciously made a decision about how to behave in the future.

The analysis of the open questions from the post-questionnaire provided little more details about their understanding of their work life and their conscious decisions of what to change in the future. Only 2 to 4 participants answered the open questions regarding their learning outcomes. Regarding the conscious decision of what to change in future one participant stated to "*focus/concentrate on a certain working outcome*" while a manager mentioned that he will "*give a greater attention to the mood of the staff in my team*". With respect to getting a deeper understanding of the own work life, one participant stated to take care about "*appropriate moments in which to do a break*". Other participants gained deeper understanding about the "*importance and influence of encouragement from the team of colleagues and especially the superiors (managers)"* and "*in the management of the business in relation to the moods, especially in times of stress*".

5.4.4.3 Level 3: Behaviour

Behavioural change (through question CB1) was rated by the participants with $M = 2.53$ *(SD* = 1.02*)*, which implies that the participants tend to slightly disagree that the MoodMap App

helped them to improve their work. While the four departments Am, Co, Qu, and Sv represent the average score, Department Pm M = 1.60 (*SD* = 0.89) strongly disagreed that the MoodMap App has helped them to improve their work. From the post questionnaire we only get confirmed that they have not noticed any improvements regarding their work.

A Pearson product-moment correlation coefficient was computed to assess the relationship between the variables from Level 2 and 3. Strong positive correlations (see Table 5.4.3) were found between making a conscious decisions of how to behave in future (CL1), gaining a deeper understanding of the work life (CL2), the improvement of work performance (CB1) and Level 1 variables like the user's satisfaction with the application, the long-term advantages, the motivation to reflect on work, better dealing with emotions and the recommendation of the application to a colleague (loyalty metric).

Table 5.4.3. Correlation between Level 2 and Level 3 variables (N = 34)

	satisfaction	**long-term advantage (LT1)**	**continue usage (LT2)**	**practicality during work-life (LT3)**	**reflect motivation**	**deal with emotions**	**loyalty metric**
CL1[5]	$r = .613^*$	$r = .619^*$	$r = .569^*$	$r = .487$ $p = .003$	$r = .665^*$	$r = .809^*$	$r = .659^*$
CL2[6]	$r = .524^*$	$r = .718^*$	$r = .737^*$	$r = .736^*$	$r = .669^*$	$r = .773^*$	$r = .608^*$
CB1[7]	$r = .685^*$	$r = .526^*$	$r = .572^*$	$r = .457^*$ $p = .007$	$r = .754^*$	$r = .775^*$	$r = .679^*$

Additionally to the feedback gathered with the questionnaires, gained insights and behavioural changes were also collected through several interviews. Seven interviews were conducted with the following participants: the manager of the Department Am, the manager and one of the two staff members (who coordinated this evaluation from Regola's side) from the Department Pm, two staff members from the Department Co, one staff member from the Department Am, and another staff member from the Department Sv. Below some representative statements of the conducted interviews are presented. The detailed results can be found in section 12.2 Interviews with participants.

Generally, the feedback from managers and staff was very positive, as it is shown for example in the following statement from a manager "…*interesting experience as a user because in a way it is my company that becomes interested in my work. They try to be aware of my goals during my work activities and I think that it is a really positive thing…*". Also statements from the staff showed the positive experience: "*Yes it is an positive experience … I think it was a useful tool to access this state of mind at any given time of the day and particularly during work and activity*" or *"I thought that the usage of the MoodMap App could help and support the entire company and team work".*

[5] CL1: After using the MoodMap App, I made a conscious decision about how to behave in the future.
*: Significant at $p < .001$
[6] CL2: After using the MoodMap App, I gained a deeper understanding of my work life.
[7] CB1: The MoodMap App helped me to improve my work performance.

An interview with one of the three persons who were responsible for introducing the MoodMap App at Regola gave us also very important insights regarding the preparation and conduction of such an evaluation. For him it was very difficult to convince the employees to use the MoodMap App on a regular base, but he still sees a clear benefit the MoodMap App could have for a company like Regola: "*Managers press you and they want you to finish in time. Usually they don't ask you if you are stressed, you are happy or in a good mood and low energy. I must say it is very positive that maybe someone now, I hope that our managers thinks about that also and not about only to finish the work and so. Because I think that if the staff is happy and working well the work is finished in time…"*

Also some of the participants perceived a clear benefit or insight for themselves e.g. "*to express my feelings and in general and at work - it is possible that the MoodMap App helped me in this regard for activities with my colleagues or my team activities in general."* or "*I am a person who is very angry after a meeting. I used to go happy to the meeting and feel angry afterwards. The MoodMap App allowed me to have a look - how can I switch my mood directly after the meeting".* On the other hand, limitations were also mentioned about how to get something relevant out of the captured data "*I collected a lot of data information and my trend. I had learned something more on my approach to the work. But then… or maybe I did not realise it or I am still missing of something that triggers and makes me shine on new perspectives."*

5.4.4.4 Level 4: Results

In order to detect differences between the situation before and after the usage period, we asked participants in both questionnaires (pre-test and post-test) about their job satisfaction and the impact of reflection in their job.

Regola could not provide KPIs from their organisation. Therefore, data regarding team specific KPIs were planned to be collected through the questionnaires, but the data from the post-questionnaire could not be obtained due to a technical problem. Consequently, a comparison with the KPIs collected before the evaluation period was not possible. Further alternatives to collect this data were not possible; therefore we focus our results on the job satisfaction, and the work improvement on an individual as well as team level, which were collected in the pre and post questionnaires.

Participants' rating about job satisfaction was improved during the app usage (see Job satisfaction in

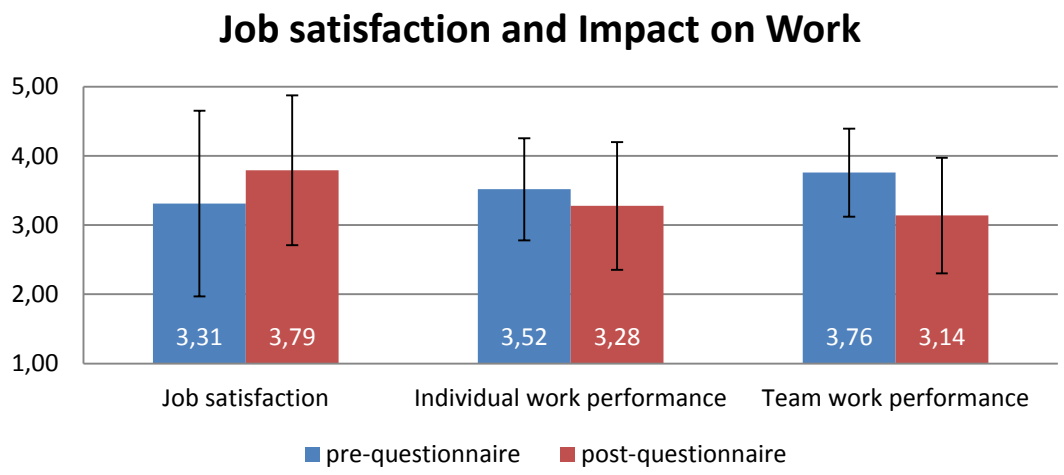

Figure 5.4.9). Participants were also asked whether reflecting on how they feel could both individually and collaboratively help them to improve their work performance. Expectation before the evaluation ("*Reflecting on how I feel could help me to improve my work performance.*") was compared with the actual experience during the evaluation ("*Reflecting on how I feel improved my work performance.*"). This comparison (see Individual work performance and Team work performance in Figure 5.4.9) shows that participants were slightly less confident after the evaluation regarding reflection on own individual work. This was also the case with reflection and discussion on their moods getting to improve their team performance.

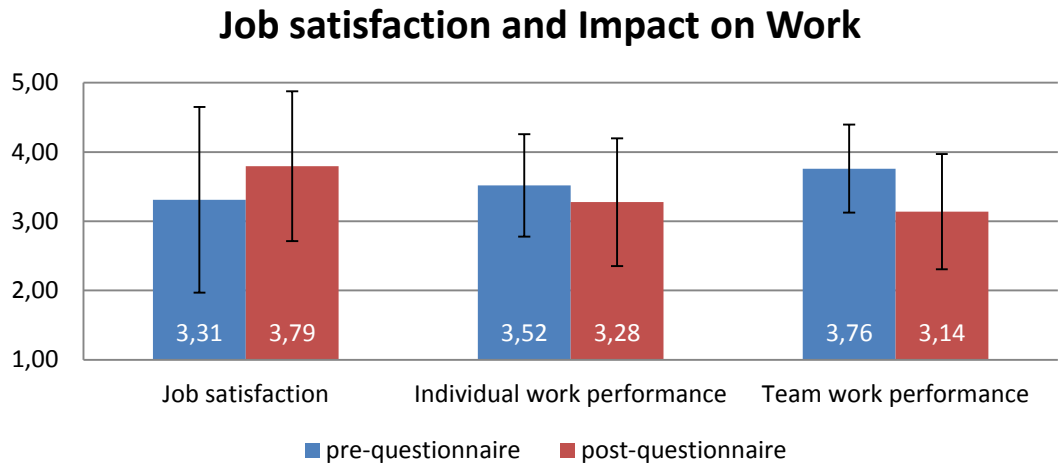

Figure 5.4.9. Job satisfaction and impact of reflection on work before and after the usage of the MoodMap App

As shown above, job satisfaction was on average improved after the evaluation. Therefore, job satisfaction at department level was analysed. Figure 5.4.10 below shows the average work satisfaction of each department, before and after the app usage. All departments except for Department Qu reported an improvement of their work satisfaction. Results show an increase in the average satisfaction value as well as a decrease of the standard deviation for each department (see Figure 5.4.10). Department Qu has shown a very low satisfaction both before (M = 2.25, SD = 1.71) and after the evaluation (M = 2.00, SD = 1.41).

Figure 5.4.10. Job Satisfaction of each department before and after the usage of the MoodMap App

Average rating to the question whether individual reflecting on how an individual feels, helps to improve the own work performance slightly varied among departments (see Figure 5.4.11). Whereas the agreement to this question decreased in 4 teams (Am, Co, Pm and Sv), in the case of the Department Qu it was increased. Details about the analysis of the answer given by each department are depicted below in Figure 5.4.11.

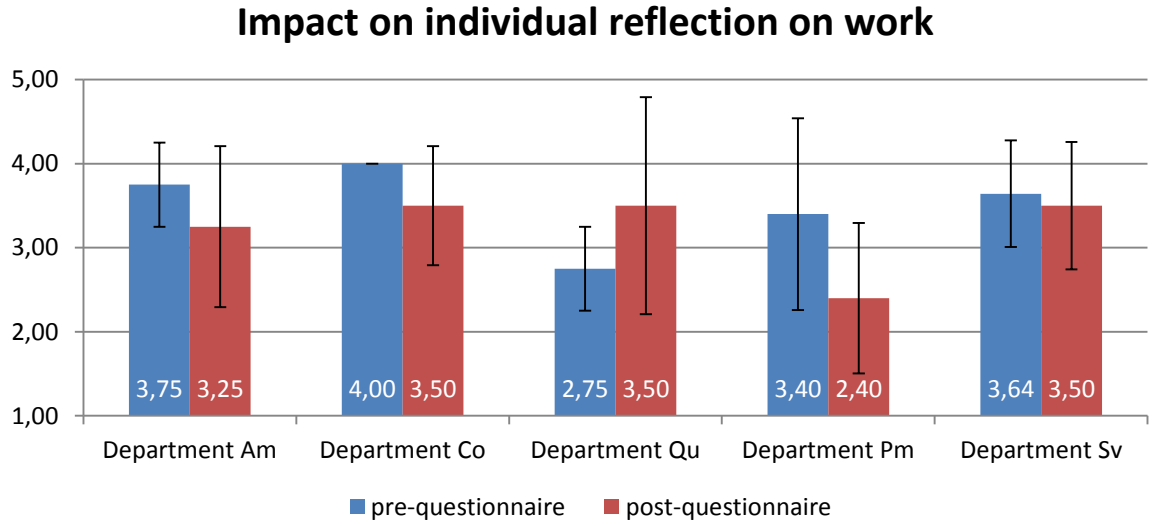

Figure 5.4.11. Impact on individual reflection on work improvement per department before and after the usage period

The results of the loyalty metrics (N = 34) looks as follows: 3% are promoters, 26% are passives and 69% are detractors. That implies a computed Net Promoter Score (NPS) of -68%. This low score reflects the difficulties unfortunately experienced at Regola to introduce the MoodMap App as a support for reflective learning.

A Pearson product-moment correlation coefficient was computed to assess the relationship between the variables from Level 4 and variables from other levels, which may have had an impact on the subjective KPIs. KPI 1 representing the satisfaction with own work, was not correlated with any of the variables investigated below, however, this was the case for the KPIs related to the impact of reflection on work (post-questionnaire answers).

There are positive correlations between KPI 2 (reflecting on how own feelings helps to improve work) and both, expression of feelings in general (r = .464, p = .007, N = 33) and expression of feelings at work (r = .684, p < .000, N = 33). This fact shows that there is a strong influence of self-expressiveness of the participants and their trust on reflection to improve their work performance.

A significant moderate positive correlation between two of the subjective post-questionnaire values of the KPIs and the average value of the app specific questions was found. The believe of reflecting on how one feels helps to improve work (KPI 2) has a moderate correlation with the app specific mean (r = .505, p = .003, N = 33). This is also the case for reflection and discussion on their moods getting to improve their team performance in relation to the app specific mean (r = .499, p = .003, N = 33).

By calculating the correlations between the KPIs and the variables from Levels 2 and 3 we can investigate the Hypothesis 3 that people who learned due to the usage of the MoodMap App and also changed their behaviour thanks to it also rated the KPIs with a higher value. This hypothesis was confirmed, as it can be shown in the following *Table 5.4.5*, as both KPIs improvement of individual work performance and team work performance correlate positively with questions of making a conscious decision about how to behave in future (CL1), of gaining a deeper understanding of once own work life (CL2) and of improving the own wor performance (CB1).

Table 5.4.4. Correlation between post-values of the KPIs and variables from Levels 2 and 3 (N = 33)

	CL1[8]	**CL2[9]**	**CB1[10]**
KPI2[11]	r = .488, p = .004	r = .360, p = .040	r = .588, p < .001
KPI3[12]	r = .456, p = .008	r = .407, p = .019	r = .583, p < .001

Strong and moderate correlations were also found with variables that had an influence on the app usage and the reflection on the data. These variables include the satisfaction with the MoodMap App (App satisfaction), the motivation to reflect on work in general (Motivation to reflect), and the belief, that the app helped them collaboratively to deal with their emotions (Dealing with emotions). Table 5.4.5 shows an overview of these correlations. Finally, the loyalty metric has a strong and moderate correlation respectively with the KPIs referring to the impact of reflection on work (see Table 5.4.5).

[8] CL1: After using the MoodMap App, I made a conscious decision about how to behave in the future.

[9] CL2: After using the MoodMap App, I gained a deeper understanding of my work life.

[10] CB1: The MoodMap App helped me to improve my work performance.

[11] KPI2: Reflecting on how I feel could help me to improve my work performance.

[12] KPI3: Reflecting and discussing on how we feel could improve our team performance.

Table 5.4.5. Correlation between post-values of the KPIs and app satisfaction, motivation to reflect, help to deal with emotions and loyalty metric (N = 33)

	App satisfaction	Motivation to reflect	Dealing with emotions	Loyalty metric
KPI2	$r = .698$, $p < .001$	$r = .480$, $p = .005$	$r = .452$, $p = .008$	$r = .721$, $p < .001$
KPI3	$r = .589$, $p < .001$	$r = .602$, $p < .001$	$r = .507$, $p = .003$	$r = .573$, $p < .001$

5.4.5 Conclusion & Discussion

The use of the MoodMap App was initially identified by the human resources manager of the company with the goal to provide a possibility for self-reflection, self-development and especially for stress detection during work. However, the evaluation did not reveal the goals expected for several reasons.

First of all, the whole summative evaluation was conducted in a time when Regola was in a very busy period (e.g. a lot of hard project deadlines fell into the trial period) which adds additional challenges to the introduction of a new application. The introduction of the evaluation was organized within a one hour lasting workshop for all employees together. After the evaluation, the project managers on site mentioned that it would have been much better to organize different meetings with each single department in order to give a more detailed and better suited introduction for the corresponding participants and their job roles, by analysing also their work processes and embedding the MoodMap App on them. Additionally they might have also given support on how to reflect on the data and how they could derive some benefits or insights for themselves. We see this as one of the major lessons learned. It is of crucial relevance to prepare the introduction of such an evaluation very well, to show all participants a clear benefit and accompany all departments during the whole evaluation process. This is necessary to guide them to include reflection in their working processes but also to support them to gain all benefits of the MoodMap App usage.

The employees of all departments have declared that their job is very stressful and that they currently have to meet many project deadlines. Nevertheless a high number of moods were captured during their working days, but they had nearly no time and/or space to reflect about the captured data. Therefore, on the one hand they did not see a real benefit for themselves but on the other hand it was not mentioned after the evaluation that the capturing of moods was seen as “waste of time”, as one participant stated in the expectations asked in the pre-questionnaire.

As part of the evaluation, very short meetings with the corresponding managers were planned in order to find out how mood or particular things according to the results of the MoodMap App can be improved. However, these meetings to conduct reflection sessions about the captured moods with their managers or within the team could not be scheduled. Such planned sessions with the project responsible would have shown the managers how to deal with the captured data and might have given them some insights regarding their whole team or single team managers. From the managers side the staff members did neither receive any feedback nor did the managers take any direct actions with regard to their team members. This could be also explained with the fact that the managers had no time to deal with the moods of their team members and to try out the whole functionality of the MoodMap App, because this in the end would have increased again their work load.

Another barrier, which did not make the evaluation as successful as we expected it to be, was that some participants were afraid to state their mood in such a tool honestly, as it may affect their working position in the company. This suggests that it is important to convince the participants that the goal of the application is not to use the data inserted against them, but to support them. Such a conviction is only possible if the managers as well as the organisation are committed to such principles. Another factor that may have influenced the usage of the application is the low number of members that some departments had. From the experience gained in other evaluations, bigger teams see an additional advantage when the manager can attend to each member individually and the MoodMap App could help to increase this awareness. However, in the case of Regola, having small teams may have led to a sense of controlling, instead of support and also in Department Sv (N=16), which is a bigger team, this result was not confirmed due to the lack of actions taken by the manager.

Regarding the MoodMap App features, the participants were very satisfied with it and they had no technical problems at all. They stated that the capturing of moods was very easy and they thought the application was complete. The favourite visualisation was the CompareMe view, because there they could directly compare their mood to the mood of the whole department at one glance. The reflection interventions and reflections amplifiers were sometimes ignored, but most of the times they were mentioned as very positive and useful.

Besides having captured 2250 moods, participants only introduced 225 notes, 207 of which were categorized as reflective items. The analysis revealed that only 14 of these notes achieved category 3, where the learner interprets or justifies an action. Therefore, the potential for reflection while capturing their moods, by immediately identifying the cause of it or relationship to work tasks, could not be totally exploited.

Summarizing the progress of the evaluation, participants were very active in capturing their moods but it seems that the purpose of this capturing was not understood. Time and space for reflection was not facilitated and reflection on the data was not achieved. As the usage data reveals, some users who were very active in capturing moods only visited the visualizations a few times, so they may have considered the capturing more as a reporting to the company instead of a way of reflecting on their work.

5.5 The Talk Reflect Evaluations at NBN, RNHA, and RBKC

The Talk Reflection App provides its users with means to document and self-assess experiences, reflect them individually and together with others and to write down outcomes from reflection.

5.5.1 Organisational context

Test bed organisation and the organisational unit

Summative evaluations of the Talk Reflection App were done in four different contexts, among which there are two testbeds of MIRROR (a ward at NBN and a care home related to RNHA) and two sites at the public administration of the London Royal Boroughs of Kensington and Chelsea (RBKC). They are explained below and summarized in Table 5.5.1:

The **first evaluation** (**E1**) took place at the Stroke Unit of the NBN testbed, which is described in section 3.4. The goal for using the Talk Reflection app in the ward was to complement the education of assistant physicians by learning about conversations with relatives of patients.

In the **second evaluation** (**E2**) we used the Talk Reflection in a British care home near Edingley, Nottingham. Although it is not a direct member part of the RNHA, it is a typical care home. Management of the home became interested in using the app when the RNHA project managers presented the app to the home. The goal of using the app in the home was to support staff in organising and conducting their work as well as in improving the service provided by the home, both with a focus on social interactions with residents, relatives and other parties

The **third and fourth evaluations** were done within the administration of the Royal Boroughs of Kensington and Chelsea (RBKC) in London, which is the public administration unit for the districts (“Boroughs”) of Kensington and Chelsea in London (see details on these testbeds in D6.4). Here, the Talk Reflection App was used to support interns in challenges they were facing in their new work in order to learn for future jobs (**E3**) and for two departments with similar duties that were about to be merged with the idea of reflecting on each other’s practices to support the merging process (**E4**).

Test users and their job roles

The evaluations were done with different amounts of participants ranging from 9 to 18 participants. It should be noted that Table 5.5.1 includes the total number of participants and that the number of active participants in the study was lower due to dropouts and users who did not return to the app after its introduction.

In E1 all participants were physicians at the Stroke Units. Among the nine physicians initially participating in the evaluation there were 6 assistant physicians, two seniors and one head physician. All participants perceived conversations with relatives as most stressful and demanding, as dealing with them needs experience, and as they were not prepared for such talks in their education. Therefore, learning about these conversations was attractive for them.

In E2 all participants were care staff. In addition the manager took part in the introduction of the tool and never used it after that – she was not included in the usage figures. They had asked to use the app to reflect on conversations with residents, who often behave unexpectedly, with relatives, who are concerned about residents and need to be informed about residents’ condition, and with third parties such as social workers and doctors, who need to be informed about residents.

Table 5.5.1. Summative evaluations of the Talk Reflection App.

	E1: Stroke Unit at NBN	**E2: Care home (RNHA context)**	**E3: Interns at RBKC**	**E4: RBKC department**
Domain / Purpose	Education in Hospital	Work organisation / Service improvement	Support for interns at work	Best practices, merging
Participants[13]	9	9	18	12
Duration (days)	42	50	51	80
Time	Jul-Aug 2013	Aug-Sep 2013	Sep-Oct 2013	Aug-Oct 2013

In E3 participants were interns working in different departments of the organization. The manager responsible for supporting these interns wanted them to be able to learn from their internship for future work. In addition from experience he was aware of the fact that the interns would be facing similar challenges but due to working in separate departments would not be able to discuss the challenges with each other. Therefore, interns were supposed to use the tool to learn how to interact with colleagues and members of the public professionally.

In E4 the manager responsible for two departments wanted to use the app to support the merging process of the departments. He had faced problems such as lacking understanding of processes in the respective other department as well as a need to improve service provided for members of the public. Thus the aim of using the tool was to share and reflect on practices of the respective other department to support the merging process as well as to exchange experiences on service conduction to improve services.

Identified need and potential for reflective learning

For all evaluations a manager or other responsible person had identified potential for improvement in different aspects of everyday work, as described above. There were some existing learning approaches in place, which according to managers and staff were not sufficient. For example, at NBN supervision was in place but could not be done often enough to deal with all difficulties arising, and at the department of RBKC meetings were held to discuss issues, but could not cover all aspects that needed to be dealt with. Therefore in all evaluations the need identified for support and the expectation coupled to using the Talk Reflection App was dealing with problems that could not be tackled sufficiently with procedures that were in place at the organisations.

Potential organizational impact

Evaluations 1, 2 and 4 were conducted within defined boundaries in the organisations (a ward, a care home and two departments). Therefore, organisational impact was not expected or intended primarily in the evaluations. However, in case of successful usage of the app in the evaluations we also expected key performance indicators to be affected (see Table 5.5.2). Therefore, applying the use of the app in each of these organisations on a larger scale (that is, in all wards of NBN, in multiple homes of the chain the care home of E2 belonged to, and

[13] The number of participants represents the number of distinct login names in the tool.

in all departments of RBKC) has a potential to create change on the organisational level (e.g., good practices for conversations at NBN or in the care home).

Evaluation 3 was conducted on a different scale concerning the organisational impact, as the interns were scattered throughout RBKC and as goals of using the Talk Reflection App were set from a manager responsible for interns from the whole organisation. Therefore, outcomes from reflection among the interns could be used to improve processes of supporting interns at the beginning and during their internship (e.g., by composing a good practice guide for interns). The fact that at the time of writing this deliverable the Talk Reflection App is being rolled out to a new group of interns supports this: from using the app with multiple groups and generations of interns common knowledge can be harvested to support interns.

Table 5.5.2. KPIs related to the evaluation contexts of the Talk Reflection App as described here.

Evaluation	**Indicators (taken from D1.5 for E1/2, given by managers and mapped on D1.5 general categories for E3/4)**
E1	Short-term: Employee satisfaction increases Long-term: Reduction of complaints, customer satisfaction increases
E2	Short-term: Employee satisfaction increases Long-term: Reduction in number and severity of incidents, decreased staff turnover, increase in happiness on residents
E3	Short-term: Employee satisfaction increases, increased quality of work, learning for later work (not in D1.5, specific to intern programme) Long-term: Decreased number of negative events
E4	Short-term: Compatibility with procedures and employee satisfaction increases Long-term: Increased client satisfaction, decreased number of stressful interactions with members of the public

Concerning Key Performance Indicators the evaluations of the Talk Reflection App can be related to different impacts (including KPIs) for the different organisations involved. It should be noted that they were conducted in a maximum length of 80 days; therefore we differentiated between short-term and long-term benefits to be taken from them – we expected to see changes in the former while we hoped to see changes in the latter. This is reflected in Table 5.5.2.

5.5.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

The Talk Reflection App provides its users with means to **document and self-assess experiences**, **reflect them individually and together** with others (by **creating comments and sharing** them) and **to write down outcomes** from reflection. Documented experiences are listed in the app and can be filtered by multiple options. In addition the app offers other

features such as voting for helpful comments and prompts to scaffold the reflection process. For a detailed description of the app the reader is directed to D6.4 (as well as D6.1-3 for earlier versions of the app).

The evaluations were done with versions of the Talk Reflection app that were slightly adapted to the organisational context and purpose the app was used in. While at NBN the core version of the app was used, for the care home we added a feature to relate experience reports to codes for residents and to filter the list of documents by resident. For RBKC we adapted the look and feel of the app to tools used internally at RBKC and we added a feature deleting content after a certain timeframe to ensure compliance to internal regulations. However, the basic process of reflective learning was the same in all cases: People followed the process of collaborative reflection as described in D6.1-4 (see also Figure 5.5.1) and they could use the app for individual as well as for collaborative reflection.

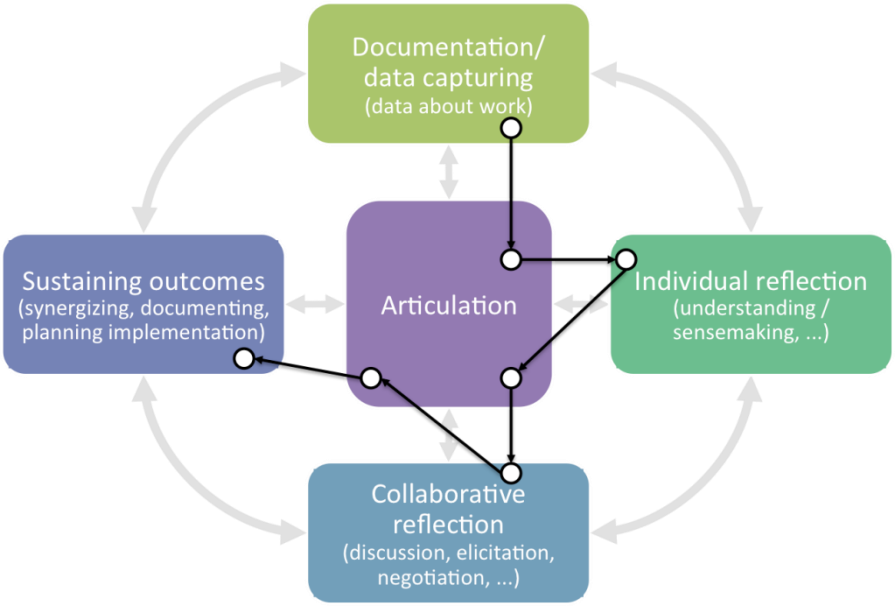

Figure 5.5.1. The cycle of collaborative reflection (Prilla, Degeling, and Herrmann 2012) describing the reflective learning approach implemented by the Talk Reflection App.

Relation to MIRROR CSRL Model

The Talk Reflection App covers both individual and collaborative cycles of reflection, which are shown in the tables 5.5.3 and 5.5.4 below. Table 5.5.8 below shows how (based on the evaluation data) the app supported the various stages and transitions.

Table 5.5.3. Stages and transitions in the CSRL model supported by the TalkReflection App for individual reflection.

Stage/*Transition*	Description
Plan and do work	Mainly for monitoring work: users capture experiences by describing them textually, making personal self-assessments (e.g., “how much does it bother me”) and adding them as documents within the app. The app supports them with this task by providing an easy to use interface with a simple text input box.

Data	The data mainly consist of textual descriptions of experiences from every day work practices and the self-assessments. This data may be enriched by adding comments as articulations of own reflection on the experience documented.
Initiate reflection	-
Frame	The app provides a space for documenting experiences.
Conduct reflection session	Commenting and reflecting on experiences made by oneself, creating topics from documented experiences to find synergies between experiences
Outcome	Creating Outcome descriptions in the app and relating them to documented experiences (own section for outcomes)
Apply Outcome	-
Change	-
Cycle transitions	From 'Apply Outcome' to 'Initiate Reflection': Outcomes are linked to documented experiences, next cycle of reflecting on experiences can be initiated.

Table 5.5.4. Stages and transitions in the CSRL model supported by the TalkReflection App for collaborative reflection.

Stage/Transition	**Description**
Plan and do work	See individual reflection, possibly multiple users can describe experiences together in the app
Data	See individual reflection
Initiate reflection	Share documentations with others and ask them to comment on those documentations, use the Talk Reflection App in meetings to find 'hot' topics to be reflected upon
Frame	Choice of recipients when sharing (frame for participants)
Conduct reflection session	Commenting and reflecting on experiences made by oneself or others, discuss experiences in the comments, rate useful comments, add documented experiences and discussion threads to topics
Outcome	Proposing outcomes / change in comments or creating outcome descriptions in the app and relating them to documented experiences (own section for outcomes)

Apply Outcome	Integration into meetings, voting on relevant or otherwise good comments in communication threads, creating topics from single discussions
Change	-
Cycle transitions	See individual reflection

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

The Talk Reflection App can be used for individual *and* collaborative reflection. The evaluation as described here was done mainly from a WP6 perspective and therefore mainly focused on collaborative reflection. Concerning the process perspective of reflection as described in (Prilla et al. 2012) and specifically the transitions and triggers from individual to collaborative reflection cycles (and within these levels) as described in D1.4b and (Krogstie et al. 2013; Prilla et al. 2013), the Talk Reflection App enables users to switch from individual to collaborative reflection by sharing data, to go from collaborative to individual reflection by looking at data in solitude and adding private data and from reflection in one group to reflection in another (or an adapted) group by changing the group of users the data is shared with. Concerning the outcomes perspective on reflection the Talk Reflection App may result in changes and ideas on individual (e.g., if an individual get insights on her work from a group discussion), group (e.g., if a group agrees on new norms for cooperation based on their reflection) or organisational (e.g., if a manager is involved an changes processes as a result of reflection) outcomes (see D 6.2 and 6.3).

Considering the transition model in Figure 10.3.1, the Talk Reflection App supports the documentation of experiences that caused a discrepancy (i.e. a trigger for reflection) and recursive reflection processes. The possibility to write down outcomes of reflection explicitly supports the application of outcomes by the group (if possible, step 3a) or by including superiors or experts (step 3b) – this implements the “push” and “pull” mechanisms of reflection as described in Prilla et al. (2012) and D 4.2/6.2/8.2.. For the implementation and sustainment of outcomes this documentation can be exported into the MIRROR spaces and made available for other tools such as task tracking tools and the like.

5.5.3 Research approach

Design and procedure

The evaluations of the Talk Reflection App were designed to run similarly as long as that was possible within the constraints of each organisation using the app. In each context, researchers of WP 6 ran an initial training session and – if possible – a closing workshop to get feedback. Initially, the app was introduced to the participants in on-site workshops including an introduction of the app and hands-on practicing. It was also discussed how the app could be used in the respective workplace, including how to use it in meetings or during the day. After this introduction, participants did not receive any further instructions to allow them to use the app in a way that would make most sense to them. In none of the cases a control group was established at the beginning of the evaluation. However, when analysing the questionnaires handed in we found several workers at the department among the participants of E4, who in the questionnaire indicated that they had not used the app after being introduced to it or only used it before workshops. This was the case for 7 participants (therefore splitting the group

into five active and seven less active users). We used this group for comparison to active users, as explained below. The duration of each evaluation varied (see Table 5.5.1).

Participants

The number of participants in each of the evaluations is shown in Table 5.5.1. It should be noted that this is the number of people for each evaluation who used the app in one more occasions. In each case some of them dropped out after the first days of using it or stopped using the app before the end of the evaluation study. Therefore, the number given in Table 5.5.1 describes the number of participants who left traces (data) in the app and the number of active users throughout the studies was lower. In addition, as can be seen from Table 5.5.5 the number of participants filling in the pre and post questionnaires as well as the participants in the final workshop was lower, too.

Table 5.5.5. Data collection in the evaluations of the Talk Reflection App.

	Pre-questionnaire	**Post-questionnaire**	**Data analysis**	**Workshop / interviews /**
E1 (NBN)	5 participants	5 participants	Usage data, content created	Workshop with 5 participants, 3 interviews
E2 (Care Home)	5 participants	- (not allowed)	Usage data (single user level), content created	Feedback workshop with 4 carers
E3 (Interns)	8 participants	- (not possible)	Usage data, content created	- (not possible)
E4 (RBKC)	7 participants	12 participants	Usage data, content created	7 users in focus group

Summative evaluation methods used

For data collection and analysis an approach of combined methods was used. Methods used included the pre and post questionnaires developed by WP 1 and adapted to the usage purpose in each evaluation (demografics, SRS, CA, CL, CB). In addition, all data created by the users of the tool was anonymized and analysed afterwards according to intensity of usage, productivity and content (see details for data analysis below). In addition, when closing the evaluation at least a workshop with users of the tool was held to get feedback from usage and first-hand impressions from users. In the case of NBN short interviews with users were conducted in addition, which was not possible in other cases. Table 5.5.5 gives an overview of tools used for data collection.

As can be seen E3 is an outlier in terms of tools used. This is because interns had short-term contracts and by the time of closing the evaluation the majority of interns who had used the app were no more available. Therefore no post-questionnaires could be given to them and time was not sufficient to arrange a closing workshop. To cope with this lack we talked to the manager of the interns to get information on how the app was used and whether the interns

had perceived it as useful – in addition, data analysis strongly suggests it was. Furthermore, post-questionnaires in E2 could not be given to the participants, as their manager was afraid they would take too much time.

Besides analysing the data from questionnaires and the feedback given by users, we analysed the data available from the app to get an impression how the app was used:

- Descriptive statistics of group application usage: We analysed how much content was created in total and on average to get a measure of group activity in the evaluations. Measures included the number of items created, the number of read events by each user and the length of communication threads created by users commenting.
- Statistics of individual activity: For E2 we analysed the activity of individual users, as from the data we became aware that usage was very heterogeneous.
- Social network analysis: To see the dynamics in usage and the role of different users we applied simple measures of social network analysis[14] to all cases. Key measures were the graph density (the ratio between the number of edges in the social network graph compared to the possible total number of edges; describes how many people communicated with each other compared to how many people could have communicated) and the unique edge ratio (ratio of unique edges compared to possible total number of edges; high value means a lot of single communication acts between people have been established).
- Content analysis: Using a coding scheme developed by WP 1 and 6 for analysis of (general and more specifically collaborative) reflection content, (see D6.4) we analysed the content created by users in each evaluation. Two coders did the coding independently. To measure the coder agreement we used Krippendorf's Alpha. Details will be given in the respective evaluation section below.

5.5.4 Results

5.5.4.1 Level 1: Reaction (Usage)

For the analysis of usage of the Talk Reflection App in the different evaluations we computed the amount of experiences documented in the app (referred to as "reports" in *Table 5.5.6*) and the number of comments created on them as items of articulation of collaborative reflection. Analysing these figures we can see that the adoption of the Talk Reflection App was sufficient in all evaluation, especially given that participants in all organisations told us that experiences they wanted to reflect on were bothersome, but did not happen every day: Each group created about 0.5 reports[15] per day, wrote 0.5 to 1 comment per day (with 0.5 as a low outlier for E2) and participants answered at least 71% of the reports shared with them. However, the figures also show that there was also room for intensifying usage in each of the cases. In particular, none of the participant groups made significant usage of the feature for documenting outcomes (therefore not mentioned in the table). In addition, we had expected more comments to be created.

[14] The social network analysis was done using NodeXL, a Microsoft Excel based tool (https://nodexl.codeplex.com/).

Table 5.5.6. Usage figures for the Talk Reflection App in the evaluations. Grey shades mark the respective low value(s) in a row, black background marks high value(s).

	E1 (NBN)	E2 (Care)	E3 (Int.)	E4 (RBKC)
Users /days	9/42	9/50	18/51	12/80
Reports	25	15	24	45
Reports per day	0.57	0.48	0.47	0.56
Reports per user and day	0.06	0.05	0.03	0.05
Comments	39	25	47	65
Answer ratio[16]	0.88	0.80	0.83	0.71
Comments per day	0.93	0.50	0.92	0.81
Comments per user and day	0.10	0.06	0.05	0.07
Average thread length[17]	1.77	2.08	2.35	2.03
Graph density	0.10	0.17	0.19	0.25
Unique edge ratio	0.62	0.18	0.21	0.60
Reports read	153	144	284	421
Reports read per user	17	16	15.78	35.08
Reports read per day	3.64	2.88	5.57	5.26

The room for improvement of usage intensity can be attributed to different constraints in the organisations the app was evaluated in. In E1 and E2 the groups were collocated, working at the same time on the same floor. Therefore, as participants told us, they were used to talk about issues personally and thus often did not use the app in favour of such personal communication. In these cases the app was used as a memory aid and way to send notes when personal communication was not possible. In E4, in which the two groups working together were located at different places, participants told us in a focus group workshop that their manager taken the lead in encouraging reflection and using the app, and that he had also focused communication on those reports the manager was interested in. Therefore they had often abstained from commenting on other reports. E3, in which participants were scattered

[16] By answer ratio we refer to the percentage of experience reports that received at least one comment.
[17] As explained above, the average length was calculated only from the number of comments, leaving out the experience report being the root of each thread.

across the administration departments, created least comments per user and day, while in E1 0.1 comments per user and day were created.

Besides time constraints and the influence of location (that is, whether participants were co-located in a way that made personal communication possible) a major barrier and success factor could be found in management or other lead user support. In E1 and E4 a lead user was present (in E1 it was the head physicians and in E4 it was the manager responsible for merging the departments), who was the driving actor in each evaluation. This led to a good uptake in terms of productivity in both cases (see usage figures in Table 5.5.6). In contrast, such a user was not present in evaluations 2 and 3, which for E2 resulted in less usage as documented by 0.5 comments per day on average as opposed to about 1 comment per day in the other cases. The situation was made worse in E2 by the refusal of the manager to actively take part in reflection. In contrast to this, usage in E3 was high despite the lack of a leading user. This can be attributed to the fact that the Talk Reflection app added value on its own by connecting interns who would not have been able to communicate and reflect together otherwise. Therefore, one conclusion we propose to take away from the summative evaluations on uptake of reflection apps is to **have a leading user to drive reflection and sustain app usage in cases in which the app is used in co-located groups**.

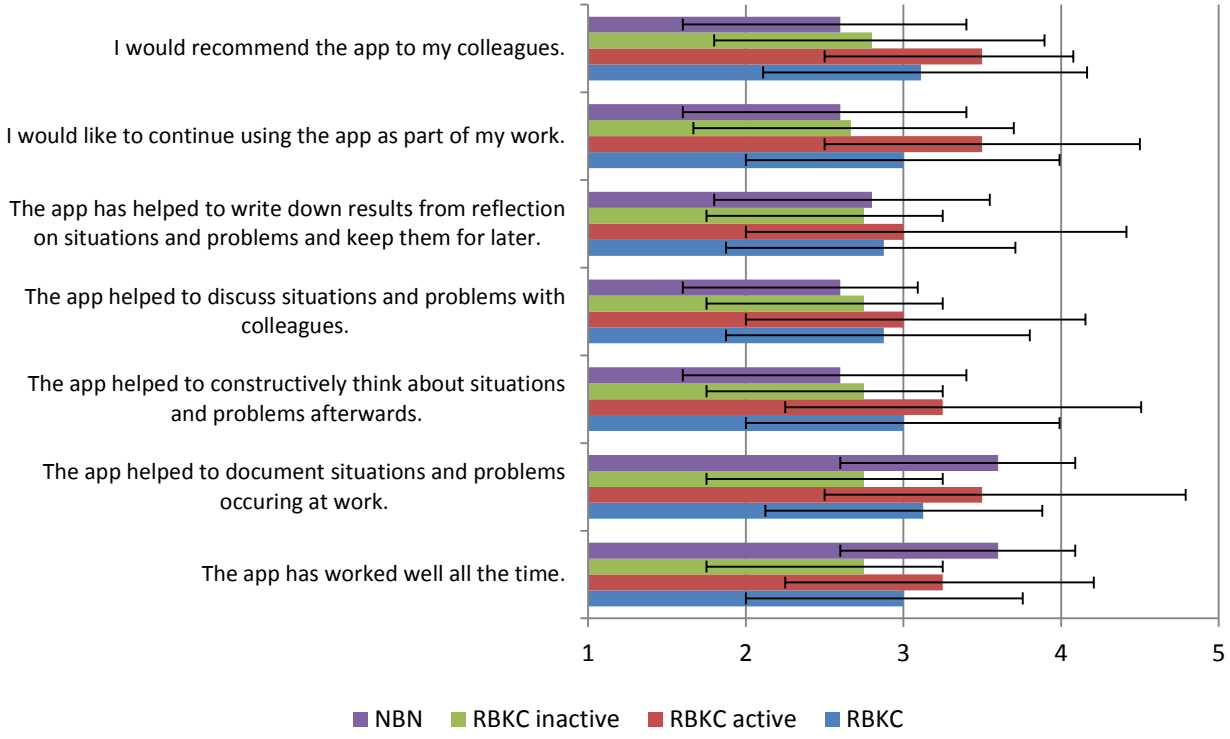

Figure 5.5.2. Reaction and Learning related items from RBKC and NBN post questionnaires.

Despite the increase in production the leading users in evaluations 1 and 4 can be credited for, the feedback we got from users in workshops and interviews (E1) differed. In E1 most participants told us that they did not perceive much benefit by the app. The main reason given was that because of the small ward all physicians knew most information on stressful experiences anyway, and that therefore the app was mainly used to document experiences, which were then discussed personally. The post questionnaire results for NBN shown in Figure 5.5.2 underpin this, as we can see scores below 3 (average) for the satisfaction with the app (e.g., “I would recommend the app to my colleagues”) but a score of 3.8 for a question on

support given by the app to document and discuss experience. In contrast, participants of E4 told us that they liked the app as it enabled them to reflect on procedures of the two departments and to exchange perspectives. This is mirrored in the more positive attitude towards the app in the questionnaire that goes along with similar values on utility as shown in Figure 5.5.2, and it becomes even clearer when separating active users from inactive users (the participants who did not use the app after the introduction or only used it before workshops, as explained above). Looking into how the app was used, we could attribute this difference in perception of its value mainly to the way it was used: In E1 the leading user had been active mainly by answering each report created by an assistant physician right away, while in E4 the leading user had created reports to reflect about, posed questions to others and answered on reports shared with him. While both leads to a focus on the leading user, the former way had caused shorter communication threads and stopped assistant physicians from commenting – what should they write after the head physician had given his opinion? In contrast, the way the leading user in E4 acted had created discussion threads rather than finalizing them. Thus we may conclude that **a leading user needs to facilitate reflection by being part of the reflection process and actively fostering it**.

5.5.4.2 Level 2: Learning

Learning was expected to take place in all evaluations, with slight differences in the topics (focus on conversations as in E1 and more diverse topics in the other evaluations) and amount, depending on the participants and the duration of the evaluations (it was expected to see slightly more traces of learning in 80 days of E4 than in 42 days of E1 etc.). In addition we expected the app to cover all stages of the collaborative learning cycles shown in Figure 5.5.1 as well as most stages (except the 'apply outcome' stage) and all transitions of the CSRL model.

As a baseline for the evaluations and an assessment of (perceived) reflection practice in all organisations we evaluated the Talk Reflection App in we used the Short Reflection Scale (CR). As can be seen from Figure 5.5.3, participants from RNHA (E2) and RBKC (E4) show the highest scores for reflection practice, while participants from NBN (E1) seem to feel they reflect less. Interns from RBKC scored their individual reflection practice high and their group practice very low.

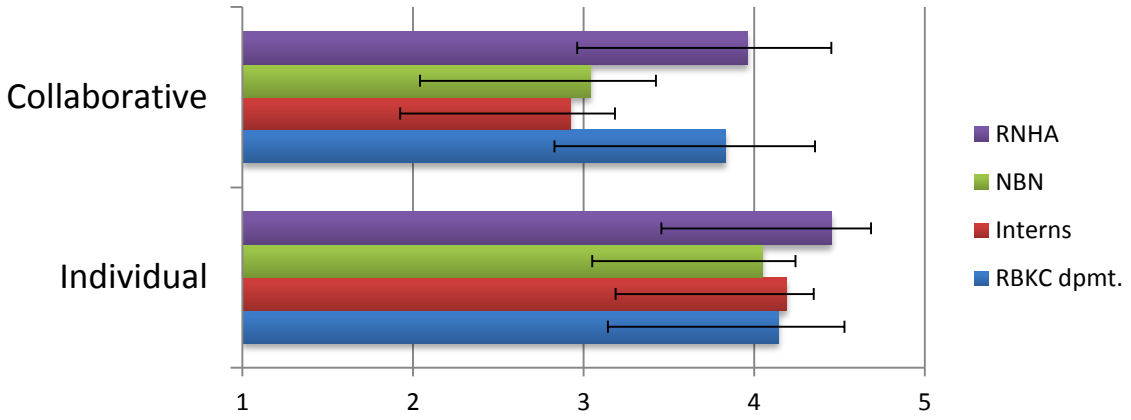

Figure 5.5.3. Answers to short reflection scale for post questionnaires in all evaluations, grouped by individual and collaborative reflection questions.

The Short Reflection Scale, for which we can compare Pre and Post data from E1 (NBN) and E4 (RBKC) shows minor differences in the perception on reflection of the participants (Figure 5.5.4), which due to the small sample sizes (see Table 5.5.5) cannot be interpreted.

Larger differences can be seen for RBKC in the importance of group reflection (e.g., *"It is important to me to discuss frequently with others about stressful situations like conversations"*; pre 4.4, post 3.3) and in the need to improve work *("I think it is important to try to improve handling stressful situations like conversations"*; pre M=4.7, SD=0.82, post M=4.0, SD=1.06). In general, the tendency of answers was negative when comparing post answers to answers at the beginning of the evaluation. This does neither fit the positive feedback we got (see section 5.5.4.1) nor the good participation in discussions (see *Table 5.5.6*). Therefore, it might be explained with a change in awareness for what reflection is and how the actual practice of it was at RBKC: By using the app and reflecting with it people might have recognized that there is a need to reflect more (efficient), which made them give slightly worse answers after the evaluation (we experienced a similar effect in an early evaluation of the Talk Reflection App at NBN, see D6.2/10.2).

For NBN, large differences can be seen in the assessment of reflection in meetings (*"In team meetings we frequently talk about how we can improve handling stressful situations like conversations"*; pre M=3.2, SD=0.75, post M=2.4, SD=0.49) and in the benefits gained from reflection (*"Reflecting on stressful situations like conversations helps me improve handling these situations"*; pre M=4.2, SD=0.4, post M=3.4, SD=0.49). The only larger positive change can be seen in the individual tendency to reflect ("*I often reflect on my work in order to improve it*"; pre M=4.0, SD=0.63, post M=4.4, SD=0.49). The negative changes reflect the feedback we got on app usage well and can be explained by the shortcomings in conducting and guiding reflection within the participants group as explained above. In addition, as mentioned above and recognized earlier (see D6.2 and 10.2) by using the TalkReflection app the participants may have recognized that they were not reflecting enough individually and in the group during their daily work, resulting in lower values. On the positive side, despite the rather negative impression people reported from using the app we can see an increase in the personal tendency to reflect.

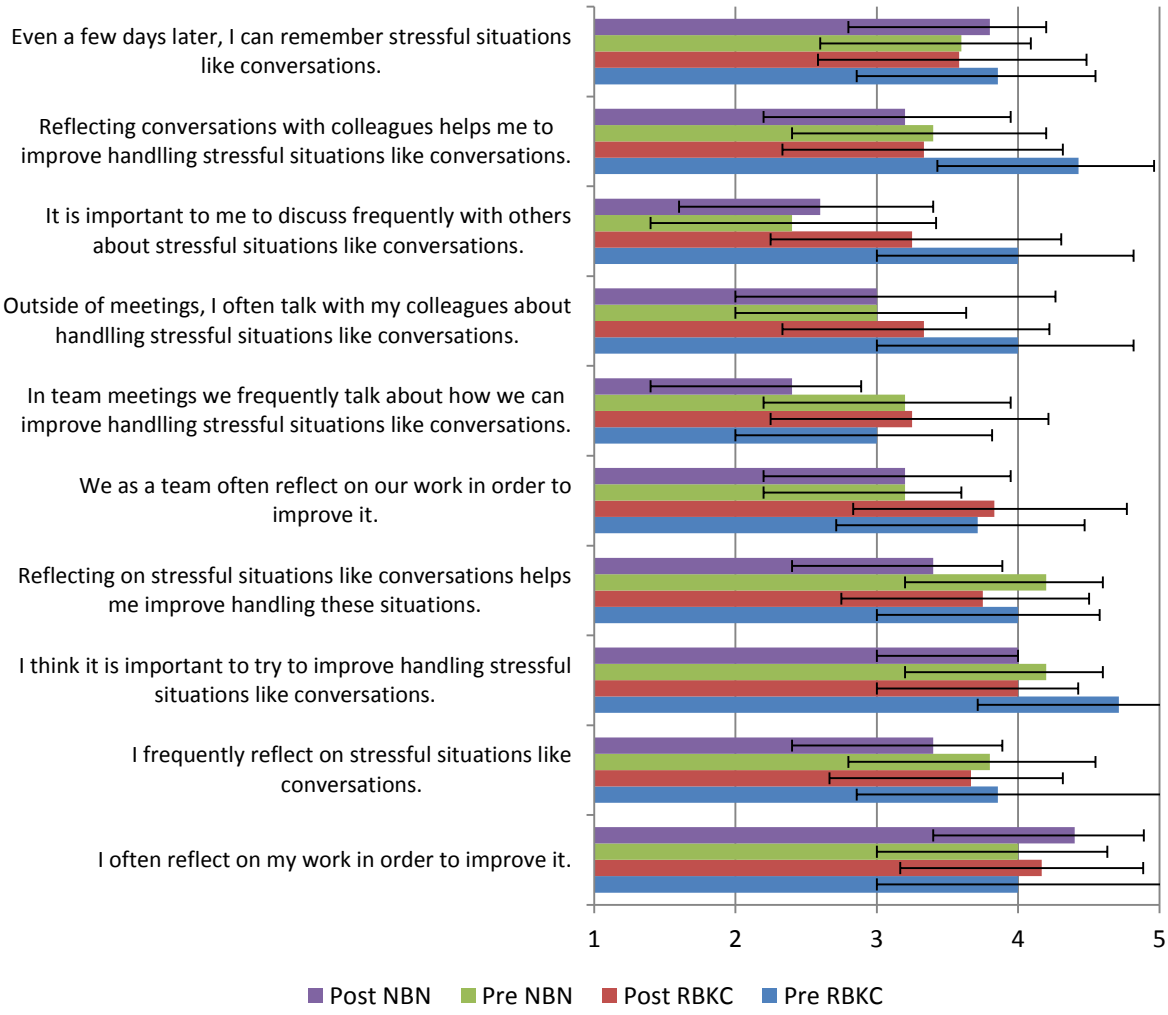

Figure 5.5.4. Comparison of the Short Reflection Scale Pre and Post results for E1 (NBN) and E4 (RBKC).

While comparing values from the Short Reflection Scale suggest that using the app did not have a positive effect on reflective learning, the analysis of other data shows such effects, as will be explained below.

Learning Process

The coverage and success of supporting stages in the WP 6 collaborative and MIRROR CSRL learning processes was measured in two ways. First, we added items to the post evaluation questionnaire that asked for participants' impression on support given by the app for different stages. Second, together with WP 1 we developed and applied a coding scheme to analyse reflective content in apps (see D6.4 for information on the development of the scheme) and used it to analyse whether the app was used to document only or to reflect and exchange perspectives. In addition, we used the feedback from participants to complement these measures and to derive insights in the cases in which the questionnaire could not been run after the valuation period.

Feedback in the cases differed, with participants from E1 and E2 reporting low learning effects, and participants from E3 and E4 reporting that they had learned by using the app. Participants in E1, due to the reasons described above, reported that they had used the app mainly to document experiences in order to keep them in mind. Reflection, according to them happened

mainly in direct interaction with others. Similarly, participants of E2 reported that they had not used the app often to exchange comments but to keep memories of issues they wanted to reflect upon. As described above the main reason for this was that they did not receive support for reflection during the day from their manager and thus felt they could not spend much time with the app. In E3 and E4 feedback was more positive and both interns (E3) and workers from the two departments (E4) reported they had used the apps to share reports and to discuss them. In addition we were told in both cases that the participants had taken some outcomes with them from reflection.

The answers to the learning process related items in the questionnaires run at E1 (NBN) and E4 (RBKC) underpin this feedback. As Figure 5.5.2 shows, participants of E1 perceived the app as supportive for reporting experiences ("*The app helped to document situations and problems occuring at work.")*, even valuing it slightly more than (active) participants of E4. This changes when it comes to support for reflecting individually and collaboratively (*"The app helped to constructively think about situations and problems afterwards" / "The app helped to discuss situations and problems with colleagues"*). Here, scores from participants of E4 are higher than from E1. Scores for sustaining outcomes are similar.

The content analysis supports the impression to be taken away from the feedback and the results of the questionnaires. As the values for Krippendorff's Alpha varied too much for individual codes (ranging from acceptable values of 0.75 and slightly below to values lower than 0.5), we aggregated codes to three levels of reflective content (see D6.4 for more details on this aggregation). When calculating the interrater agreement on these levels, we arrived at good agreement values (97% for stage 1, 96% for stage 2, 80% for stage 3). Although the level of details is lower for these stages compared to the coding scheme, the quality of the resulting data is better. To further enhance the quality of the coding for stage 3, the coders discussed differences in using codes 8a, 8b and 9, resulting in a coding agreed upon the two coders, thus improving the agreement on level 3 to 100%. This data was used for the following analysis (Prilla and Renner 2014).

Table 5.5.7 shows the results of this aggregation: For E1 and E2, for which we received less positive feedback, we can see the lowest results for traces of learning in the content – for E1 we did not find any such traces at all. It is surprising that for E1 we found 95.2% of the conversations to contain reflective content, although the feedback of participants indicates that the app was used mainly for documentation purposes. This shows that potential was lost in E1, as even marginal perceived value of using the app resulted in traces of reflection in the app, which was then not taken up by its users. For E3 the value is higher and in E4 one third of all conversation threads included traces of learning. It should be noted that this only relates to content in the Talk Reflection App and does not allow to make conclusions on the learning that took place in E1-4 in general. However, it shows that users in E3 and E4 arrived more often at learning outcomes by using the app than users in E1 and E2.

Table 5.5.7. Results of the analysis of reflective content in the evaluations.

	E1 (NBN)	**E2 (Care)**	**E3 (interns)**	**E4 (RBKC)**
# Conversations analysed	21	12	17	24
Stage 1: **Provision and description of experience**, but no (explicitly) traces of reflection	100 %	100 %	100 %	95.8 %
Stage 2: **Reflection on experiences**, including analysis and potential solutions, but no (explicit) mentioning of learning or change	95.2 %	83.3 %	94.1 %	95.8 %
Stage 3: **Learning or change** resulting from reflection explicitly mentioned	0 %	16.7 %	23.5 %	33.3 %

Based on these results we can conclude that support given by the Talk Reflection differed in the evaluations, with basic support for documentation and sharing in all evaluations and additional support for individual and collaborative reflection in E2, E3 and E4. A mapping to the CSRL model stages and transitions shows a similar result (see Table 5.5.8).

Table 5.5.8. Support for the CSRL model stages in the evaluations of the Talk Reflection App.

	E1	**E2**	**E3**	**E4**
Plan and do work Transition: Data	++ *Reports*	++ *Reports*	++ *Reports*	++ *Reports*
Decision to reflect Transition: Frame	+ *Sharing*	+ *Sharing*	++ *Sharing / reading*	++ *Sharing / reading*
Conduct reflection session Transition: Outcome	- *(Direct interaction)*	+ *Comments*	++ *Comments*	+ *Comments*
Apply Outcomes Transition: Change	-	*(-)*	+ *Comments*	++ *Comments*

Learning Outcomes

As a direct measure of perceived learning outcomes we can use the questionnaires filled in by participants of E1 and E4. As the feedback received from participants of these evaluations on learning outcomes indicated, we can see differences in the learning specific items of the questionnaire (see Figure 5.5.5): RBKC participants in general rated learning effects to be higher than NBN participants, and active participants from E4 were most positive about this. It has to be mentioned, however, that most values on learning effects are below 3 (with the exception of active RBKC participant), which indicates low overall perception of learning in both evaluations.

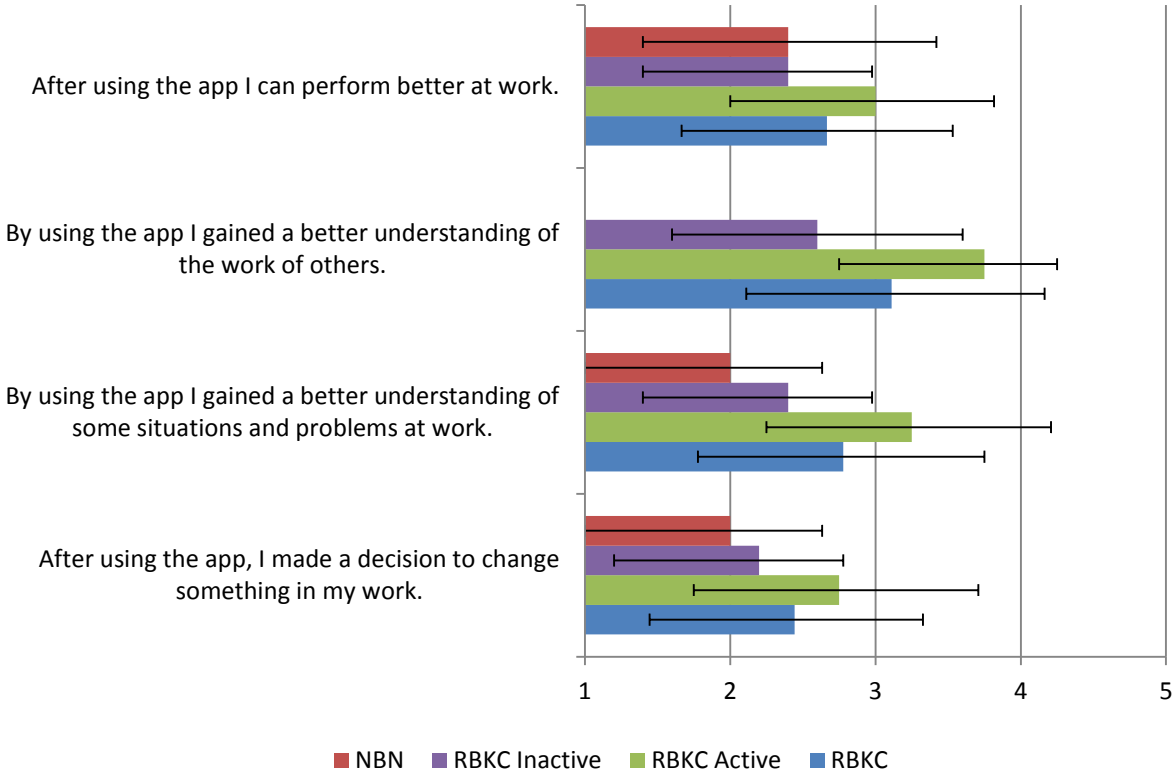

Figure 5.5.5. Learning items in the post questionnaires for NBN (E1) and RBKC (E4).

These results do not fit the results of content analysis as shown in Table 5.5.7: From this analysis we can see that in all evaluations people frequently engaged in reflective activities and that at least in E3 and E4 (as well as partly in E2) a good percentage of discussions actually contained learning outcomes documented by the users. This is backed by feedback we got from participants and managers of these evaluations, who valued the usage of the app concerning the learning effects it had.

We may explain this difference in figures and qualitative data describing learning effects by the lack of systematically deriving and converging outcomes from the usage of the app in all cases as well as by the lack of organisational integration and process support in E1 and E2. We could see learning effects documented in comments used in one third of all conversation threads in E4, but answers of items in the questionnaire indicate that participants rated effects low to medium (with the exception of active users, who were more positive). This suggests that some participants were not aware of such outcomes or that at least they had not related them explicitly to their work when the questionnaire was run. In contrast, in content from E1 we did not find such content and therefore the answers to the items shown in Figure 5.5.5 do not come as a surprise. As a result of this observation we first suggest to better **support making explicit outcomes in a way that people become aware of them**. Second, we suggest to **more explicitly focus on the connection of work and goals to reflection outcomes** in reflection support tools.

5.5.4.3 Level 3: Behaviour

Feedback on behaviour change was medium to positive in most evaluations both in items and verbal feedback. In line with the feedback received before, participants of E1 rated the impact on behaviour rather negative, which fits the feedback as described below. Participants of E2

saw potential in deriving changes in the work from the app but reported that because of lacking support and possibilities to use the app during their work they had not tapped from this potential sufficiently. Consequently, in E1 we found no traces of learning or behaviour change in the content created in the Talk Reflection App and for E2 we found the second lowest value for this in the evaluations. For E2, however, we also were told about positive effects of the app, in which it had triggered discussions on certain issues as we had already seen in an earlier evaluation of the app in the care home (see D6.3 and D10.2).

In E3 the manager responsible for the interns told us that based on the feedback he got the interns took away some lessons from reflecting together and by adapting certain work practices actually felt they could do a better job afterwards – there was no chance to get direct feedback from participants to prove this. However, we found content in the app such as for example "*The list of FAQ's/useful contacts is actually really useful - I can't think of the amount of times I have picked up a call, then wasted time trying to find the correct procedure - will definitely be using that."*, which includes intentions to change the way they would work (in the example it refers to a report describing difficulties in handling different requests by members of the public). In addition we were told (and could see traces in the content) that the interns group had organised meetings among them to further discuss certain issues based on the experience they made with the app. Further observations on behaviour change could not be made – the main problem in observing mid or long-term behaviour change among participants of E3 is that (as described above) interns have short-term contracts and therefore change may affect work after their internship, in which they usually have a new employer and are not available for gathering data. In practice, the fact that most interns left RBKC after they used the app even hindered us from running a post questionnaire.

For E4 we received positive feedback from a group of participants in a focus group run after the evaluation period that they had used issues reflected upon in the app and ideas stemming from this to restructure their procedures. In addition they told us that they used the app to prepare their meetings to make sure the most pressuring issues would be discussed face-to-face. In line with this, we found several traces of change or intentions to change in the content created by the participants such as *"It can be really frustrating not getting a response to a question as it holds you up and just causes more work. I do try and remember to cc you guys in if I am asked to respond"* as a resulting comment from a reflection thread on why people would not answer emails sent to them. This not only changed the discussion culture in the existing departments but also, as the manager of the group later on told us, did the app fulfil the function of *"establishing communication links between the departments"*, thus changing the way people communicated with each other.

These rather negative impressions on behaviour change in E1 and the positive aspects from E4 are mirrored by the items in the post questionnaires run at E1 and E4 (see Figure 5.5.6). The upper four items shown Figure 5.5.6 not only ask for insights but also for effects on work practice. It can be seen that participants from NBN (E1) scored each of these items well below average (3), indicating that there was little impact on their behaviour by using the app. While the average scores of participants from RBKC are already higher, especially the scores for the active participants in E4 show that in this group the app had a perceived effect on behaviour. In E4 users profited for especially in situations where cooperation is important, whereas ratings are also below average for items concerning the individual.

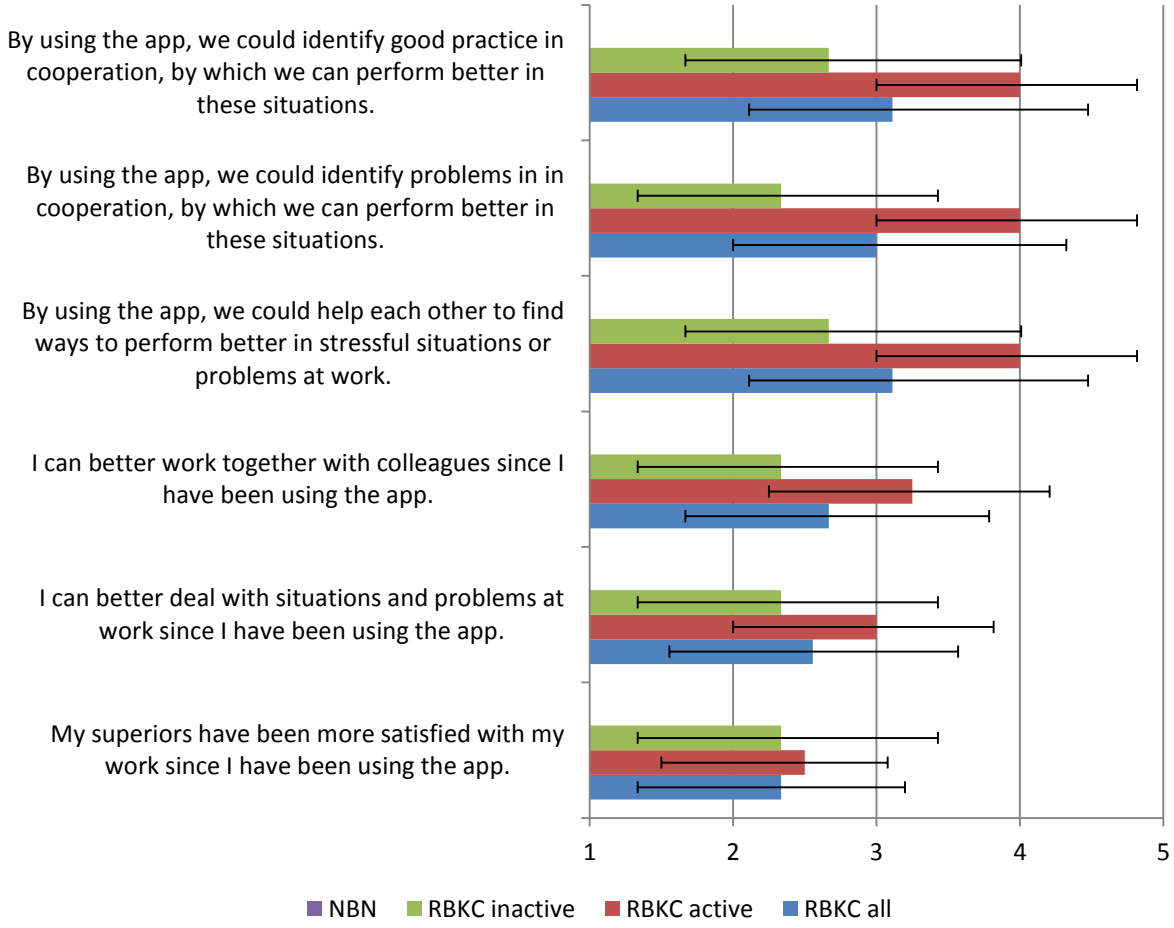

Figure 5.5.6. Post questionnaire items related to changes in behaviour (upper four items) and KPIs defined by the organisations (lower two items) for E1 (NBN) and E4 (RBKC).

5.5.4.4 Level 4: Results

Measurement of results had to be restricted to those KPIs for which we had foreseen short-term visibility (see Table 5.5.2), as the timespan for the evaluations was too short to expect long-term aspects (e.g., decreased number of negative events in different contexts) to become visible. For all evaluations, increased employee satisfaction was among these KPIs. We received rather negative verbal feedback on this for E1 and neutral feedback in E2 – for the latter, staff seemed to see potential in the app but was set off by lacking opportunities for using it. As reported above, for E3 and E4 we received positive feedback from managers about staff being satisfied with using the tool. For E3 feedback also included hints that individual interns performed better after reflecting on certain experiences, and for E4 participants reported on improvements in procedures, which also included compatibility among the procedures of the two departments. Answers from the post questionnaire indicate that participants from both E1 and E4 were reluctant about the improvements created by using the app (see the lower three items shown in Figure 5.5.6).

Another indicator for the results level we see the intention of the participating organisations and other, external users to further use the Talk Reflection App. After evaluating it at the care home used for E2 we discussed with different representatives of other care homes and created

interest for using it there. In this context, we are also discussing with a Dutch care group on how they would like to use the app in their work (a sustainability test has been done already). For RBKC usage in the departments used for E4 came to an end, as they have now been successfully merged. For the interns starting their internship at the time of writing this report, the app is currently rolled out again; there is also interest from other departments. Besides this the formative and summative results of our evaluations also created interest by other hospitals, care organisations, administrative units and companies, which we are currently exploring.

5.5.5 Conclusion & Discussion

The four evaluations of the Talk Reflection app show that the app can support individual and collaborative reflection, but that there is also room for improvement in this support. Such improvement may take place in different areas such as

- **features of the tool:** While we found support for documentation, sharing and discussion of experiences to be well received and used at least in some evaluations, other features such as the explicit documentation of results were used less. Further work on reflection tools should therefore create other ways of support for persevering outcomes.
- **socio-technical embedding of reflection:** We found that the successful and frequent usage of the Talk Reflection was closely related to which emphasis was on reflection in the organisations and whether people felt it was embedded in their work (i.e. the positive example of embedding it in a change process in E4 and the negative example of not using content created in the apps in face-to-face meetings in E1 or E2).
- **group and workplace constraints:** The evaluations show that for co-located small groups the value added by the Talk Reflection App was lower than for dislocated, larger groups. There is a need to add extra value in the former setting, for example by enabling users to prepare face-to-face (reflection) meetings by using the app (e.g., by picking topics to be discussed collaboratively). The latter setting points to enabling reflection in wider communities, possibly transcending department or organisation boundaries.
- **facilitation:** The evaluations point to the positive effects created by a motivated and skilful human facilitator and/or driver of reflection. On a process level we suggest to include such a driver into the planning of reflection if possible. This may be amplified or complemented by using features that facilitate reflection such as prompts (see D6.3) or questions posed by users sharing reports and explicitly asking others for their view on it.

On a methodological and measurement level we learned that single scales cannot be used to determine reflection success in all cases, but should always be seen as one component among other, possibly qualitative data. This is demonstrated by answers to questionnaire items in E4, which imply that using the Talk Reflection App had little effects on participants. Direct feedback indicated much more positive effects, which would not have been visible by relying on the survey only.

In addition, for collaborative reflection support we experienced an observer problem that is caused by a lack of documented insights or ideas in tools, which seems to imply that little or no reflection has taken place. In our evaluations (e.g., E1) feedback from users backs this conclusion taken from scales and content analysis, but other cases (e.g., E2 or the evaluation of the Talk Reflection App at RNHA in year 3, see D6.3/10.2) feedback from users was positive while little content could be found in the app. Gathering data in a socio-technical system that can be analysed like content from apps is more difficult and needs time. However, it is indispensible if the support of reflection apps is to be assessed.

Moreover, the question remains whether the relation between user activity and perceived benefit in using the app expresses a causality or not: While this differentiation seems to imply that those using the app more than others also benefited more, it may also be interpreted to show that those who were more positive about the app and its benefits used it more than others. Future evaluations need to include additional data gathering that allows the clarification of this.

In general we draw positive conclusions from the evaluations. Even in cases in which problems occurred (E1, E2) the app has supported certain aspects of reflection and created awareness for reflection. In other cases it was received well and used frequently, which shows that it has the potential to support all necessary stages (except the explicit documentation of outcomes) of reflection. The differences in constraints and contexts of the evaluations also allowed us deeper insights into the socio-technical aspects of reflection support, enabling us and others to embed reflection and supporting tools better into practice.

5.6 The DoWeKnow Evaluation at Infoman

The DoWeKnow App allows users to document their experiences with presentations at customers. Documentation and reflection can be done individually and together with others by commenting and rating slides.

5.6.1 Organisational context

Test bed organisation and the organisational unit

For a description of Infoman we refer to section 3.3.

Test users and their job roles

The ten test users were part of 3 departments:

1. Sales and Business Consultants work with clients and regularly use slides to present INFOMAN's products and services during pitches and later while working on projects.
2. The marketing department is responsible for designing and preparing standard slide sets that consultants can use. They also coordinate which slide sets can be used and collect and implement changes proposed by those that use slides.
3. Management is involved as they coordinate efforts of the different departments and also themselves are working as consultants from time to time.

A more extensive description of work roles at Infoman can be found in D6.1

Identified need and potential for reflective learning

The need for collaborative reflection on slides was identified during the user studies in year one and workshops conducted in year two. Slides are considered the main artefacts which are used by consultants at INFOMAN and to which they link experiences. The potential benefits for individuals are in a reflection about what went good or bad in a presentation and how it connects to a slide used. As a collaborative effort discussion of these individual reflections can result in outcomes related to the own performance as well as to the insight that some aspect presented in a slide is not communicated clearly enough. Participants can benefit from each other's experience and collaboratively collect improvements for slides and presentation strategies.

Potential organizational impact

The goal of using the DoWeKnow app is to implement standard processes around creation und publication of slides within INFOMAN so that all consultants only use slides validated by the marketing department. This may have two major impacts on the organization: First it should lead to higher customer satisfaction due to better slides and second should save time for consultants in preparing presentations as the good slides should be easier to find since slides are hosted and exchange via an online document repository instead of individual storage solutions.

5.6.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

The DoWeKnow App allows users to document their experiences with presentations at customers to reflect individually and together with others by commenting and rating slides.

Ratings are also thought to lead to descriptions of why a slide proved as especially useful (or not) and therefore inform others about applicability or lead to suggestions for improvements for a slide.

The app was developed by INFOMAN and integrated in their internal Knowledge-Management Framework that is based on Microsoft Sharepoint. Within that sharepoint instance a document repository was adapted and extended to allow users to rate and comment on single power point slides. It is therefore not a standalone application but more a combination of features within the existing infrastructure.

For a detailed description of the app see D6.4

Relation to MIRROR CSRL Model

The app supports users in their daily work (do work) by allowing them to browse through existing slides and search for new slides based on topic descriptions provided. Reflection sessions are not explicitly initiated but can be easily conducted by rating a slide or posting a comment. Users are encouraged to articulate experiences they made when presenting certain slides at events with customers. Others can be involved if they decided to subscribe to e-mail notifications for new comments. Comments and ratings are evaluated by members of the marketing department who then apply the suggestions that can be the outcomes of a reflection session and propose new versions of slides to other consultants.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

Reflection on slides starts with individual experiences that can be expressed in a condensed form as a rating or with a more explicit comment that describes a case where a slide was used during a presentation. These experiences are discussed collaboratively; others can add their own stories or can endorse statements. They are also encouraged to propose suggestions for improvements of slides. In the evaluation slides were also discussed in workshops and the participants decided which slides were most important to be changed. The organizational changes are initiated by the marketing department that collects comments on slides and improvement suggestions. Slides are then either changed and adapted to new use cases or larger changes like a whole new standard slide set e.g. for company presentations are developed to change the organizations image.

5.6.3 Research approach

Design and procedure

The evaluation was conducted in ten weeks from February till April 2014. Before the first and after the second workshop participants were asked to fill out the reflection scale questionnaire. After a workshop in which the procedure was discussed the first weeks were used to identify slides that are regularly used and that have clear potential for improvements. In a second step the participants were asked to use the app for 2 weeks to comment on the slides and make suggestions for improvements. These suggestions were evaluated and discussed during a second workshop to sum up the comments and decide what has to be changed. Afterwards the slides were edited by the marketing department and the app was used again to collect feedback and suggestions.

Participants

10 INFOMAN employees participated in the evaluation including 4 from Sales department, 3 from Business and Management Consulting and 3 from marketing. Among them described themselves 7 as men and 3 as female. The majority (9 out of 10) are between 30 and 39 years old. They work in the current position and team between 1 and 3 years while the overall experience with similar jobs ranges from 1 to 15 years.

Summative evaluation methods used

The evaluation methods used were a documentation of the workshop outcomes and a pre and post questionnaire based on the Core Questions Short Reflection Scale (CR) and KPI related questions. The latter were differentiated for consultants and marketing. Sales and business consultants answered the questions: „The standard slides from the repository help me to create good presentation."(KPI1_sales), „With the slides from the repository I can make good presentations at customer meetings." (KPI2_sales) and "I often use slides from the repository for my presentations at customers." (KPI3_sales) While participants from the marketing department were asked "I often receive feedback from colleagues regarding standard slides." (KPI1_marketing), "I can understand the feedback I get regarding standard slides." (KPI2_marketing) and "The feedback I get regarding standard slides helps me to improve them." (KPI3_marketing). In addition the post questionnaire included Core Question App-Specific Reflection (CA), Core Question Learning (CL) and Core Question Behaviour *(CL)*.

5.6.4 Results

5.6.4.1 Level 1: Reaction (Usage)

The app usage was embedded in a socio-technical evaluation that also contains workshops and discussion rounds. Therefore app usage was concentrated in the two weeks between the two workshops where participants were asked to add their comments and ratings and make suggestions using the app. During the usage period participants commented on nine different slides. The users were asked to comment on each slide in the repository.

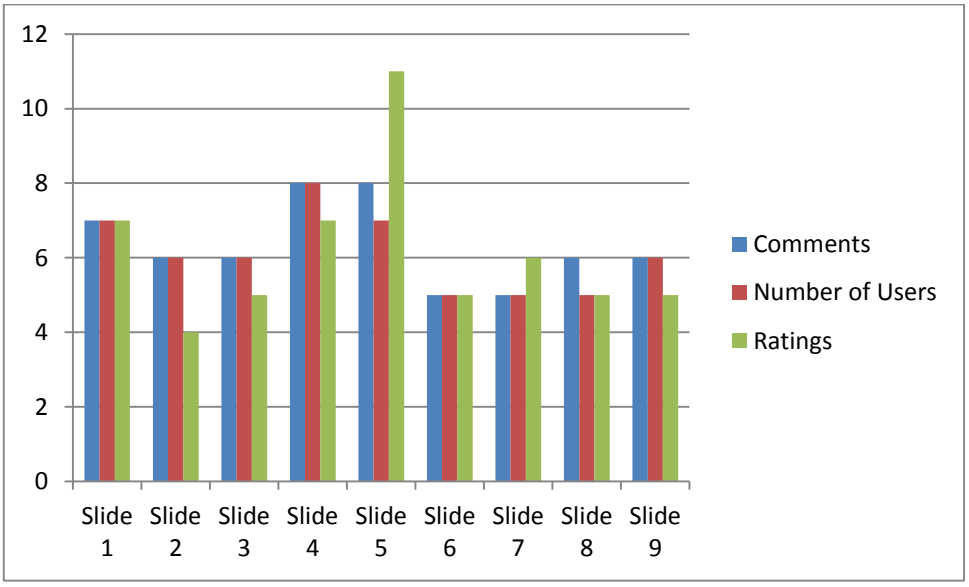

Figure 5.6.1 Participation in commenting, users that made comments and ratings on 9 slides

Figure 5.6.1 shows the number of comments and ratings made on a slide, the number of users that commented. Overall 57 comments and 55 ratings were created. An average of 6,33 comments and 6,11 ratings per slides. The average of all average ratings is 3,03 with a standard deviation of 0,41. In general 4 users commented on all slides while the others varied in particpation.

What was not measured were usage statistics that would indicate how often the app was used as a repository to find slides for a presentation. But the answers to KPI3_Sales of the post questionnaire indicate that it was used frequently. The question “I often use standard slide from the repository for my presentation” got an average agreement of 4.2 within the group participants.

5.6.4.2 Level 2: Learning

A comparison between the results of a pre and post questionnaire indicate that the app and the corresponding workshops had a positive effect on the reflection culture, although there are differences between the groups of consultants and those working in marketing (see Figures 5.6.2 and 5.6.3). While both groups see an increase in collaborative reflection in general (CR02) the values are in general higher for the consultants group than within the marketing group. Nevertheless we can also see a decrease in frequency of individual reflection (CR01) for the three participants from marketing. An explanation for the negative effects on this group could be, that most of the reflection outcomes and suggestions for improvements resulted in more work for them, while the consultants group benefited from improved slides without having the duty to implement the changes.

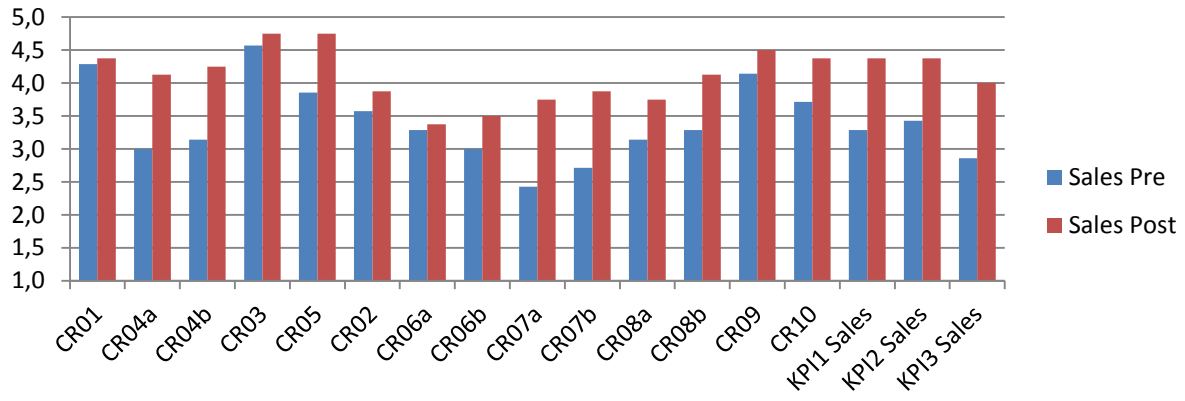

Figure 5.6.2 Comparison between results from pre and post questionnaire of participants in sales/consulting

Figure 5.6.2 shows a comparison of results for the group of consultants. The questions where answers increased most (CR04a/b, CR07a/b) are related to the frequency of reflection about slides and presentations individually and collaboratively. Here the app and workshops had a positive effect since they triggered discussions about slides and their applicability and potential. Consultants also see a benefit in this reflection (CR05) were they responded to the questions whether it helped them to reflect on slides to do better presentations with an increase from 3.86 to 4.75.

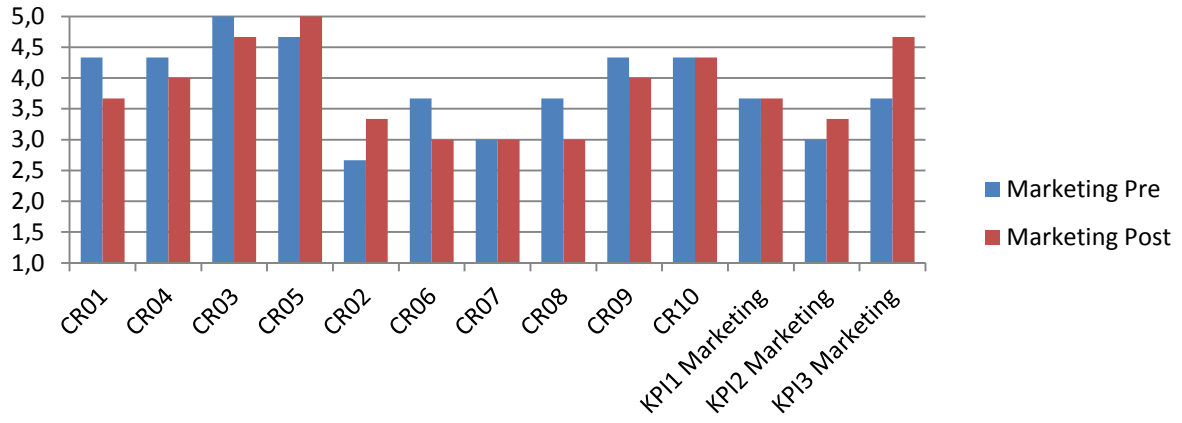

Figure 5.6.3 Comparison between results from pre and post questionnaire of participants working in marketing

Learning Process

The app in combination with the workshops clearly helped participants to identify potentials for improvements. In relation to the CSRL model we can say that it helped to initiate reflection and guide discussions for conducting collaborative reflection session. Average of CA questions was 4.12 (SD=0.19) with the highest ratings for questions asking if the app helped to discuss (4.27, SD=0.78) and improve (4.45, SD=0.82) slides.

Nevertheless negative answers from participants of the marketing department indicate that they still have a need for more communication about the topic (CR06&CR07). From the logs we can see that not all of the participants used the app and most of the comments only contained critique (e.g. “this slides is not self-explanatory”), but only few comments made explicit suggestions how they could be improved. As the marketing department was supposed to apply those outcomes they might have felt left alone with the questions how slides can be enhanced.

Learning Outcomes

Questions about learning outcomes asked in the post questionnaire were “After using the app I have deliberately chosen to change my presentations” (CL01) and “After using the app I have a better understanding of my work” (CL02). While the first question was answered with an average of 4.18 (SD=0.12) the latter received an approbation of 3.73(SD =0.11). We trace this back to the relatively small purpose of the app that has a positive impact on learning for the specific purpose of slides and presentations but not on the general level of work of consultants which consists of more complex tasks.

5.6.4.3 Level 3: Behaviour

Users report that the app has positive effects on their work. They answered with an average of 4.09 (SD=0.26) and 4.07(SD=0.27) on the questions “My work is easier since I use the app” (CB01a) and “I have a better overview about the content and products of Infoman since I use the app” (CB01b). Still there is again a gap between the perceived benefit of different user groups. While consultants answered those questions with an average of 4.25 to both questions (SD=0.70 and 0.88) the marketing staff answered with 3.66 (SD=1.15) and 3.33 (SD=1.52).

5.6.4.4 Level 4: Results

Similar results and increases are visible for the KPI_sales questions that ask more specific whether the app helps them to find good standard slides, improve their presentations and increase customer satisfaction with delivering better presentation. The comparison between pre and post questionnaire shows an increase by 1.06 in average for the KPI_sales questions from 3.19 to 4.25. The situation is more complex for the three participants from the marketing department which are less satisfied, with an average of 3.89 (SD=0.69) on each questions. Especially one participant that joined the test later, seemed not convinced. Nevertheless employees from marketing fully agree with the statement “The feedback I got helps me to improve slides” (KPI3_marketing) with an average of 4.67 (SD=0.65).

5.6.5 Conclusion

The DoWeKnow App has clear positive effects on the work with slides and presentations at INFOMAN. The evaluation presented here shows that especially the sales and business consultants benefit from a single source of presentation slides which are controlled and improved on their remarks by the marketing department. While the first are encouraged to reflect on slides and the way they used them the marketing staff has little to reflect upon since their main task is to improve the slides based on suggestions made by others. Due to the small number of the participants and the short time of the study we cannot determine to which percentage these benefits are related to the increased amount of collaborative reflection sessions. Clearly the largest contribution is made by standardizing the process of sharing and improving slides within the company which was an unstructured and individual task in the past. Also the amount of reflection by uttering experiences with slides is, at the moment, mostly a one way process as changes are delegated to the marketing department instead of using the app as a communication tool to improve not only slides but the presentations as a whole.

5.7 The Issue Articulation and Management App Evaluation at BT

At BT the Issue Articulation App is intended to capture reflection outcomes on advisors' level in order to incorporate them into the coaching sessions. The overall aim pursued is to improve customer satisfaction. The reflection outcomes are then visualised in the Issue Management App with automated analysis of relations between the outcomes and business objectives.

5.7.1 Organisational context

A general description of the testbed BT and the test users and their job roles can be found in section 3.1.

Although the evaluation should involve all job roles, it was hard to get the managers involved. Accordingly, the evaluation focused on the relation of advisors and coaches.

Advisors work in shifts and their workday is populated by one call after the other. Accordingly they have a stressful workday and not much time in between the calls. Coaches supervise the advisors by rotating between them and engaging randomly into calls in order to observe their behaviour.

Identified need and potential for reflective learning

Learning on the job plays a big role in the call-centres of BT. In order to increase customer satisfaction, BT pursues the optimization of the call handling of each individual advisor. At present, they arrange coaching sessions with advisors and train them according to their guidelines. However, this is more a top-down approach and BT likes to shift to a bottom-up approach, i.e. to increase the involvement of the advisors. The efforts made in this coherence are based on a change in organizational culture at BT aiming to improve quality in call centre support. Motivating the call centre advisors to reflect on their own experiences was perceived as a perfect match to this organizational campaign by the management and coaches. Making their reflection outcomes and lessons learned available for higher levels in organizational hierarchy results in bottom-up flow of reflection outcomes. Although there is an open-minded culture which encourages advisors to contribute to the optimization of the customer satisfaction, there is currently no defined process for this kind of communication which makes it more difficult to handle. The managers of BT see great potential in the implementation of learning by reflection with support of IT.

Potential organizational impact

The implementation of learning by reflection using the Issue Articulation and Management App is intended to increase customer satisfaction in long term. In order to reach this aim, there is a need to capture the knowledge of the individual advisors which are directly linked to the customers. Furthermore, a communication channel to the coaches and managers needs to be created to allow more transparency for the communication of the experiences. By adopting this approach, the coaches are then able to incorporate individual needs into the coaching sessions and therefore increase the quality in the calls which in turn increases customer satisfaction.

5.7.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

The Issue Articulation App is intended to capture reflection outcomes of the advisors in order to incorporate them into the coaching sessions to pursue the aim of improving customer satisfaction. Therefore the advisors are encouraged to reflect after each call and at the end of

their working shift in order to derive problems or ideas for improvement. In parts of the evaluation, the advisors have got a 15 min timeslot at the end of each working shift to reflect on their calls. The reflection outcomes are then visualised in the Issue Management App with automated analysis of relations between the outcomes and business objectives. Coaches and Managers are then able to reflect on their experiences incorporating the reflection outcomes of the advisors to derive topics for the coaching sessions.

Relation to MIRROR CSRL Model

The BT Evaluation of the IAA and IMA aimed to identify, support and proof organizational reflective learning. Since reflection does not take place on an organizational level but is always an individual or collaborative process (see D4.2, 4.3, 6.2, 6.3, 8.2 and 8.3), the relation of our evaluation approach to the CSRL model is split in different reflection cycles.

We expect an individual or a group to reflect before the actual articulation of a coaching need or a call of the day. The elements articulated in the IAA can therefore be seen as a reflection outcome of an individual reflection process, where the articulation in the IAA can be seen as the “apply outcome” phase in an individual reflection cycle. The articulation triggers a new reflection cycle that touches the organizational level.

Although there is still an individual or a group of individuals reflecting, the following part of the reflection cycle can be seen as on the organizational level. Since characteristic objectives and organizational units are involved, the “initiation of the reflection” session is the first phase in this new reflection cycle.

The “conduction of the reflection session” is characterized on the organizational level by the information made available in the IMA. The relation between different coaching needs is visible, the coaching needs can be filtered, sorted or just browsed to understand the meaning of the articulations. The IMA supports this phase of the reflection cycle especially.

The “outcome” can then be applied as a decision on what to coach, on changes in the behaviour of advisors or in the decision to address the coaching need later. The IMA as used in the evaluation scenario at BT allowed the transfer of coaching needs into a so called coaching form. In this form the coaching needs can be marked as resolved and there is room to document the steps taken and the expected impact of these steps. The transfer in the CSRL model from “apply outcome” to “plan and do work” marked with “change” can therefore only be documented in the app, but it is not directly supported by the app.

From the usage scenario at BT we see the IAA supporting the “outcome” transfer in an individual reflection process and the transfer from application of these outcomes to a new reflection cycle on an organizational level supported. The IMA then supports the “initiation of a reflection session”, the “conduction of the reflection session” and the “application of outcomes” on an organizational level. The change can then only be monitored but not supported by the apps.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

The apps support individual and organizational reflection.

Transition model (Figure 10.3.1):

The Apps cover especially step 1 (reflection of the advisor), step 2 (reflection of the coach) and step 3b (application of outcomes by the coach/manager).

Push and Pull mechanism:

The Apps support a push mechanism as well as a pull mechanism. The use case in the evaluation, presented in the paper at hand, represents the push mechanism. The advisors can articulate coaching needs and issues and push them forward to a different organizational level where they can be addressed or solved. The Pull mechanism was not part of the prototype for the evaluation but was developed independently as an add-in feature for the IAA and IMA in parallel. It allows organizational deciders to notify IAA users and request input for certain elements of the underlying process.

5.7.3 Research approach

Design and procedure

The data collection for the evaluation reported in this deliverable took place from beginning of March to mid-April 2014 (week 10 to 16). Though, the usage of the app still continues and an end date has not been set yet. The situation before and after the evaluation have been measured by questionnaires. A control group, who also filled out the questionnaires, has been used for comparison aspects. Employees of the DFKI introduced the app to the project manager of BT and we put additional development effort into the integration of the apps into the organisational environment. The project manager of BT then introduced the app to the coaches, managers and advisors. Before the participating advisors, managers and coaches got their own access to the apps, the project manager at BT introduced the pre-questionnaire. The users filled in the questionnaires and we opened the app for testing afterwards.

Participants

We had 68 users in the experimental group and 10 users in the control group. According to the pre-questionnaires (40 participants), about 40% of the participants were between the age of 20 and 29, 20% between 30 and 39, 25% between 40 and 49, 10% between 50 and 59 and about 5% were above 59. 58 % of the participants were female and 42 % male. The participants were on average 3.2 years in their current job position (SD = 3.14).

According to our log files from the IAA, 57 participants articulated at least one coaching need or call of the day. We still count the 11 people that did not actively participate as part of the experimental group instead of the control group, since the app was introduced to them and they had insights in the usage process that may have influenced their answers in the post-questionnaire.

Regarding the questionnaires the control group did not fill in enough questionnaires to compare results. We therefore decided to limit the comparison with the control group to the KPIs.

Summative evaluation methods used

The methods used for summative evaluation of the IAA and IMA at BT were a pre- and a post-questionnaire. In addition a detailed logging of usage data from the evaluation participants gave us insights on usage behaviour. The questionnaires were anonymized and therefore there is no clear connection between questionnaire answers and user behaviour. We therefore asked participants in the post questionnaire for permission to connect their answers to their tool usage logs. In total only eight participants approved the connection of the data. For clarification and further insights there are management interviews planned with coaches and managers that participated in the evaluation. The interviews will be held after the creation of

this document and the insights are presented in D8.3 and will be presented in the WP8 session at the final review.

The pre- and post-questionnaire both contained all demographic questions from the toolbox.

Level 1: Reaction – There is objective data available in the usage logs of IAA and IMA and subjective answers on the questions from the toolbox are available. The wording of the level 1 questions in the toolbox had to be adapted to the understanding of the evaluation participants. We were especially interested in questions CU01 and CU02 and therefore integrated these.

Level 2: Learning – Both questionnaires contained the short reflection scale (CR) from the toolbox. Furthermore, the post-questionnaire contained 13 app specific questions[18] (CA), and 2 questions regarding learning outcomes (CL). We added a set of use case specific questions to deepen the insights from the questionnaire. The questions asked on the influence of own wishes and needs on the coaching process and organizational change in general, as perceived by the participants. Further we used additional questions from the toolbox (LT).

Level 3: Behaviour – We used the core behaviour question (CB) from the toolbox and added several use case specific questions in the post questionnaire to ensure a deepened understanding of the behavioural changes.

Level 4: KPIs – The BT call centre allowed us access to KPIs describing the performance measures for the participating staff members in three sets. One set is from before the evaluation period, one was measured within the testing period and one was measured after the testing period. These KPIs were available for all participating groups and for a test group not participating in the evaluation.

5.7.4 Results

5.7.4.1 Level 1: Reaction (Usage)

According to the questionnaires, advisors estimated that they have used the apps on average 1.7 (SD = 0.45) times during the evaluation period and that the average usage time was 1.88 minutes (SD = 1 .58). Not all users who filled out the questionnaire have answered the question which results in the low usage value. As only two coaches have answered the questionnaire, the average value of 15 times is not resilient. However, only 57 out of 68 users actually articulated coaching needs or calls of the day throughout the testing period and are therefore considered as “active” users. The log files of the active users show that the mean objective usage frequency is 5.06 (SD = 4.66) times. The following table gives an overview of the total usage figures of different functions. Though, the table does not represent a complete list of functionalities and therefore the average value differs from the usage frequency depicted above. Due to the fact that only few coaches answered the questionnaire, the average value of the log files is higher than the overall average value resulting from the questionnaire.

Table 5.7.1. Usage frequency of main functionalities of the Issue Articulation and Management App

Function	Mean	Standard Deviation
Save Coaching Need	1.96	1.40

[18] The toolbox questions used from CA were: 2,7,10,13,17,26,31,32,33,34,37,38,41

Save My Call of the Day	1.5	0.76
Logins of Coaches/Managers	38.45	60.04
Show Coaching Observation Form	24.11	38.44
List Coaching Needs	19.86	14.99
Show specific issue	23.63	27.88

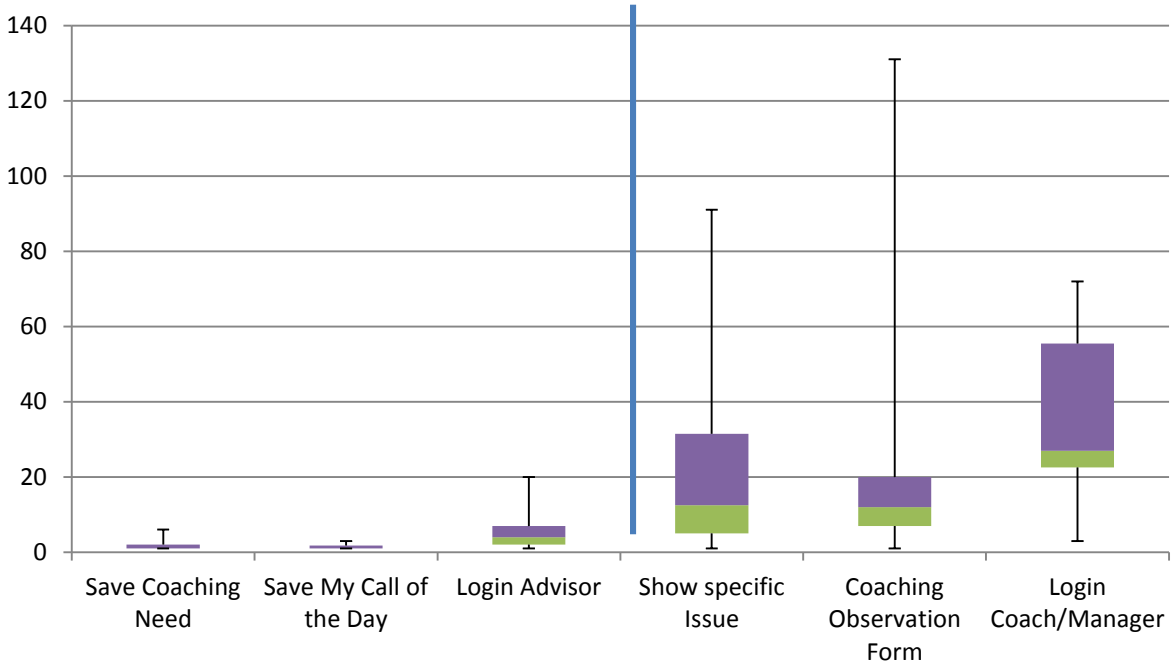

Figure 5.7.1. Usage of the main features of the Issue Articulation and Management App

Figure 5.7.1 shows the usage of the main features of the Issue Articulation and Management App. The left side of the diagram visualizes the main features used by the advisors, whereas the right side of the diagram shows the usage by coaches or managers. It is clearly visible, that the coaches and managers respectively were far more active. This corresponds with the answers in the questionnaire regarding the benefit of organisational learning by reflection (cp. 5.7.4.2). Though, it is also clearly visible that there have been huge differences between the single employees, especially regarding the coaches. It seems that some coaches/managers were far more active than other coaches/managers. In general we observed a much higher frequency in app usage for the coaches/managers then for the advisors. We interpret this as a sign, that the deeper understanding at this higher organizational level motivated participants to use the apps more often.

We could also clearly identify this trend in the questionnaires. Advisors stated a much lower number of usages then coaches/managers.

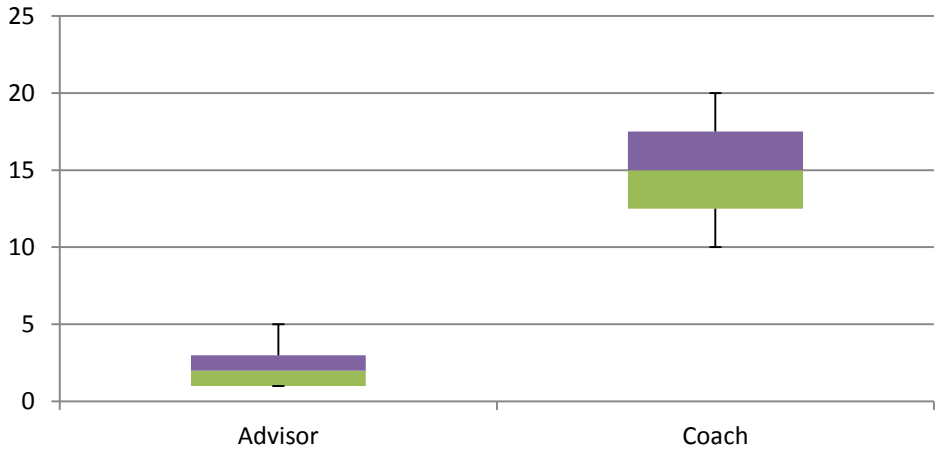

Figure 5.7.2. Times of usage according to questionnaires

From open questions in the questionnaires we experienced, that the advisors who did not use the app continuously, had different reasons for that:

- There was no need to use the app in the timeframe of the evaluation (no coaching need occurred. Some participants were very experienced and needed little to no coaching).
- Some advisors sit next to the coach, so it was faster to contact the coach bilaterally
- Lack of time

Due to the fact that only 5 participants responded to the question, why they didn't use the app, it is not possible to draw general conclusions.

The Questions USE01 to USE04 focused on identifying barriers which limit the usage of the apps (cp. Figure 5.7.3). The figure shows that regarding the questions one to three there is a great diversity. One assumption is that this results from the different handling of the various teams that participated. It also seems to be the case that time pressure (USE01) is an influencing factor, but with a mean value of 3.00 (SD = 1.41) it doesn't seem to be the case for everyone. Though, the high standard deviation shows that it is the case for some participants. The results of questions three (no advantage: M = 3.17, SD = 1.17) and four (lack of motivation: M = 3.00, SD = 0.89) are not surprising. First of all the advisors do not benefit by using the app in the first place, but coaches and managers do. On closer inspection the coaches and managers do not agree with question three, whereas advisors rather tend to agree to this question. Moreover, the advisors are rated according to the number and quality of their calls, so they will always choose to do one more call rather than using time on the app.

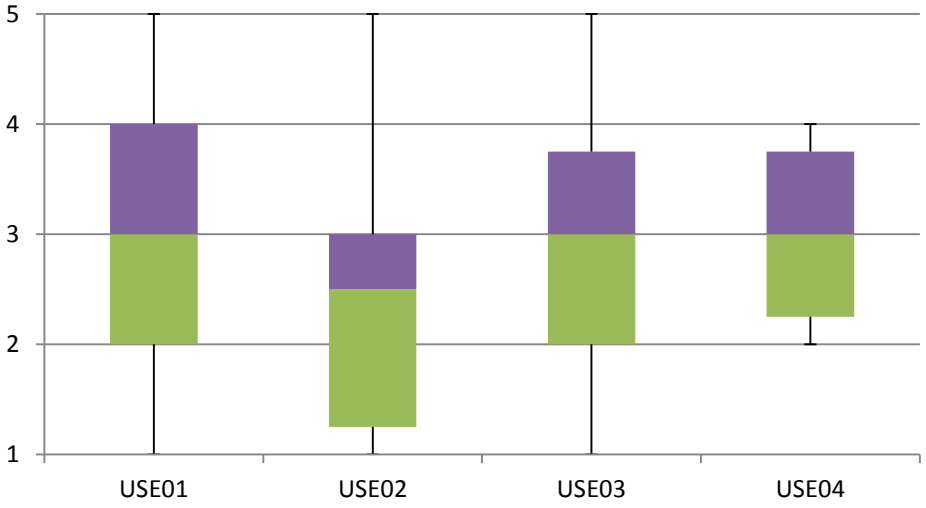

Figure 5.7.3. Insights on barriers

USE01: I did not have the time to use the app.

USE02: I did not have the physical space (e.g. necessary privacy) to use the app.

USE03: I did not see an advantage in using the app.

USE04: I was not motivated to use the app.

In general there is a slightly positive tendency whether the participants want to use the app as part of their work-life (LT02) (Mean = 3.5, SD = 0.88).

5.7.4.2 Level 2: Learning

Learning Process

The Short Reflection Scale (SRS) is slightly lower in the post-questionnaire than in the pre-questionnaire (cp. Figure 5.7.4). Though, it is not significant. While in the pre-questionnaire the SRS of the advisors is about 4.15 (SD = 0.73) it decreases in the post-questionnaire to 3.89 (SD = 0.89). A very little decrease could also be observed for the coaches/managers from 4.45 (SD = 0.55) in the pre-questionnaire to 4.35 (SD = 0.65) in the post-questionnaire. Nevertheless the SRS for the coaches/managers is in general higher than for the advisors.

It seems that no significant change regarding reflection took place during the evaluation period but the general level is clearly positive.

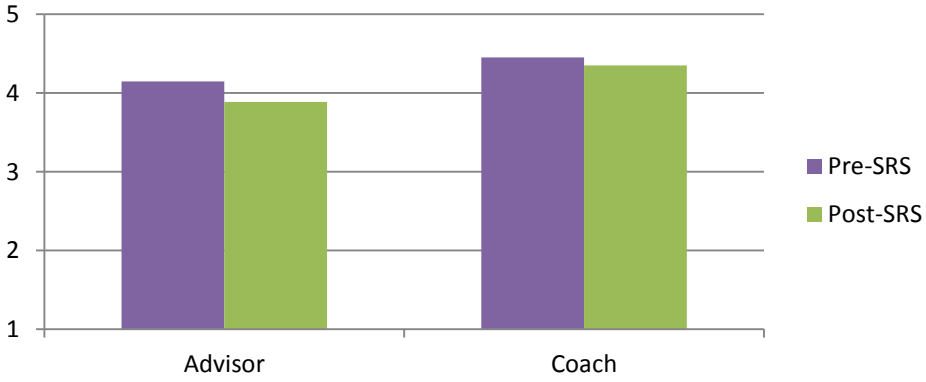

Figure 5.7.4: Short Reflection Scale

It could also be observed that the coaches/managers have a higher awareness regarding the importance of reflective learning for the organisation than the advisors (cp. Figure 5.7.5). The difference between the pre- and post-questionnaire is just marginal: The mean value for advisors in the pre-questionnaire is 3.55 (SD = 0.96) and in the post-questionnaire 3.59 (SD = 0.91). Compared to the coaches, the awareness decreased slightly from 4.10 (SD = 0.68) in the pre-questionnaire to 4.05 (SD = 0.78) in the post-questionnaire.

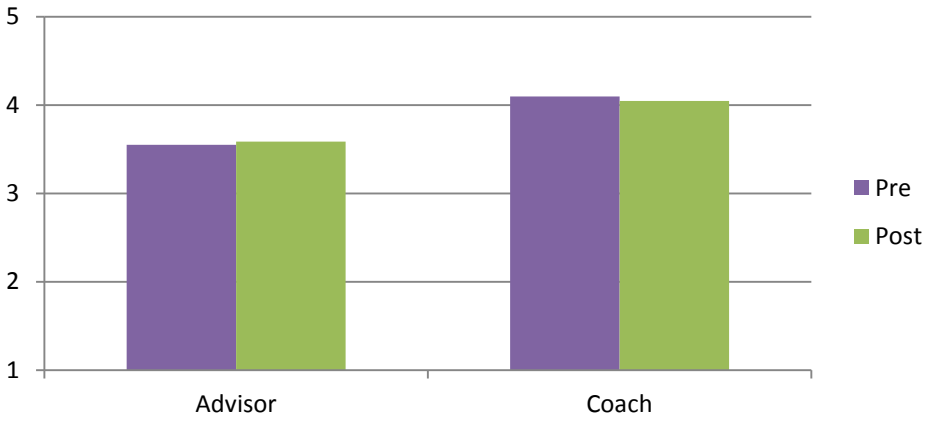

Figure 5.7.5. Awareness of importance for organisation

Regarding the app-specific questions the participants also had a more or less neutral attitude (Mean = 3.39, SD = 0.90).

Learning Outcomes

The questions regarding the learning outcomes showed a neutral attitude of 3.28 (SD = 1.07) regarding a 5-point Likert Scale from 1=”strongly disagree” to 5=”strongly agree”. Though, as we observe the organisational perspective we need to differentiate between the person that reflects and the person who applies the outcome (cp. ICO transfer e.g. D8.3). Apparently the advisors will not directly be able to observe whether there was a change, because there will first be a recursive reflection by the coach/manager and then the outcome will be applied. Both questions begin with the statement “after using the app”, though they will not recognize the impact directly after using the app. According to what we could observe from the coaching feedbacks, which represent the reflection outcomes of the coaches, there are definitely learning outcomes regarding the reflection sessions of the advisors. It seems that the ICO transfer hinders the perception of the advisors regarding learning outcomes.

5.7.4.3 Level 3: Behaviour

On the basis of the questionnaire the application of the learning outcomes could not be proved in general. The participants answered the question, whether the app helped them to improve their work performance, between disagree and strongly agree. On average the questions have been rated with neutral (Mean = 3.21, SD = 1.01), so it is difficult to make conclusions out of that result. Though, this may be a result of the ICO transfer. As already stated in 0 the ICO transfer seems to hinder the perception of the advisors regarding the learning outcomes, so it might be that due to the non-direct feedback the advisors are not aware that the app has an influence on their behaviour. Based on the KPIs we received from BT regarding one team we saw a tendency towards correlation between the app usage and the Net Promoter Index (cp. Figure 5.7.6) as well as the Advisor Sat (cp. Figure 5.7.7).

Net Promoter Index (NPI): The Net Promoter Index is based on customer advocacy, i.e. "how likely are you to recommend our services to others based on your recent experience with us".

Advisor Sat: All customers receive a post call SMS message asking how satisfied they were (between 1=Poor and 10=Excellent). The percentage is calculated dependent on how many customers score the advisor. The current internal target of BT is 90 %.

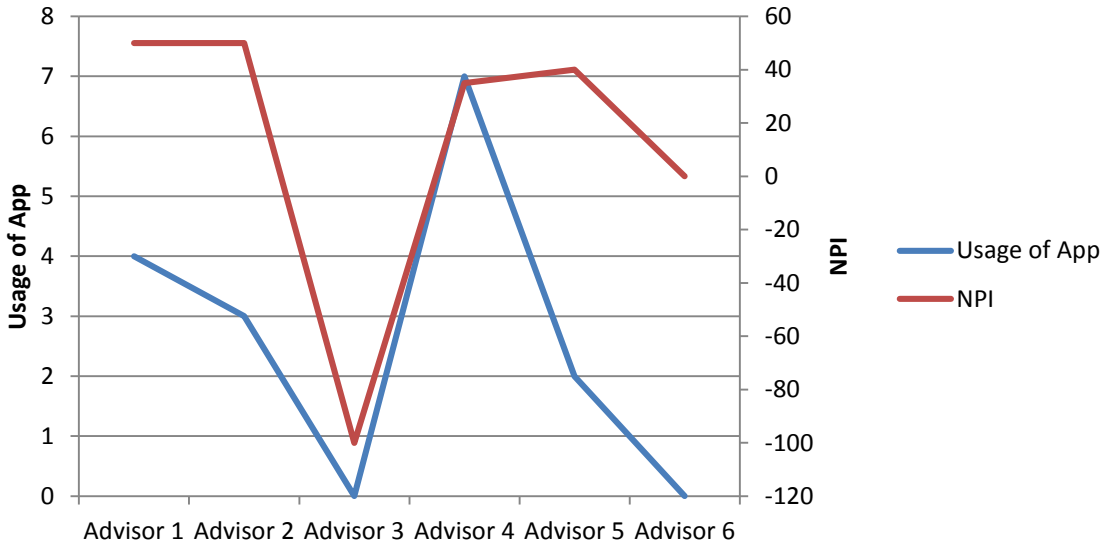

Figure 5.7.6. Correlation of App Usage and NPI

In Figure 5.7.6 the blue line represents the number of app usage during the evaluation period of the single advisors. The red line represents the Net Promoter Index regarding the corresponding advisor. Although there is no linear correlation between the degree of the deviation of the NPI and the usage of the app, the tendency, whether the NPI/Usage of App is low or high seems to correlate.

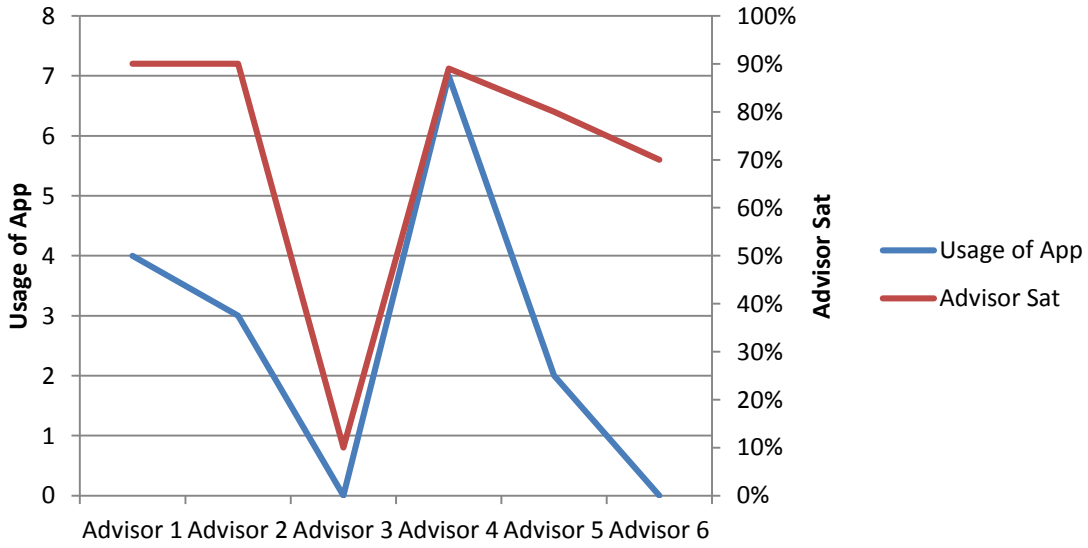

Figure 5.7.7. Correlation of App Usage and Advisor Sat

The same can be stated for the correlation between usage of app and Advisor Sat (cp. Figure 5.7.7). There is no linear correlation between the degree of deviation, but the tendency seems to be correspondingly.

5.7.4.4 Level 4: Results

According to the KPIs we received from BT, the teams which participated in the evaluation show a better performance than the control group (cp. Figure 5.7.8). Three KPIs have been monitored: Net Promoter Index (NPI), Advisor Sat and Repeat Calls. The first two KPIs have already been introduced in 5.7.4.3. Furthermore the KPIs were reported three times. Once before the testing period, once during the testing period and once shortly after the testing period. Since the impacts of learning processes may take time to establish, the numbers will have to be observed further in future KPI controlling to gain reliable conclusions. For now the results are only to be understood as tendencies and immediate effects that could not be observed longer due to the deadline of the deliverable creation. We will try to keep track of the KPIs until the end of the MIRROR project and show possible results at the review. Since several measures are taken in the call centre teams on a daily basis to improve performance, it is very difficult to identify and isolate the effect of the apps in use. We compare the results to a control group but we still know that the isolation will stay a challenge.

Repeat Calls: This represents the percentage of calls an advisor dealt with compared to the number of callers who call back within a 7 day period stating the same issue. The current internal target is 17.5 %. In contrast to NPI and Advisor Sat, a low percentage is preferred here.

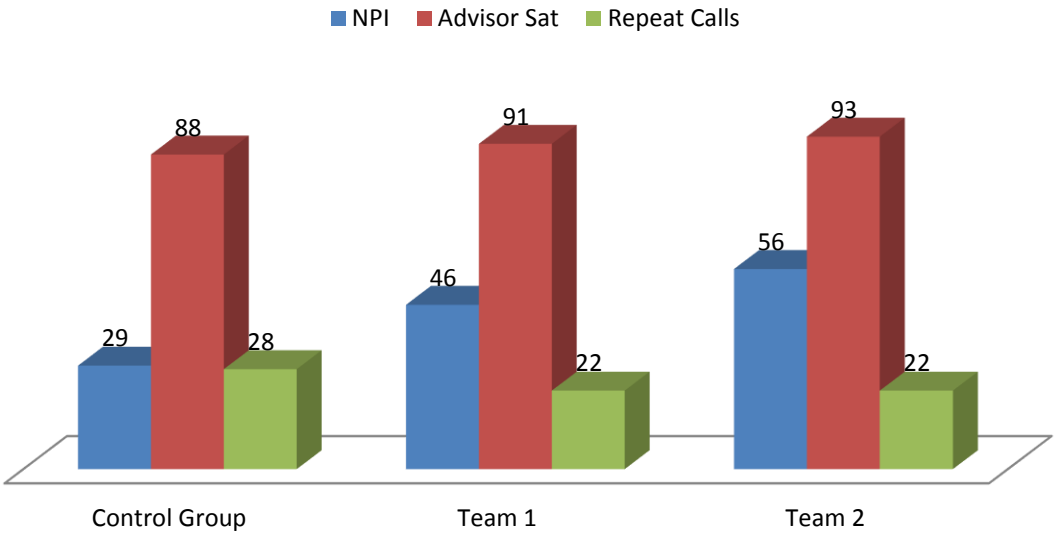

Figure 5.7.8. KPIs in evaluation period compared to control group

During the testing period the three reported KPIs were distributed as shown in Figure 5.7.8. The NPI of the control group is notably lower than of the two participating teams. The Advisor Sat of team 1 and 2 are also higher than of the control group. Whereas for NPI and Advisor Sat a high value shall be achieved, repeat calls should be kept low. Although there is no big difference, team 1 and 2 have less repeat calls than the control group. In summary, it can be stated that the teams which participated in the evaluation performed better than the control

group. However, we need to be careful with the interpretation of this data. Although a correlation between the usage of the app and NPI/Advisor Sat could be shown in one team, there were also other influencing factors which impact the measured KPIs. Accordingly, the better performance of team 1 and 2 may be a result of the app, but it does not have to. In order to observe organisational learning a much longer evaluation period is needed, especially in such a large organisation as BT. Decision processes in large companies take much more time than in smaller companies, so changes can only be observed in longer term evaluations. Still, a tendency towards improvement in task performance over the testing time and therefore organisational learning was shown based on our measures.

Interviews with the project coordinator from the BT call-centre showed that they were very positive about the app. They decided to further use the app for their own evaluations. In the mid of May we had additional 30 users compared to the evaluation period, who are working with the app. We agreed upon to keep the app up running at least until the review in September. This is a very strong statement as we did not ask them to further evaluate the app. Accordingly they see great benefit in using the app.

The KPI data was made available by BT for three measurements. One before the testing period, one during the testing period and one after the testing period. A comparison of the NPI before, during and after the testing period shows a slight downward tendency in the control group and a slight upward tendency in the testing teams. The three teams started almost at the same level and significant changes were observable through the testing time and after the tests (cp. Figure 5.7.9).

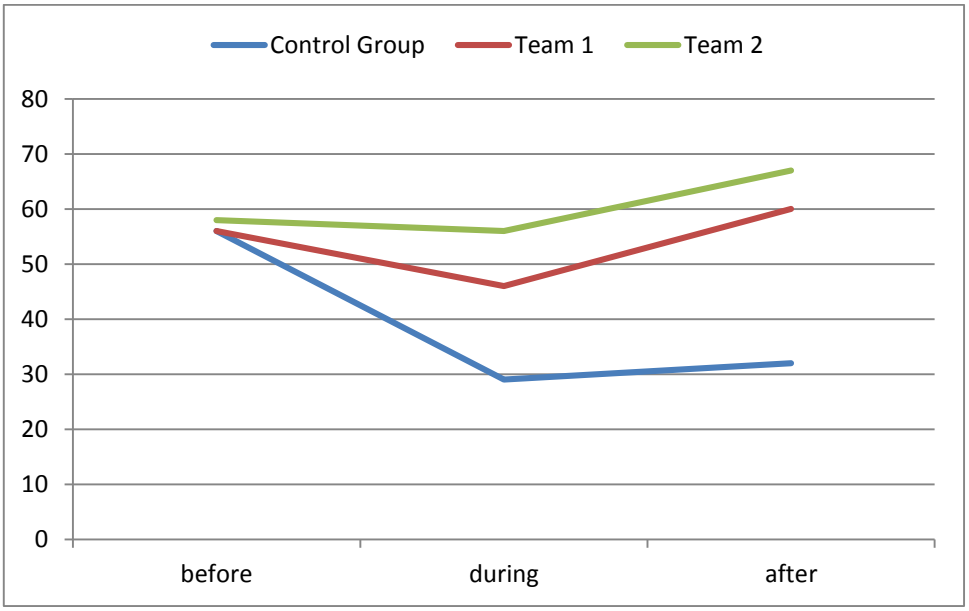

Figure 5.7.9. NPI over time

Comparison of the Advisor Sat KPI over the three timespans shows a notable downward trend in the control group although this team had started with the highest value in this KPI. The testing teams stayed at the same level or raised the level of Advisor Sat in the testing phase. It has to be mentioned though, that this is a KPI, the advisors are very much aware of. It showed in the interviews at our first test bed visit in Dundee that the advisors are very proud of a high level of Advisor Sat and are very much aware of the way they can influence this KPI.

The slight tendencies stated above therefore appear on a generally very high level of this KPI and the changes observed are not significant.

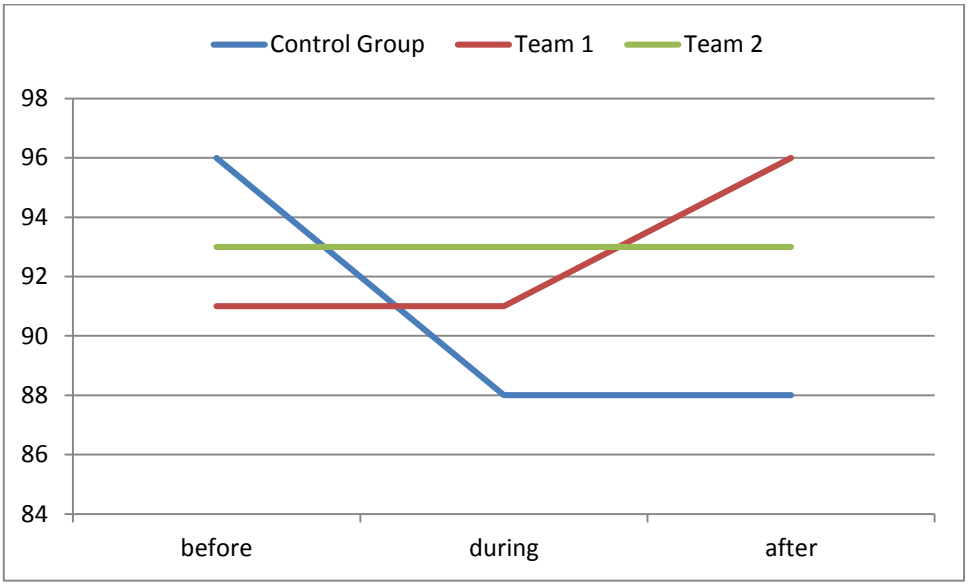

Figure 5.7.10. Advisor Sat over time

In the Repeat Calls KPI, no real tendency is notable. The testing groups started at a lower level and maintained that performance; the control group started at a higher level and was able to improve the performance by lowering the KPI through the timespan observed. Still the control group remained on a higher level than the testing groups. On average we don't see an impact of the app on this KPI based on this observation. Still there is a notable tendency when the KPIs are compared in more detail. Based on the data made available for us by BT we can see the KPIs for individuals within the teams. Especially in Team 2, the person that had the most intensive app usage based on the log data, managed to lower the 7 day repeats from 23.6 before to 16.2 during and 17.7 after the testing period. This cannot be proven to be an effect of app usage.

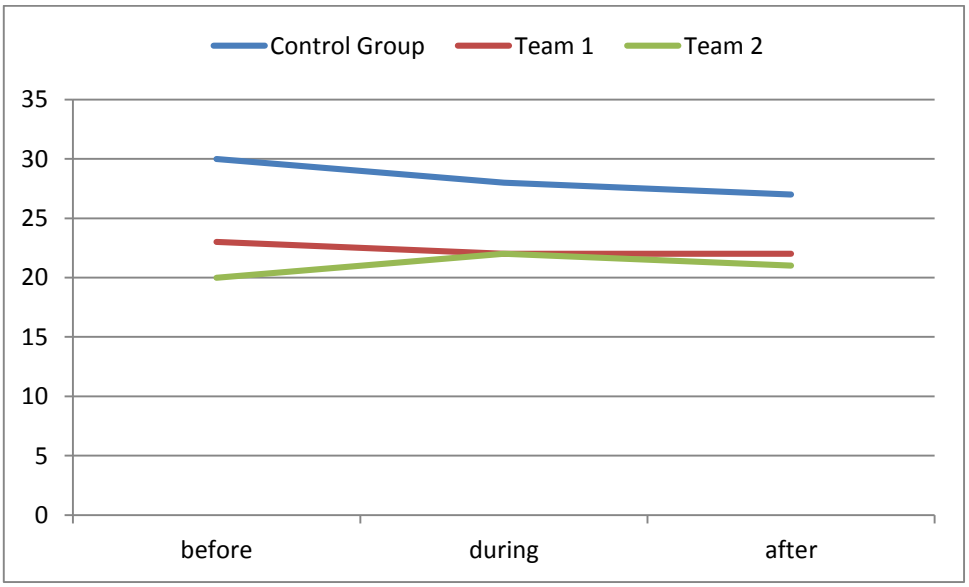

Figure 5.7.11. Repeat Calls over time

Over all the effects observed cannot be assigned to app usage without any doubt. Nevertheless there is an overall positive tendency in the teams that were in contact with the app and partly there were significant improvements in these teams compared to the control group. To verify our perception of an overall positive tendency, we interviewed coaches and managers after the testing period to ask for their impression on usefulness and usability of the apps. The results of these interviews are not available by the time this deliverable is created but will be reported at the final review. The direct feedback from users in the open questions of the questionnaires was diverse. We had feedback from advisors on the benefit of the app which varied from “I do not see any benefit of the app” to “Good App; would be really useful in large teams where there is little opportunity to ask direct questions to team members/coaches and managers when stuck”. Coaches saw use in the app for large teams that can be very challenging for coaches: “I think it is suited to a call centre based team where the coaches do not get much investment time. This way they can schedule time daily and complete all of the questions/calls and coaching within that time.”

5.7.5 Conclusion & Discussion

In general the evaluation revealed what we expected. We could show a tendency that organisational learning took place based on the integration of reflection and support with our app.

The feedback in the interviews with the project coordinator was very positive so they also seem to believe that there is a correlation between the performance of the teams and the app. At least they see great potential in reflective learning in their organisation in combination with the Issue Articulation and Management App. One very promising result of the work with the coaches and managers is the intention of the BT call centres to keep working with the app and to leave it embedded in the working process for the teams. We will therefore discuss options for further support and we will stay involved in the development process of the app after the MIRROR project.

On the one hand we had to do a lot of customizations of the app in order to fit it seamlessly into the work processes, but on the other hand this enabled us to inspire the coaches and managers of the idea of reflective learning and its benefits for BT.

Over all we had hoped for more feedback from advisors, but the reactions in the questionnaire on usefulness and usability of the app were all neutral or slightly positive. There were certain limitations in the app (e.g. only three coaching needs could be active at a time), but we generally observed decreasing participation in the test over the testing period. This showed us how important it is to create benefits for the individuals to foster organisational learning. We got the impression that in a real life implementation in all teams and probably even in different call centres, the articulation of needs might become part of organizational culture and usage of the app for articulation could become natural. In our tests the influence of a person in charge, motivating the participants was immense. In the evaluation phase the project manager at BT did a very good job in motivating participants and his engagement became directly visible e.g. in the number of issues articulated. After he talked to team managers in rather inactive teams, the number of coaching needs articulated rose immediately.

Over all we see the app evaluation as a full success especially because the test bed is willing to work with the app after the testing phase. From all our interviews and discussion on the potentials of the app, we see our research approach and the ideas of the MIRROR project over all confirmed.

5.8 The Medical Quiz Evaluation at NBN (Workshop and Stroke Unit)

The medical quiz is an application especially developed for nurses working at a stroke unit, thus the questions concern knowledge relevant for their specific work. The goal of the quiz is to serve as a trigger for reflection, especially for connecting theoretical knowledge to practical work experiences. Therefore reflective questions are integrated into the game to serve as reflection amplifiers and to provide guidance for the reflective learning process.

5.8.1 Organisational context

Test bed organisation and the organisational unit

After a formative evaluation of the Medical Quiz v1 in 2012/2013 (see D10.2), the improved and extended version of the Quiz was assessed by means of a summative evaluation. This summative evaluation of the Medical Quiz was carried out in the Neurological Clinic Bad Neustadt (NBN) during the qualification course „Spezielle Pflege auf Stroke Units“, which is organized by NBN and on the Stroke Unit of NBN itself (for a more detailed description of NBN and the Stroke Unit we refer to section 3.4

Two evaluations of the Medical Quiz were conducted at the Neurological Clinic Bad Neustadt (NBN), and addressed mainly the needs of nurses working at a Stroke Unit.

The first evaluation of the Medical Quiz was conducted during the qualification course for nurses, who are working at stroke units in different German hospitals and which was organized and located at the Neurological Clinic Bad Neustadt. The second evaluation took place at the Stroke Unit at NBN itself.

The qualification course takes place once a year. In 2013/2014 about 21 nurses have participated in this course. The course started in October 2013 and ended in February 2014. In each month one course week has taken place, so the course consisted of altogether 5 weeks. The evaluation started with the beginning of the course in the middle of October 2013 and ended with a workshop during the course week, which was held in the middle of January.

In the second evaluation, three nurses working directly at the Stroke Unit participated. They used the Medical Quiz for starting in February for two months directly during their working shifts.

Test users and their job roles

The work of nurses in the stroke unit is generally divided into three shifts comprising early, late and night shifts. While during early and late shifts, usually about six to eight nurses are on duty, in the night there are only up to four. The responsibility of nurses is to ensure medical treatment of patients as well as assuring their physical and mental well-being. The tasks they perform in order to fulfil these responsibilities are manifold and include the implementation of directives e.g. for medication given by physicians, the organization of patients' days including their transport to examinations and the documentation of both care given and physiological data measured during the day.

The early shift of nurses is structured by the day structure of patients, including tasks such as waking them up and serving them meals, and physicians' needs such as the ward round. Other influences they have to react on are unforeseen incidents such as emergencies and external influences such as therapists entering working with patients or patients being collected for external examination.

In the morning and late shifts, each nurse usually is responsible for patients in two rooms. One nurse per shift is responsible for the emergency ward and carries around a telephone to be informed about people being moved to this ward.

In both settings, the nurses were learners. During the workshop the nurses were learners who wanted to become more qualified for their work at a stroke unit, including the special needs of stroke patients. In this setting the Medical Quiz gave the nurses the possibility check their newly gained knowledge, to think about their learning progress and to establish a relation between the newly gained knowledge and their own working experiences with the help of the posed reflection questions during the quizzes.

At the stroke unit at NBN, the nurses used the Medical Quiz during their work. Here the quiz should help them to keep their knowledge up-to-date, to refresh and strengthen knowledge they already have and also to establish a relation between the newly gained knowledge and their own working experiences with the help of posed reflection questions.

Identified need and potential for reflective learning

In both settings the gaining of new knowledge as well as keeping professional knowledge up-to-date is amongst the major challenges for participating nurses. In the health sector it is particularly critical for professionals to know about new treatments or new developments with respect to medical and care procedures. Relating theoretical knowledge to practical experiences and in the end to learn from this reflections is seen of crucial relevance for nurses at a stroke unit. Within the Medical Quiz reflective learning was induced in posing reflection questions before, during, and after a quiz. The potential of these questions lies especially on their occurrence while playing the quiz and that they refer directly to work-related knowledge and experiences.

Potential organizational impact

If the Medical Quiz achieves its objective, the long-term usage at NBN would have the potential to provide nurses with continuous access to new knowledge, and to bring together theoretical with practical knowledge. From an organisational point of view this means, that hospital staff would be continuously up-to-date with new knowledge. For the hospital as organisation, this would help maintain a high quality of care. In turn, this should improve the satisfaction of employees, patients and their relatives.

5.8.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

The goal of both long term evaluations was to find out, if the Medical Quiz, including reflection guidance components in the form of reflection questions, has the potential (i) to trigger reflective learning on an individual level, (ii) to connect theoretical knowledge with work experiences and (iii) to improve the nurses' working behaviour (e.g. better care for their patients, higher patients and relatives satisfaction).

The Medical Quiz consists of four different quizzes: a "Quiz against time", a "20er Quiz" (answer twenty questions), a "10er Quiz" and a "5er Quiz". For all the quizzes content-based questions are randomly chosen. The content-based questions are multiple-choice or single-choice questions covering knowledge relevant for nurses working at a stroke unit. Additionally reflection questions (open questions) were added on top of the usual content-related quiz questions at the beginning, during, and at the end of a quiz. At the beginning of the quiz the

reflective question makes aware of the current knowledge status, depending on the gained quiz results on previously played quizzes. The in-between questions put the focus on the content-based questions and how they refer to past working situations and working experiences. The questions at the end of the quiz asked explicitly for gained insights or new knowledge with regard to the currently played quiz. After having finished a quiz, the quiz results are directly presented to the players, which again is worth reflecting on.

Because users answer all types of reflection questions while they are playing the quiz, reflection is initiated on an individual level. The given answers can be seen as insights or reflection outcomes. At the moment neither collaborative nor organisational reflection are directly addressed by the quiz, however discussing one's own experiences and quiz results with colleagues can initiate collaborative reflection.

Relation to MIRROR CSRL Model

Regarding the entire reflection process, using the quiz can be seen as self-directed learning activity, and reflection is triggered by the quiz with reflection questions. The creation of learning outcomes is also supported by the reflection questions. In the case of the qualification program for nurses "working" is viewed as learning activity. Since the quiz is part of this learning activity, the "working" stage is addressed by all nurses. If the nurses are going through the other stages depends on the individual learning goals of the nurses.

For the SU Evaluation, plan and do work is carried out by the participants independently of the quiz. However, situations at work might trigger or motivate nurses to play the quiz (e.g. because they realized some knowledge gaps) and on the other hand, playing the quiz (i.e. gaining new knowledge and linking the theoretical knowledge with their practical work) might influences nurses' working behaviours.

Considering the single phases of the CSRL model, the quiz supports the following two types of reflection cycles:

Cycle triggered by content-based questions: While playing the quiz (Plan and do work), the nurse can become aware of individual knowledge gaps or needs (initiate reflection). After playing the quiz, she might find out in more detail, by having a look at the quiz results, which knowledge is missing and where she has to deepen her knowledge (conduct reflection session). After having deepened the missing knowledge, she might play the quiz again and reflect about here learning progress. This can be directly checked with the quiz results. Wether the new insights or knowledge gained by playing the quizzes were applied during work, cannot be supported or checked by the application itself. (Missing: apply outcomes).

Cycle triggered by reflection questions: While playing the quiz (plan and do work) reflection questions are presented, with the goal to directly trigger reflection (initiate reflection). Depending on the type of the reflection question (at the beginning, during or at the end of a quiz play), a reflection session might be conducted to, for example, reflect about individual learning progress or relate working experiences to the content of the quiz. Insights and outcomes can be directly inserted into the quiz by answering these questions during the quiz play. Whether the users actually apply their gained insights or outcomes during work cannot be checked by the quiz itself. (Missing: apply outcomes).

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

Individual users use the quiz to check their newly gained knowledge, reflect on their current knowledge, keep-themselves up-to-date and relate content-based questions to working

situations or experiences. Although the quiz might also have potential to initiate collaborative reflection, when providing comparison possibilities of the gained quiz results, usage statistics or learning progress or some kind of competition functionalities (e.g. knowledge battles) for motivational purposes, these features are not supported yet. However, for the individual participants it was possible to discuss their gained insights and outcomes with their colleagues during the workshop or at the stroke unit as well as to apply them during work.

Regarding the transfer model (Figure 10.3.1) it can be stated that the content-based and reflection questions of the quiz serve as trigger for reflection (Step 1). Answering these questions could lead to a consecutive recursive reflection process (Step 2). The gained insights or outcomes can also be stored within the quiz (by answering the corresponding questions). Applying these insights our outcomes during work (Step 3a) is not directly supported by the application but leaves the task to the user.

Since the quiz was designed to supports individual reflection, push- or pull mechanisms to initiate communication and collaborative reflection do not apply.

5.8.3 Research approach

Design and procedure

Evaluation during the qualification program (QP): The Medical Quiz was introduced to the participants of the qualification course „Spezielle Pflege auf Stroke Units“ in the first week of the course. The presentation for the introduction was prepared by KNOW and the introduction was held by the NBN project leader of MIRROR in the mid of October 2013. During the introduction the participants were asked to fill in the consent form, the demographic questionnaire and general questionnaire. Afterwards the participants got an introduction to the quiz itself including a neutral account. From the introduction of the quiz to the next qualification program week (middle of November 2013), the participants were asked on the one hand to play the different quizzes as often as possible and on the other hand to create new questions for the quiz. During this time all quiz attempts including all quiz results and answers to the reflection questions were logged. In the second week of the qualification course, the clinic representative presented the results of the first filled-in questionnaires and the number of quiz accesses. Then participants filled in an in-between questionnaire and were asked again to play the quiz and create questions for the quiz until the third course week (first week of December 2013), in which the same procedure was repeated. In the fourth week of the course (middle of January 2014) two researchers from the KNOW visited NBN and conducted a workshop with the study participants. During the workshop the overall statistics were presented to the participants, group discussions and some interviews were conducted and a final questionnaire was distributed.

Altogether the evaluation lasted from October 2013 to January 2014. The quiz itself was played from the middle of October until the beginning of December. Participants used their spare time to play the quiz and create questions.

Evaluation at the stroke unit (SU): The evaluation at the Stroke Unit followed a similar design and procedure as the evaluation conducted during the qualification program. However, the following deviations were necessary: Due to the lack of an internet connection at the stroke unit for security reasons and because there are only few computers on a ward which are used for the patients’ documentation, the quiz was installed directly on a notebook, which was made available to the participants in their staff room. The quiz was available for about 7 weeks in February and March 2014. The participants were asked to play the quiz as often as possible

during their working time (e.g. during silent night shifts). They received only one in-between questionnaire after four weeks of the trial, but also the final questionnaire at the end of the evaluation. It was planned to conduct interviews with the participants, but no nurse agreed to be interviewed regarding their experience with the quiz.

Baseline data regarding general reflection was collected for both trials.

Participants

QP evaluation: 21 nurses participated in this evaluation, 2 men and 19 women; 66% were aged from 20-29 and 33% between 30 and 59. 95% were nurses and 5% were head nurses. The average time in the current position was 4 years.

SU evaluation: 3 nurses participated in this evaluation, 1 man and 2 women, all of them aged from 20-29. The average time in their current positions was 1.5 years.

Summative evaluation methods used

We used for both evaluations nearly the same procedure. The main difference is that at the end of the trial we conducted a workshop with group discussions for all participants of the qualification program.

Demographic Information: The pre-questionnaire contained all demographic questions from the toolbox, except for date and Team-ID, because there were no teams in this setting.

Level 1: Reaction: all usage data can be received from the logs of the Medical Quiz.

Level 2: Learning: The short reflection scale (CR) was presented in the pre- and post-questionnaires. Additionally, the post-questionnaire contained 8 app specific reflection questions (CA) which fitted best to our approach and the two Learning Outcome questions (CL1).

Level 3: Behaviour: We adjusted the core behaviour question (CB1) to the Medical Quiz by asking how much the Medical Quiz helped to improve one's work at the stroke unit.

Level 4: KPIs were measured by several questions, encompassing work improvement in general, employees' work satisfaction (KPI), improvement of work performance (WK), patient satisfaction (KPI), and loyalty metric.

Additionally, with the pre-questionnaire we measured IT attitude (TI), learning strategies (LS), and posed questions about the work at the stroke unit in general (4 questions asking how participant perceive their work-relevant knowledge, problem solving abilities, patient satisfaction, and the quizzes' potential to improve their knowledge), and upcoming expectations regarding the quiz. The two in-between questionnaires covered subjective usage of the quiz and experiences with the quiz. The post questionnaire contained – in addition to the core questions - learning effect (GAE), self-efficacy, and questions regarding the future of the quiz.

Furthermore, the workshop with the participants of the qualification program was used to get deeper insights in participants' experiences with the app by means of interviews and group discussions. The gained insights concern for the main part evaluation levels 3 and 4, covering the following topics: for the interviews usage barriers, awareness of knowledge and behaviour change, transfer to job, satisfaction, and quality of work; and for the group discussions reflection guidance, influence into the organisation, potential for the future, and the quiz's relation to the CSRL model.

5.8.4 Results

The following results refer only to the QP evaluation of the Medical Quiz. The results for the evaluation of the SU are presented in section 5.8.5 'Additional results of the evaluation at the Stroke Unit'.

5.8.4.1 Level 1: Reaction (Usage)

The introduction of the quiz took place at the 14th October 2013. From this date on until the 2nd of December 2013, the quiz was regularly used by the participants. Based on the collected log-data, Figure 5.8.1 shows the application usage during this period. The four bar groups refer to the four different types of quizzes offered to the participants.

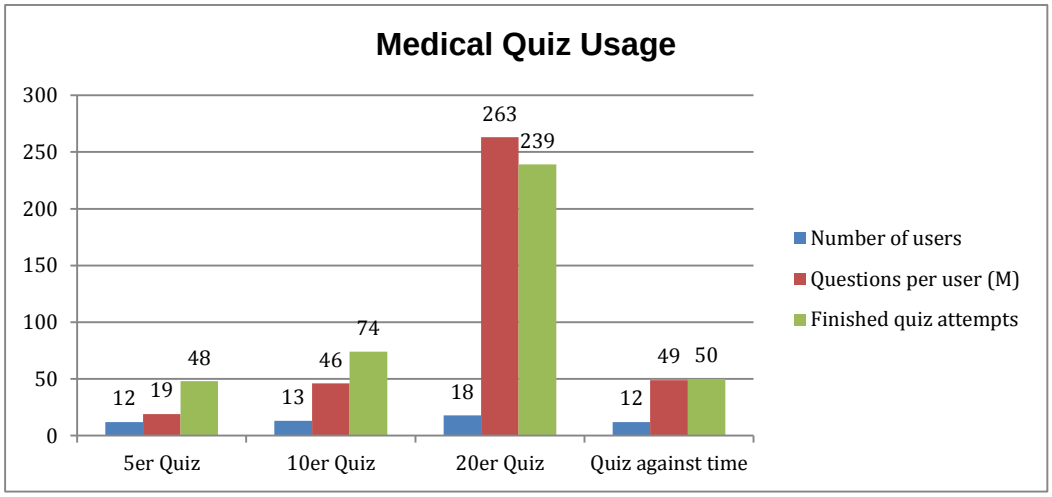

Figure 5.8.1. Medical Quiz Usage

The "Number of users" shows, that the 20er quiz was played by 18 different users, the other quizzes by only 12 or 13. "Questions per user" represent the average number of questions each user has answered per quiz type. This value varies from 18.5 (*SD = 27.8*) questions for the 5er Quiz to more than 263 (*SD = 290.9*) questions for the 20er quiz. The average questions answered for the 10er Quiz and the Quiz against time were nearly the same (M = 45.5 (*SD = 61.3*) – M = 48.8 (*SD = 70.5*)). The "Finished quiz attempts" show how many quizzes per type have altogether been played and finished during the evaluation (summed up across all users). Three participants never played the quiz due to different reasons discussed below. Thus, all users who have tried out the Medical Quiz have played the 20er Quiz. Altogether for the 20er quiz 239 finished quiz attempts were counted, followed by 74 finished attempts of the 10er quiz. The finished attempts of the 5er Quiz and the Quiz are almost identical with 48 vs. 50.

During this evaluation, the participants have answered altogether 8314 questions. While one user answered more than 1120 questions by playing the 20er quiz, other users only answered 24 questions during the entire evaluation period. Overall, the results regarding the usage of the app show that most of the participants have played the quiz very often, which indicates that they perceived it as useful; otherwise they would not have played it that often.

Barriers for using the app: Two of the three participants, who have not used the quiz at all stated in the questionnaires that they had no internet access and one of the participants did simply not bring herself to try it out. The following other barriers for not using the quiz or for using the quiz in the beginning of the trial more than in the end were mentioned in the interviews: a lack of time especially in November and December, a loss of interest in the quiz after passing the exam and because of too many recurring questions and too little user-

friendliness especially for nurses with lack of computational skills. On the other hand, several statements included the wish, to have the quiz directly available at their ward especially for the night shifts.

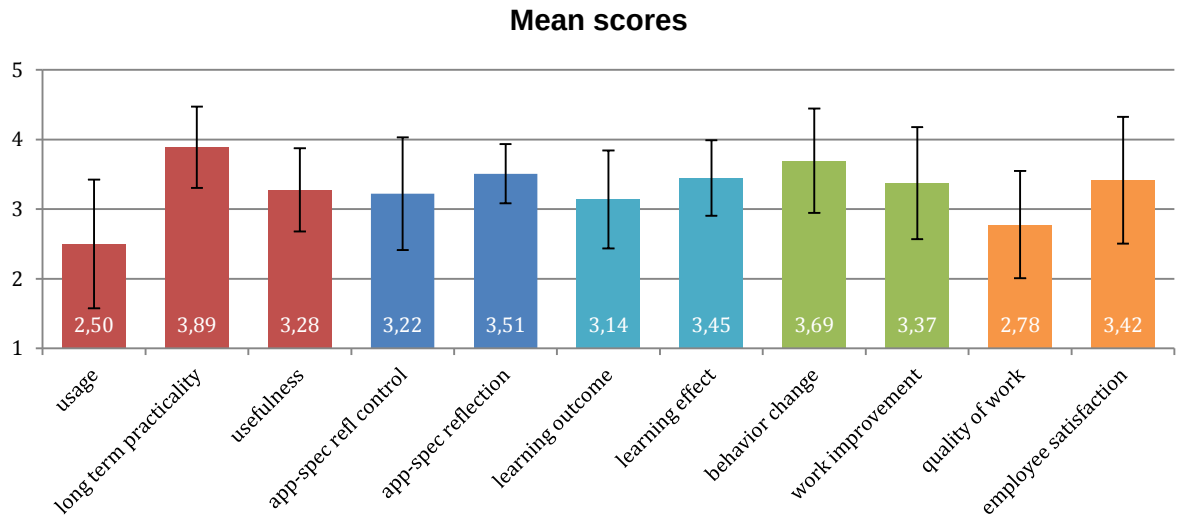

Figure 5.8.2. Mean ratings (SDs) after using the quiz (from 1-totally disagree to 5 – totally agree); different colours indicate different evaluation levels (starting with level 1 on the left hand side of the figure).

Further data concerning level 1 have been gathered via the questionnaires. Figure 5.8.2 shows the mean ratings (from 5pt. scales) of summarized values of the post questionnaire. The red bars on the left hand side present the values for “Level 1: Reaction” of Kirckpatrick’s model. The four blue bars represent the “Level 2: Learning”. The green bars can be related to “Level 3: Behaviour”, while the orange bar shows the result for “Level 4: Results” with regard to the measured KPI. Each of these values will be described in detail in the corresponding section of this deliverable.

The first three red bars of Figure 5.8.2 relate to “Level 1: Reaction”. Regarding the subjective estimated usage frequency of the application, the participants rated their individual usage frequency rather low (M = 2.5, *SD = 0.92*). If we compare the subjective impression of the participant’s with the real usage retrieved from the log data, we are of the opinion that some of the participants underestimated their quiz usage. Interestingly, there is no correlation between subjective and objective usage in terms of number of questions played (r=.324, p=.19, N=18). Most of the participants agreed to long-term practicality (M = 3.89, *SD = 0.58*), which means that the application can be used to complement professional training for nurses. The usefulness of the quiz (M = 3.28, *SD = 0.60*) was rated neutral or at least slightly positive from most of the participants. This attribute indicates that the users see the long-term advantage of the quiz during work as well as that they are interested in using the application during work. We found no correlation among the four “level 1” variables objective and subjective usage, long-term practicality, and usefulness.

5.8.4.2 Level 2: Learning

The goal of the Medical Quiz during the qualification program is to provide an easy possibility for the nurses to check their newly gained knowledge, to detect possible knowledge-gaps, to make them reflect about their current knowledge, and to relate the content of the posed questions to work-related situations or experiences. Within the qualification program the

participants used the quiz to be well prepared for the course's examination, which was explicitly stated by one participant.

With regard to the nurses using the quiz directly at the stroke unit, the quiz should make them aware about their current knowledge and refresh it if necessary, relate the content of the questions to work-related situations or experiences but also provide a possibility to keep-up-to date with new treatments and medications.

Learning Process

CSRL model and reflection questions

Relating the Medical Quiz to the CSRL model, playing the quizzes corresponds to the "Plan and Do work" stage with the goal to learn something or get insights for the practical work. Especially with the help of the implemented reflection questions, the "Initiate reflection" phase is triggered. By answering the posed reflection questions the "conduct reflection session" phase is covered including noting down any insights or outcomes. The last phase "apply outcome", is not covered directly by the Medical Quiz, because we have no possibility to check whether the gained insights our outcomes have been tried out or applied during work.

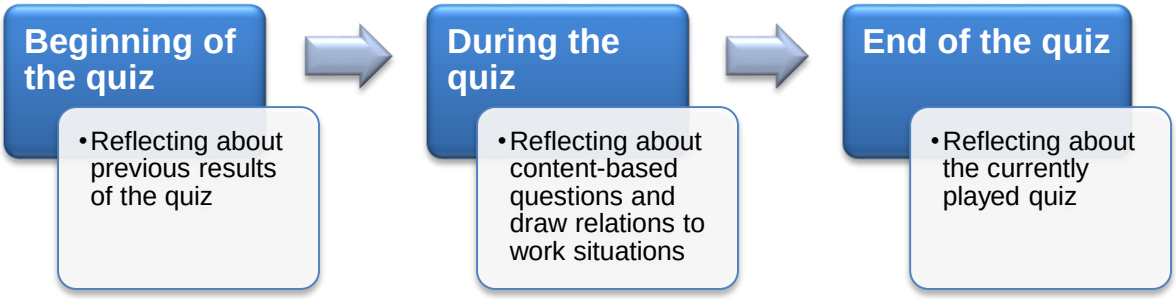

Figure 5.8.3. Reflection guidance components

In order to better support the "initiate reflection" phase during the quiz play, we implemented a part of the reflection guidance concept (see D4.3, section 5). We added reflection questions (open questions) additionally as well as on top of the usual content-related quiz questions at the beginning, during and at the end of a quiz. At the beginning of the quiz (implemented in all quiz types) the reflection questions make aware of the current knowledge status, depending on the gained quiz results on previously played quizzes. Additionally, participants are asked to think about the reasons for their past results and, in the case of low knowledge, levels how they can improve in the future (see Table 5.8.1 for some example questions). The in-between questions (only available in the 20er Quiz), put the focus on the content-based questions and how they refer to past working situations and working experiences (see Table 5.8.3 for examples). The questions at the end (20er, 10er and 5er Quiz) of the quiz asked explicitly for gained insights or new knowledge with regard to the currently played quiz (see Table 5.8.2 for examples). After having finished a quiz, the quiz results are directly presented to the players, which again is worth reflecting on.

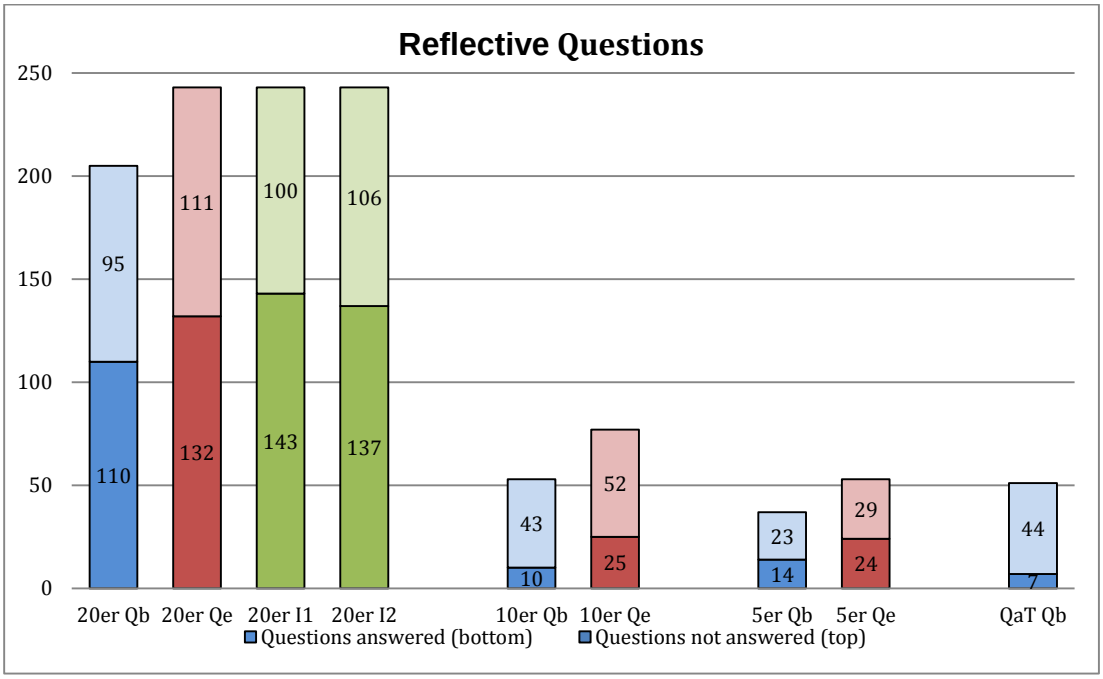

Figure 5.8.4. Number of presented (full scale bars) and answered (bottom part) reflection questions: split into questions shown at the beginning (Qb), during (I1, I2) and at the end (Qe) of the quizzes.

At the beginning of each quiz, one entry reflection question was automatically posed, which was based on the previous quiz results of the current user. These questions were displayed as soon as the user has played the corresponding quiz for three times. The questions referred to the current knowledge state of the user and motivated the user to think about how the quiz might influence or support the user's learning. Altogether we had 9 different reflection questions. Examples of these questions are presented in Table 5.8.1.

For the 20er-Quiz, altogether 205 reflection questions at the beginning (blue bar) were shown and more than 50% were answered by the players (Figure 5.8.4, bar "20er Qb", bottom bar). For the 10er Quiz only 18% of the 53 posed questions were answered (Figure 5.8.4, blue bar "10er Qb"), for the 5er Quiz 38% out of 47 questions (Figure 5.8.4, blue bar "5er Qb") and for the Quiz against time 13% were answered out of 51 questions (Figure 5.8.4, blue bar "QaT Qb").

Table 5.8.1. Summary of the reflection questions posed at the beginning of all types of quizzes.

Question at the beginning	Frequency of occurrence	Frequency of answers	Answers
Your knowledge is very constant. How could the quiz help you to learn?	65	26	Practice, Nothing, Yes, Repetition
Unfortunately your state of knowledge is very low. In what respect does the quiz help your to learn?	59	37	Learning, Yes, Practice, Repetition, Retain Knowledge
Unfortunately, your state of knowledge is very low. What is your success recipe?	63	35	Yes, Learning, Repetition, Practice

Table 5.8.1 shows three possible reflection questions posed at the beginning of all quizzes. The “Frequency of occurrence” shows how often which of these questions were shown during the quiz and “Frequency of answers” shows how many of them were meaningfully answered. (Unfortunately some participants only inserted some letters, in order to get the question counted as correctly answered.).The “Answer” column shows a summary of the words most often occurring in the answers.

At the end of each quiz, except the Quiz against Time, a reflection question was presented with the goal to motivate the users to reflect about the currently completed quiz and if they could gain any benefits or insights out of the currently played quiz. These reflection questions were chosen randomly out of altogether 8 questions. Examples of these questions are presented in Table 5.8.2. At the 20er Quiz 243 of these reflection questions were presented to the players and 54% of them were answered (Figure 5.8.4, red bar, 20er Qe). For the 10er Quiz 77 questions were presented and 32% were answered (Figure 5.8.4, red bar, 10er Qe), and for the 5er Quiz the users answered 45% of the 53 posed questions (Figure 5.8.4, red bar, 5er Qe). In the Quiz against time no reflection question was presented at the end.

Table 5.8.2. Summary of the reflection questions posed at the end of three types of quizzes.

Question at the end	Frequency of occurrence	Frequency of answers	Answers
Reflect on the currently played quiz. Have you perceived any special insights for yourself?	54	26	Yes, Retain Knowledge, No, Repetition
Reflect on the currently played quiz. What do you intend to do with regard to the quiz results?	53	30	Learning, practice, use theory in practice
Reflect on the currently played quiz. In what respect does the quiz questions support you to learn for the qualification program?	44	23	Yes, very much, recognise progress and repetition of the learned knowledge

Table 5.8.2 presents three of the posed reflection questions at the end of the quiz. “Frequency of occurrence” shows how often the question was posed at the end of the quiz and “Frequency of answers” presents how many of them were filled in. The last column contains often received answers.

The two in-between reflection questions were only added to the 20er Quiz. Nine different reflection questions were randomly presented. Their task was to motivate the player to relate

the presented content-based question to possible situations during work. Examples of these questions are presented in Table 5.8.3. For both in-between questions, the participants had also answered more than 50% of the presented questions (Figure 5.8.4, green bars, 20er I1 and 20er I2).

Table 5.8.3. Summary of the reflection questions posed during the 20er quiz.

Question during the 20er quiz	Frequency of occurrence	Frequency of answers	Answers
Does the question above remind you on an interesting situation/discussion during your work and if yes, on which?	51	23	Yes, No
In what respect is the above mentioned knowledge relevant for your work?	68	40	Yes, Very relevant
Could the question mentioned above support your learning? If yes, how? If no, how should this question be chanced, in order to support your learning?	60	35	Yes, No it's too easy

Table 5.8.3 presents three of the in-between reflection questions posed during the 20-er quiz. "Frequency of occurrence" shows how often the question was posed during the quiz and "Frequency of answers" presents the number of how many of them were filled in. The last column contains reoccurring answers.

From the evaluation of the log data we saw that the guidance concept was accepted and more than 50% of the reflection questions were answered in a meaningful way. Nevertheless from the interview and group discussion conducted, we also received some further feedback regarding the reflection questions. Answering or not answering the reflection questions should not have any influence on the quiz result. Sometimes the questions were found as disturbing during the quiz play, referring especially to the in-between questions, because the randomly chosen questions did not always fit very well to the content-based question above. When introducing the quiz to the workshop participants, the sense of the reflection questions should be explained in more detail.

General and app-specific reflection

Referring back to Figure 5.8.2 (blue bars) one can see that the participants of the qualification program only slightly agreed that the application triggers reflective learning (M = 3.51, SD = 0.42). This is in line with the general impression that participants viewed the quiz mainly as learning support and that reflection was only of secondary importance. The result is not surprising considering that these participants played the quiz in the context of a training course. However, correlating the app-specific reflection ratings with usage data, shows that participants with higher objective (number of played questions) or higher subjective usage (rating) also showed higher ratings concerning the app's potential to support reflection (r =.712/.535, p=.001/.022, N=18 for objective/subjective usage).

The "app-specific reflection control" question asked whether the sharing of experiences is supported within the Medical Quiz. With a mean rating of M = 3.22 (SD = 0.81) participant also slightly agreed, which was rather surprising, because no sharing at all is implemented within the application. We believe that the participants misinterpreted the question by referring to sharing experiences "about the quiz" instead of sharing learning/working experiences "within" the quiz.

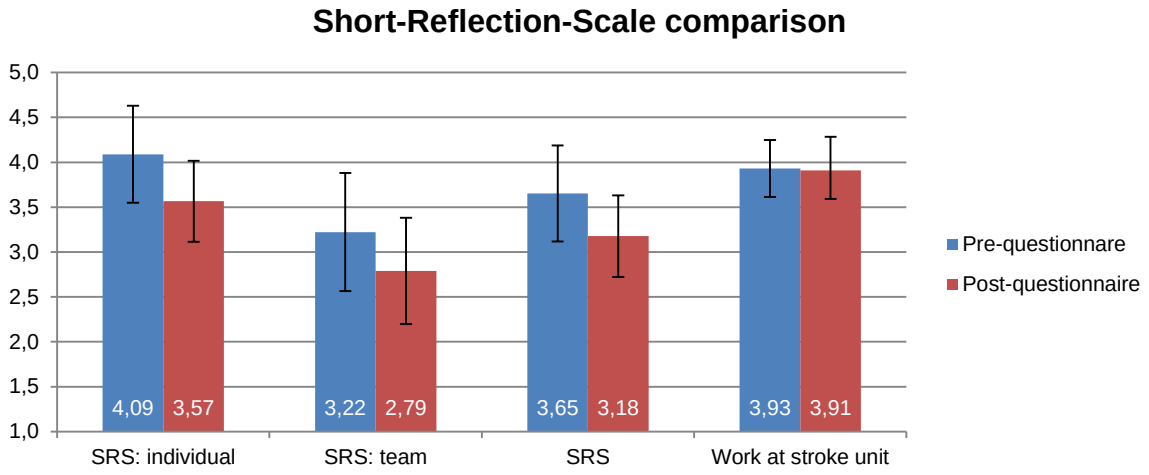

Figure 5.8.5. Short Reflection Scale and "Work at a Stroke Unit" before and after playing the quiz

Figure 5.8.5 shows the mean ratings obtained for the Short Reflection Scale (SRS) as well as the two subscales concerning individual and team reflection only. Comparing the scores (SRS) of the pre- and post-questionnaires, Figure 5.8.5 clearly shows that the general tendency to reflect decreases significantly for the overall scale, as well as the two subscales (for all 3 comparisons, related t-test revealed significant differences with all $p \leq .003$ and N=18). One reasonable explanation of this phenomenon is, that at the beginning of this evaluation all participants thought that they were rather reflective practitioners. However, after becoming aware of how reflection is defined within the MIRROR project, they might have changed their understanding of the concept of reflection as well as their reflective practices. Further comparisons of the post-SRS values with usage data and other learning outcome scales (see below) show a positive relationship between high individual reflection and perceived usefulness (r=.522, p=.026) as well as long-term practicality (r=.536, p=.022) of the quiz. Otherwise, there was no relationship between usage, learning outcome, or behaviour change.

Learning Outcomes

Regarding our implemented reflection guidance concept, by presenting reflection questions at the beginning, during and after a quiz play, we have evidence that reflective learning can be triggered. 52% of all posed reflection questions were answered in a meaningful way, which proofs that the quiz players have at least thought about the posed question. Some of the answers, which give more insights about the player's thoughts, show clear insights or benefits for the individual player. Answers like "I can recognize my state of knowledge by answering the questions several times and enhance my knowledge accordingly"[19] or "I partly better understand medical orders"[20].

In the interviews participants also indicated, that they could change their state of knowledge and their learning behaviour with the help of the quiz. They liked that the quiz was integrated in the qualification program and it increased the motivation to learn.

Referring back to Figure 5.8.2 (cyan-coloured bars) the learning outcomes stated within the post-questionnaire were seen as nearly neutral M = 3.14 (SD = 0.70), i.e. participants could not decide whether they gained a deeper understanding of their work-life and what to change

[19] German Original Statement: "ich kann durch vielfaches Beantworten der Fragen meinen Wissensstand erkennen und das Wissen entsprechend erweitern"
[20] German Original Statement „ich verstehe ärztliche Anordnungen teilweise besser".

about their work behaviour. The average learning effect, which was assessed by 12 questions, was rated as slight agreement (M = 3.45, SD = 0.45). Regarding the single question mean ratings range between M = 2.78, (SD = 0.81) for “talking about the quiz helps me to reflect upon my learning behaviour” and M = 4.22 (SD = 1.06) for “the quiz supported me when preparing for the exam”. Again, the agreement was higher for questions concerning the gain of new knowledge than for questions concerning the reflective behaviour itself.

5.8.4.3 Level 3: Behaviour

The results presented in Figure 5.8.2 (green bars) show that the behavioural change was rated with M = 3.69 (SD = 0.75), which implies that the participants tend to agree that the quiz helped them to improve their work at the stroke unit. Also the 5 further questions on work improvement received an average rating of *M* = 3.37 (*SD* = 0.80). Inter-correlations between the scales from the post-questionnaire show that ratings of app-specific reflection questions and learning effect (both on level 2 – learning) are positively related to behaviour change and work improvement (both level 3). Thus, participants who perceive the quiz as helpful for supporting reflection and learning are also more positive that it helps them to change their behaviour at work (all correlations show *p*- values < .01).

The interviews confirmed these values and showed that with the quiz behavioural changes have taken place and that these changes are very relevant for their future work. First, the nurses emphasized that they gained a lot of new knowledge during the qualification program and by playing the quiz, altogether too much at a first glance to be able to reflect on it. Bringing together the theoretical knowledge with their working practice was seen as very relevant for them, especially when they can use the new knowledge during work. What was also mentioned is that they have now much more background knowledge in general. And finally they reported that the more they know, the higher is their self-confidence during work. These statements are really promising results for the behavioural changes.

5.8.4.4 Level 4: Results

With the Medical Quiz we focused on the KPIs “employee satisfaction” and the improvement of “quality of work” (see Figure 5.8.2, orange bars). The employee satisfaction was rated with M = 3.42 (SD = 0.92) whereas the quality of work was rated with M = 2.78 (SD = 0.84) in the post-questionnaire. Regarding the first KPI the participants stated neutral and slightly agreed, that the employee satisfaction was positively influenced by the quiz. The quality of work, including the improvement of the medical care for the patients and to better solve problems occurring during work, was rated from neutral to slightly disagree, which means that it was nearly not influenced by the quiz play.

First, with this summative evaluation we can prove that the participants liked the Medical Quiz and gained a lot of new knowledge relevant for their work.

Second, we can prove reflective learning was initiated. With the help of the implemented reflection guidance in the form of reflective question we could show that reflective learning was initiated, when analysing the answers of the reflection questions.

Third, by relying on the statements of the interview, we got first evidence that participants changed their behaviour according the newly gained knowledge and the insights they got from reflection. Unfortunately we have no further or deeper insights on this, neither in the quiz itself nor in the statements from the interviews.

The evaluation of the Medical Quiz within the qualification program has no direct influence on an organisational level, because the participants came from different stroke units spread all over Germany. But in the group discussion about the possible organisational influence they saw high potential for the quiz. Possible influence on the organisation was seen with regard to the optimisation of patient care, improvement of the quality of patient care and improvement of the employee satisfaction. Having more theoretical knowledge leads to a better understanding of their work, improves the quality and finally results in a better patient satisfaction. In order to achieve this with the Medical Quiz, there need to be more practice relevant questions within the quiz, real case studies with corresponding questions and more questions in general referring to practical work.

From the group discussion with respect to the future development or features of the quiz, we got a lot of suggestions. These encompass different difficulty levels, rewarding system, over knowledge battles nurse vs. nurse or clinic vs. clinic, until to the creation of a Facebook group or other type community. As prerequisite it is necessary to maintain the quiz by adding consequently new questions (removing outdated questions) and motivational features, in order to keep the motivational question to use and learn with the quiz constantly high.

The results of the loyalty metrics looks as follows: 5.6% are promoters, 38.9% are passives and 55.6% are detractors. That implies a net promoter score (NPS) of -50%.

5.8.5 Additional results of the evaluation at the Stroke Unit

Below we give only a short summary of the evaluation directly conducted at the Stroke Unit, for only three users were involved in this evaluation and they did not play the Medical Quiz very often. The analyses are based on the log data of the Medical Quiz as well as the pre-, in-between and post-questionnaires. No participant agreed to be interviewed regarding their experience with the quiz.

Level 1: Reaction (Usage): Three participants took part in this trial at the SU and had the possibility to play the quizzes during their shifts at the SU for 6-7 weeks. Due to a technical problem, the quiz was not available for about one week during the evaluation period and the offline version of the quiz was very slow, which reduced the motivation to play the quiz for the participants. Additionally, each of the three participants took a holiday of 2 weeks during the evaluation period. Thus the actual evaluation time was about 3-4 weeks. Altogether the three participants played the “10er Quiz” once and the “20er Quiz” ten times. Overall they answered 210 content based questions, on average $M = 70$ ($SD = 26.46$) questions.

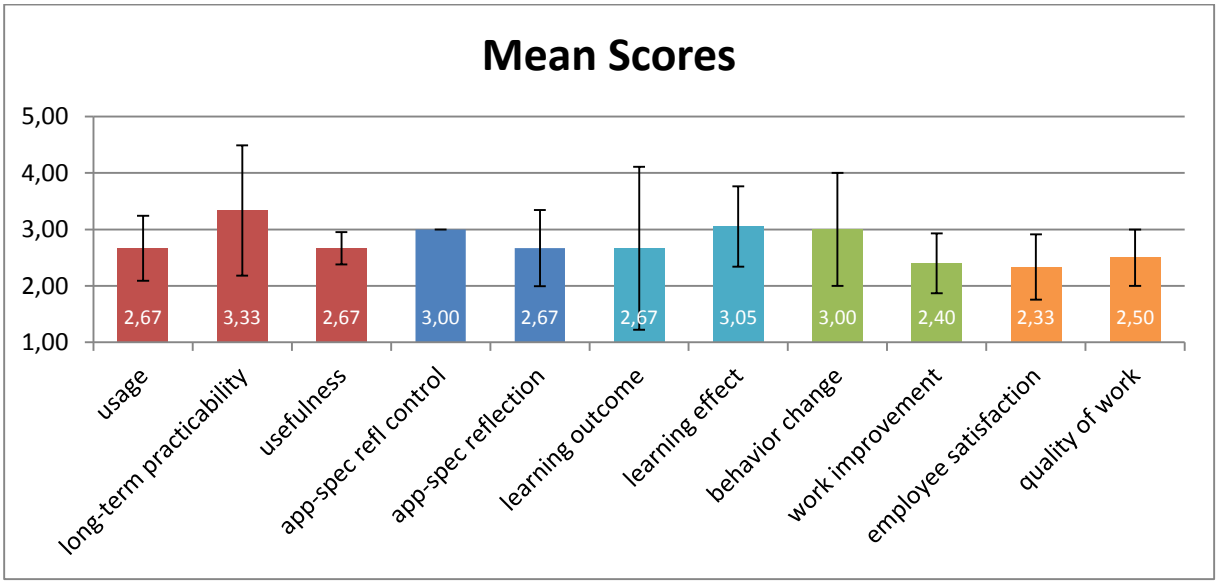

Figure 5.8.6. Mean ratings (SDs) after using the quiz (from 1-totally disagree to 5 – totally agree); different colours indicate different evaluation levels (starting with level 1 on the left hand side of the figure).

The first three red bars of Figure 5.8.6 relate to “Level 1: Reaction”. Asked, whether they played the quiz very often, participants subjective estimated usage frequency of the application was rated rather neutral, while the objective usage of the participants was rather low (compared to the QP evaluation). The participants rated the long-term practicability neutral or at least agreed to it, while there was slight disagreement regarding the usefulness of the application.

Level 2: Learning: Altogether 36 reflection questions were presented to the participants at the beginning, during and at the end of the quiz. Only 22% of these questions were answered in a meaningful and useful way. From the in-between questionnaires we know, that the participants did not perceive the reflection questions as very useful to initiate reflection.

Referring back to Figure 5.8.6 (blue bars, app-spec reflection) one can see that the participants slightly disagreed that the application triggers reflective learning. The “app-specific reflection control” question, asking whether the sharing of experiences is supported within the Medical Quiz was rated neutral, i.e. even higher than the app-specific reflection questions. This result indicates that the obtained responses have to be interpreted with much caution, since it is not clear whether the participants only misinterpreted the control question or if they did not put too much effort and seriousness into answering the questions, in general.

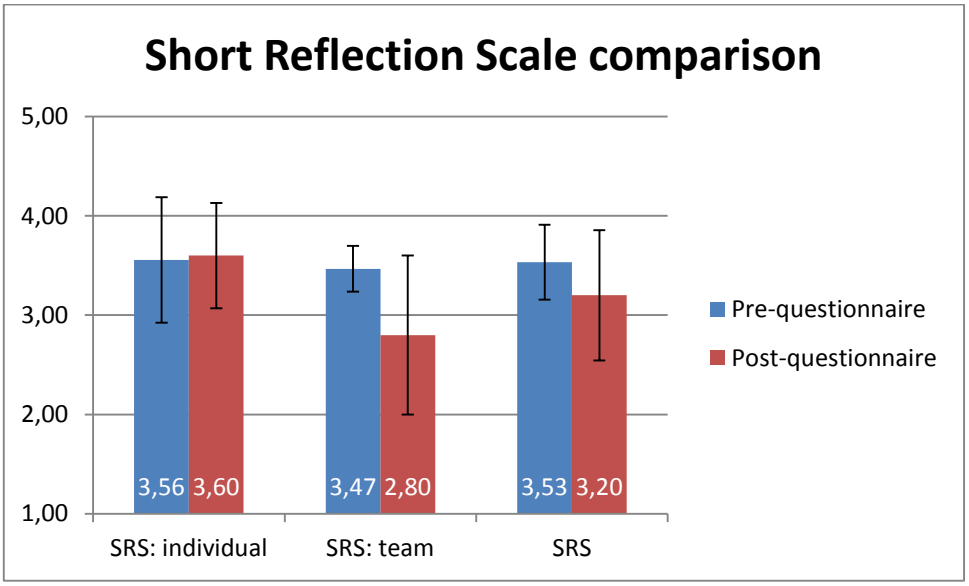

Figure 5.8.7. Short Reflection Scale before and after playing the quiz

Figure 5.8.7 shows the mean ratings obtained for the Short Reflection Scale (SRS) as well as the two subscales concerning individual and team reflection only. Comparing the scores (SRS) of the pre- and post-questionnaires, *Figure 5.8.7* clearly shows, that the general tendency to reflect decreases for the team scale, which also effects the overall SRS score. There is no change in individual reflection.

In Figure 5.8.6 (cyan-coloured bars), the mean learning outcomes stated within the post-questionnaire were rated with slight disagreement, while the learning effect was rated neutrally. Again, the agreement was higher for questions concerning the gain of new knowledge than for questions concerning the reflective behaviour itself. For both learning outcome and learning effect a closer look at the individual responses shows, that 2 participants rated both aspects rather positively (with individual means of 3.5 for outcome and 3.4 and 3.5 for learning effect), whereas the third person perceived no learning outcome at all (both items rated with 1) and a mean learning effect of 2.2 (across 13 questions). Interestingly, the person with the lowest agreements regarding learning outcome and effect shows also the lowest score on the short reflection scale (2.5 as compared to 3.3 and 3.8 for the other two nurses).

Level 3: Behaviour: The results presented in Figure 5.8.2 (green bars, behavioural change and work improvement) show that the behavioural change was rated neutral, which implies that the participants did not agree but also did not disagree to take any behavioural changes. The five further questions on work improvement showed that the quiz has no influence to improve their work.

Level 4: Results: With the Medical Quiz we focused on the KPIs “employee satisfaction” and the improvement of “quality of work” (see Figure 5.8.6, orange bars). Both, the employee satisfaction as well as the quality of work were rated negatively in the post-questionnaire. This means that neither the employee satisfaction nor the quality of work, including the improvement of the medical care for the patients and to better solve problems occurring during work, were influenced by the quiz play.

Because of the missing interviews, we have no evidence of why the quiz was not perceived as useful as in the setting of the qualification program. Explanations could be, that the offline version of the Medical Quiz was very slow on the notebook, that some technical reasons made

the quiz inaccessible for a while or also that within such a setting more guidance and feedback for the participants is necessary with regard to reflective learning.

5.8.6 Conclusion & Discussion

In general the results of the evaluation showed that the Quiz was well accepted by the workshop participants, most of them really used it very often, especially to prepare themselves for the workshop examination. They perceived the quiz as useful and they also stated that, if the quiz would be extended and maintained, it would be great to have it available during the work especially during the night shifts. They also mentioned that the quiz brings in more motivation as well as fun aspects with regard to the learning of new knowledge. In contrast, they also had some ideas for improvements. Although we had in the end of the trial more than 150 different content-based questions, the participants stated that there were too few of them in the quiz and that they reoccurred too often. They also mentioned that they would like to have not only content-based factual questions but also case studies from real work situations and corresponding questions to it. And they also would like to have some difficulty levels, which might bring more motivational aspects into the quiz.

By integrating a part of the reflection guidance concept in form of reflective question at the beginning, during and at the end of the quiz, we are able to proof, that asking the right questions at the right moment can trigger reflective learning. The participants mentioned that they were able to gather new knowledge with the help of the quiz, which is in the end very useful for their work. As a result the participants mentioned that they feel more self-confident during work, because they were able to answer more of the questions posed by physicians, patients or relatives. They also stated that because of more available background knowledge, they understand the treatments and some conclusions taken by the physicians in a better way. They also confirmed by answering the reflection questions in the end of the quiz, that they gained clear benefits and insights for themselves but unfortunately these learning outcomes were not inserted into the quiz.

Clearly, there is also room for improvement with respect to the reflection questions. First, some participants stated, that the meaning of the reflection questions was not clearly evident in the beginning. Second, the questions posed at the beginning of the quiz, were sometimes inconsistently composed due to a bug in the quiz code. For example “Your knowledge is rather low. What is your success recipe?” does not really serve as a motivational question. Third, the reflection questions were randomly chosen out of a set of questions, which did not always fit especially to the content-based questions. The willingness to reflect on the reflection questions was given more with the questions presented at the beginning and at the end of the quiz. At the same time the in-between reflection questions were perceived as more disruptive during the learning process.

Some participants stated that they have tried to take some behavioural changes and applied them during their work as a consequence of using the quiz. However, since the quiz does not capture these data, we can only rely on participants’ subjective reports.

Concluding, we see that the Medical Quiz has the potential to trigger reflective learning and to facilitate combining theoretical knowledge with practical experiences. We also see a lot of possibilities how to improve the Medical Quiz, in order to become a very good exploitable for the whole MIRROR project with regard to individual reflective learning at work.

6 Evaluation reports of short-term interventions reports

This section contains the reports of all completed summative evaluations which are short-term, that is their duration was one to several days and they were more designed as a reflection campaign. 6.1 reports about an evaluation at work, whereas 6.2 and 6.3 describe training evaluations.

6.1 The CaReflect App Evaluation at RNHA

CaReflect supports reflective learning by measuring and visualising face to face interaction between carers and residents. The system builds on the proximity sensing method that was evaluated in 2012.

6.1.1 Organisational context

Test bed organisation and the organisational unit

For a general description of the RNHA care homes we refer to section 3.5. The CaReflect App was evaluated at care home S, being the site of the first technology test of the proximity sensors in 2012. The care home is a purpose-built nursing home that is specialized in dementia care. The Nursing Home consists of two floors with eleven residents' rooms on the ground floor and nineteen on the first floor. There are 20 single and 10 shared rooms. The nursing home uses an innovative approach to organize dementia care. Residents are grouped by their level of dementia to account for the specific needs at each level. As result, the care home is separated into 4 wards.

The four monitored wards consist of the residents' single sleeping rooms and the common rooms. Residents that are able to leave their bed spend most of the day time in the common rooms, either sitting at one of the tables or wandering around. Documentation takes either place in one office or in the common rooms using a laptop at one of the tables.

Test users and their job roles

Most staff members in a care home are care staff. There will be a small number of specialist auxiliary roles e.g. cook, handyman, cleaners, but most staff will be direct givers of care to residents, providing for their personal care needs, and their recreational and therapeutic activities. Care staff members are mainly 'care assistants', and are typically managed by senior care assistants, who provide regular supervision and guidance in professional matters. Care staff often work in teams, led by a senior care assistant providing support to particular groups of residents, who may be organised by location e.g. a wing of a care home, or a specialist unit e.g. for residents with advanced dementia. Hierarchically, above Senior Care Assistants are Registered Nurses, who are medically qualified staff, and are statutorily required in sufficient numbers according to a ratio of residents to staff. A care home will be led by a Registered Manager, who is often a Registered Nurse by background.

Most of the care staff, except for recently qualified nurses, are not educated to degree level and only have National Vocational Qualifications. This means that staff without formal training can be confronted with complex situations to resolve. Work is organized in 2 day and 1 night shifts with handovers; protocols document every treatment and activity.

In the target care home, 3-16 carers and nurses are working on each shift. There is always at least 1 registered nurse on each shift. Care assistants are mainly assigned to a specific ward and its residents. As residents are bound to stay on their ward, carer assistants stay on the ward as well. Lunch and dinner is served in the separate common rooms on each ward. During night shifts 2 carer assistants and 1 nurse are present. In the following, we refer to both, care assistants and nurses as carers for the sake of brevity.

Identified need and potential for reflective learning

Although often paid around the statutory minimum wage, a new care worker is expected to undertake an induction period and then training in some 13 or more mandatory areas of professional knowledge in their first two years of work, ranging from 'manual handling' to dementia care and 'end of life' care. Induction will involve the shadowing of experienced carers, as well as knowledge-based training. While e-learning is increasingly a part of this training, most training is still of the traditional small-group type with a specialist trainer presenting for a half-day or more. However, such general approaches cannot hope to cover all the variants of challenges likely to be faced by staff from their residents and their unique demands – that often requires some reflection, and some help, for example, asking experienced staff, or using creative thinking for solutions.

A growing challenge for nursing homes is the higher proportion of increasingly elderly residents suffering from dementia when admitted to the homes. This can lead to instances of challenging behaviour where the elderly people are confused and react, sometimes aggressively and irrationally, to their unfamiliar surroundings. Reflective learning on the side of the carers and nurses working in the homes is seen as a potential, as there is no one-size-fits-all solution when dealing with personalities approaching the end of their lives with their individual and complex life-histories.

Differences between residents lead to different needs. Some residents actively demand attention while others stay calm. Care staff and managers in the target care home were interested in analysing how this behaviour affects the attention that is provided by carers. Are quieter residents neglected or are demanding residents avoided?

Potential organizational impact

This was the second trial of FZI's CaReflect at care home S, being the site of the first operational trial in 2012. Accordingly the management were aware of possible uses and benefits, and set 3 objectives which determined the allocation and deployment of the limited number of sensors. These were:

- Where are nurses all day?
- How much carer contact do different residents receive? Is this appropriate? Who provides this care? How is it provided, i.e. as many short contacts or few long contacts?
- How often are staff 'doubled up'(did one resident require attention from more than one carer at the same time)?

6.1.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

CaReflect supports reflective learning by measuring and visualising face to face interaction between carers and residents. The system builds on the proximity sensing method that was evaluated in 2012. The proximity sensors are wearable devices - either in form of a badge or wrist watch - that capture the proximity between wearers of the sensors. Every 10 seconds the environment is scanned for other sensors that are worn by residents, carers or placed at important positions such as the documentation desk. The sensors store all contacts and can be read afterwards by the CaReflect system. The data provides an objective perspective on the daily interaction by quantifying the contact times. Reflective learning should be triggered by provoking cognitive discrepancies between own perception and measured data.

Every carer has to pick up a sensor before starting their shift. After the shift the sensors are returned and the collected data is presented to each carer individually by a researcher using the CaReflect system. Discussions among carers after this individual reflection session are desired but cannot be facilitated because these discussions have to take place during or in between care activities.

The selected visualizations aim to represent a shift as simple as possible. The data was shown without judging or selecting specific parts or content. Our formative evaluations with the proximity sensors have shown that carers can understand even raw data and are able to come up with own interpretations.

Relation to MIRROR CSRL Model

According to the WP3 perspective, the CaReflect app aims at collecting data in the “Plan and Do Work” stage. The data is visualized to trigger reflection and become part of the frame of the reflection session. The subsequent stages "conduct reflection session" and "apply outcome" are not directly supported by App, but it is expected that this follows from the App usage, if the visualized data provokes cognitive discrepancies. We aimed at facilitating these steps in interviews and active presentation of the data. Furthermore, this part of the reflection cycle may well be supported by other apps, e.g. the TalkReflect app.

The proximity sensor technology measures contacts between selected carers and residents. The CaReflect App was built on top of this technology to visualize the data as pie charts and timeline so that they can be used for reconstructing work experiences.

CaReflect aims at triggering a reflection session by presenting a different view on the past day. The sensors “capture data relevant to reconstructing and reflecting on experiences from work”. The contained data includes a large amount of information that might conflict with the carer’s own perception. In the first test in 2012, carers were able to understand their day from raw data. In this study, additional visualizations were included.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

CaReflect targets individual reflection by care staff by providing quantitative data on their daily work with residents. Although the app might also have potential to initiate collaborative reflection these features were not evaluated during this summative evaluation. Nevertheless, participants discussed among each other about their data and these discussions will have influenced the final feedback.

CaReflect is focused on the crucial Step 1 to initiate reflection which may trigger recursive reflection when digging deeper into the available visualizations. Step 3 was not evaluated during the short time of the study.

CaReflect provides no explicit support to support transitions between individuals, groups and organizational reflection session. However, the visualizations can be used across all levels and individual data can be used as evidence and reference for discussions.

6.1.3 Research approach

Design and procedure

The study was conducted in collaboration with FZI, Tracoin and RNHA. FZI provided the technology and support. Tracoin was trained to operate CaRefelct during a meeting in Karlsruhe. With this knowledge Tracoin conducted the study and provided the local technical support for carers. RNHA organized the study, supported Tracoin and conducted the final interviews.

28 sensors were allocated to 9 residents, 5 locations (4 medicine cabinet areas and the nurses' office), and to 44 separate carers (nurses and carers) over a 4 day period. The carers were monitored for just one shift or a number of shifts, depending on the rota, with some 3-12 carers being monitored at any one time. At the end of the shift, the data was read from the sensors and first visualizations were shown to the carer. Afterwards, Tracoin, which provided the technical support, administered 'end of shift' questionnaires. Furthermore, reports for each carer and the overall care home have been created using CaReflect. RNHA conducted overall feedback in short interviews and questionnaires one week later. For each of the available carers, around 20 minutes was spent to:

- feedback their own data,
- show some of the aggregated data slides,
- administer the end of trial questionnaire,
- ask about :
    - usefulness (for triggering reflection),
    - the usability of actual physical sensors,
    - privacy.

In a concluding interview with the manager, the organizational defined objectives were reviewed and the possible impact of the CaReflect system regarding the planned objectives was discussed.

Participants

44 staff members have worn a sensor and 40 (2 male/38 female) filled in the short end-of-shift questionnaire at least once. The majority of participants (31) are carers. 6 nurses and 3 care coordinators complete the picture of the active staff. Participants came from all age groups. The experience of carers varies as shown in Figure 6.1.1. The majority of carers (27) has less than 5 years of experience.

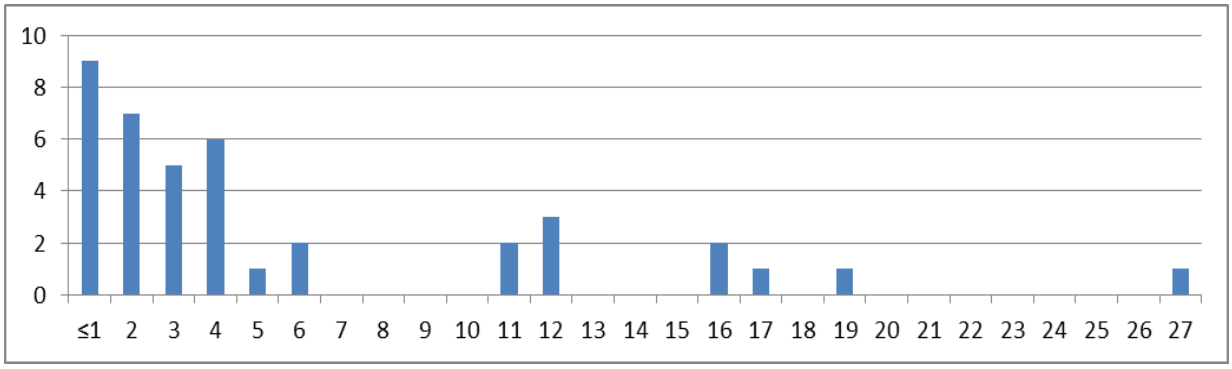

Figure 6.1.1. Care experience in years

17 of the participants were available for the concluding interviews because of the shift system.

Summative evaluation methods used

Due to the specific constraints at RNHA the methods from the summative evaluation toolbox had to be adapted to our participants. The number of questions was reduced and the wording was simplified to ensure acceptance of the questionnaire. The final questionnaire was administered as part of an interview to remove the barrier of writing long texts. Participants needed time to think and complete the open-ended questions. The end of the shift, when the carers need to pick up children, get home, etc. makes this a rush, and is unlikely to get the best responses. However, it is difficult to see easy alternatives, because the study should not interfere with care activities.

Demographic Information: The 'end of shift' questionnaire collected the demographic questions from the toolbox.

Level 1: Reaction: usage data was collected by observing carers when interacting with the system and comments made while exploring the data.

Level 2: Learning: The short reflection scale was not applicable due to the short time of the study (which matches the intended use of CaReflect). The app specific question CA12 was asked. Adapted versions for questions CL1 ("I learned something by looking at this data") and CL2 were used to measure learning outcome. We used two similar questions for CL2 in the end of shift questionnaire (CL2a "I have now an idea what I could change.") and concluding questionnaire (CL2b: "I have an idea now, how we can improve our work.").

Level 3: Behaviour: We asked participants for plans to change their behaviour because a change in behaviour is not possible within the 4 days.

Level 4: KPIs were not measured yet in the care home and would most likely not change within 4 days. Therefore, management and staff were asked about the potential impact of CaReflect. Especially the concluding interviews with the manager highlighted this aspect.

6.1.4 Results

6.1.4.1 Level 1: Reaction (Usage)

Overall, more than 45000 contacts were captured during the 4 days of the study. The overall time covered by sensors adds up to 1200 hours. Carers asked several times for more sensors to equip all residents with sensors. As visible in Figure 6.1.2 no sensors were distributed to carers during the first two night shifts.

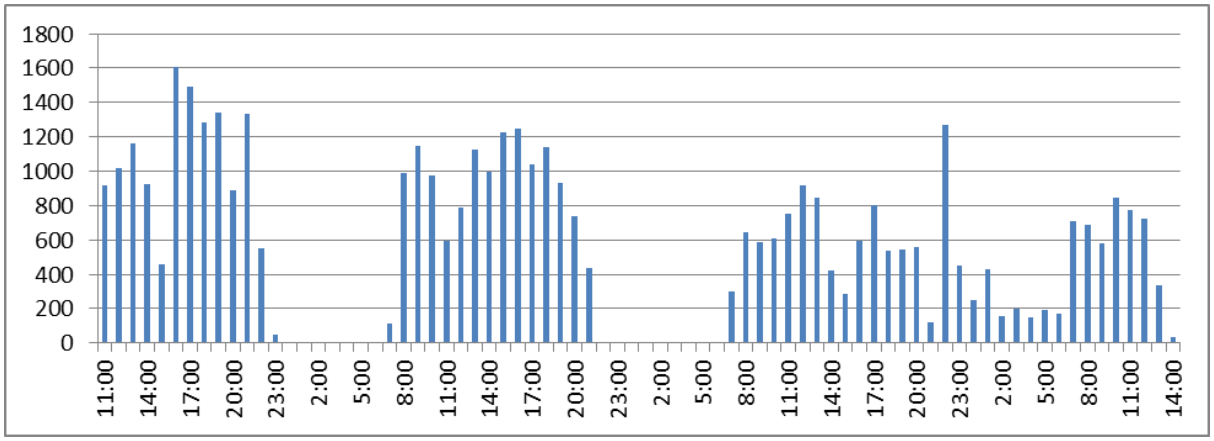

Figure 6.1.2. Contact count during study

Most observed carers were interested in seeing the absolute and relative time given to residents, spent with other staff, and at various locations, particularly "the office". A number of carers said it was difficult to remember all their contacts over an 8 hour shift, particularly when encouraged to work in the "butterfly" mode i.e. a large number of small contacts, rather than large blocks of single contact.

Observed team leaders or senior carers appreciated the overview graphs, showing the aggregated time given to individual residents from all carers. This aggregated level of data was seen as useful to reflect on performance, equity, volumes of service, frequency of contact, need and amount of 'doubling up' given for heavy, difficult, or highly dependent residents.

The vast majority of carers stated in the concluding interviews that they had no concerns about privacy and expected their data to be seen by the senior carers. Certainly this was seen as important information by the seniors to monitor individual and team performance – amounts of time given to residents, amounts of time spent with other carers, or at 'locations' e.g. the office, amount of time 'unaccounted'.

6.1.4.2 Level 2: Learning

Learning Process

CaReflect helped participants to capture data in the "Plan and Do work" stage to reconstruct work experiences. The presented data portrayed their work properly (CA12 M=3.77 SD=0.93). The data, shown on a day by day view, often stimulated the carer to provide a narrative of this specific day (e.g. "this was the day Allan died", "this was the day Doris didn't want to get up", "this was the day I spent ages in the office talking to John's daughter", etc.).

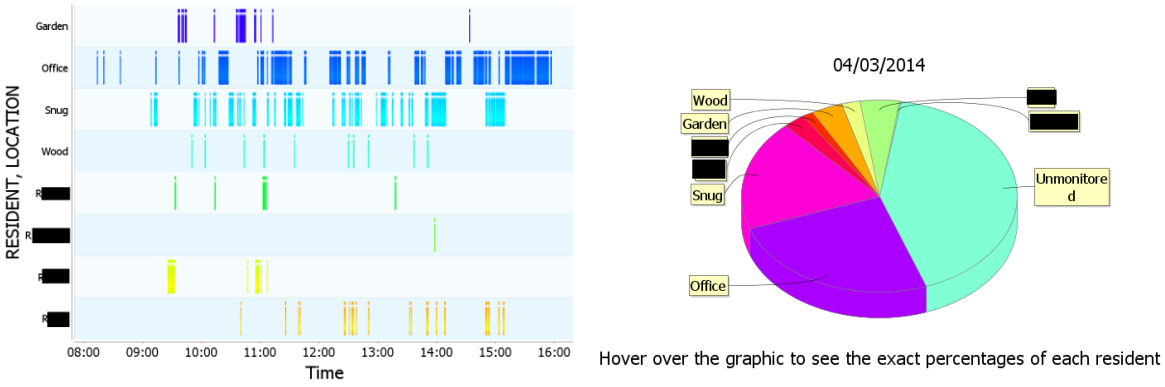

Figure 6.1.3. Screenshot presenting captured data

Carers analysed the data as depicted in *Figure 6.1.3* and were interested in answering the following questions:

1. How their time was spent differentially – more, or less, with specific residents?
2. How much time was spent with another (specific) carer?
3. How much time was spent with a specific resident?
4. How much time was spent in the office, or documentation?
5. How many different contacts were made over a shift?
6. How much time was ‘undocumented’?

Learning Outcomes

The insights according to questions mentioned above were along those lines, e.g. “I need to spend less time in the office”, ”I need to spend more time with Mrs R”, ”I seem to spend much of my time with L (colleague). I didn’t realise it was so much”, ”It’s good that I’ve spent some time with all the residents, at least some time, today” and ”I wonder why so much of my time was undocumented”.

The quantitative results of the questionnaires indicate that carers in general agree on the impact on learning (CL2: M=4.03, SD=0.55). When asked for examples of insights, carers talked about the time spent on documentation or the differences between residents. For instance, one carer was surprised how much time she needed to assist a resident with meals and wanted to discuss with colleagues about their experiences.

However, fewer staff members had directly a specific idea how they could change their own behaviour (CL1a: M=3.66, SD=0.85). If the data is split into experienced carers with at least 5 years of experience (12 out of 41) it becomes apparent that especially the experienced carers see more benefit in using CaReflect.

Table 6.1.1. Responses split by experience

Question	All (n=40)	At least 5 years experience (n=12)	Less than 5 years experience (n=28)
CA12: The graph showed my work properly.	3.77 (SD=0.92)	4.00 (SD=1)	3.68 (SD=0.88)
CL1: I learned something by looking at this data	4.03 (SD=0.55)	4.18 (SD=0.39)	3.96 (SD=0.60)
CL2: I have now an idea what I could change.	3.66 (SD=0.85)	3.77 (SD=0.91)	3.61 (SD=0.82)

6.1.4.3 Level 3: Behaviour

During the short time, carers had not enough time to change their behaviour. However, after one week we asked participants, if they now have an idea how to improve as team. There was only a slight indication that changes are planned (CL1b: M=3.35 SD=0.68).

Participants suggested in the interviews that wearing a sensor, and knowing others are wearing one, could affect a worker's behaviour; in this case, giving residents with a sensor more attention. This effect may have influenced the result of the study but can be a starting point for a long term coaching approach based on self-tracking.

6.1.4.4 Level 4: Results

The concluding questionnaire revealed that the participants were satisfied with CaReflect (SAT01 M=3.82, SD=0.6). 82% said they would like to use CaReflect again with their team. Only 24% would also use it individually. The interviews showed that carers are eager to compare their data to others and improve together.

"The overall net promoter score in an adapted way (How likely is it that you would recommend CaReflect to a friend or colleague colleague if it included more content (e.g. over a longer time span)?) was negative (NPS -29%). This result is due to the many young detractors (7) among inexperienced carers (9 of 17) that did not yet see a value in the collected data. Experienced staff members (8 of 17) were neutral (NPS= 0%). These experienced staff members include all nurses and care coordinators who work closely with the management."

In the concluding discussion with the manager about the technology, the approach was seen very positive and several ideas emerged:

- The aggregate of all individual and team reflections are seen as beneficial to the organisation as a whole, for improved performance, better overview perspectives on service contacts, volumes, timings, and for coordination purposes. Access to the level and detail provided by the sensor database addresses many organisational issues.
- The question of how often staff are 'doubling up' was seen as particularly important, because the government provides a higher level of subsidy for residents who require 1:1 care, or who regularly require 2 carers to provide proper care. Estimates are given

but they change over time, and no exact data is currently collected on a routine basis. (And are there residents who are funded for 1:1, but don't get that level of care?) The idea has been considered further to provide data to allow differentiated levels of funding/charging according to the levels of basic support required.

- CaReflect can be used to provide reassurance to relatives, and evidence of good service. It is often heard that residents will not remember staff carer visits over the previous hours, and will claim they've been alone "all day". It is also the case that for care home inspectors, "if it's not documented, it hasn't been done". Both scenarios are a source of concern for busy carers who don't have the time to write everything down. The data provided by CaReflect provides this reassurance to relatives and evidence for inspectors, showing the detail of these visits – who, when, for how long. This is seen as a major source of evidence of performance for management and individual carers – all without input.
- Finally, managers showed interest in a product based on the used technology. The issue of how to market this 'product' was raised by the manager. Is it to be sold, rented, or part of a consultation exercise? Is it for monitoring care levels for relatives, or reflective learning for staff, or used for evidence for differential payments to government?

6.1.5 Conclusion & Discussion

Participants were able to learn from the visualized data. They quickly understood the visualized context data and their narratives of events related the new perspectives to their own experiences. Carers were able to reconstruct specific situations and gained new insights, while doing so. They learned about their own work patterns as well as general organizational issues, e.g. how much time is spend on documentation. A small number had already plans to change their behaviour. However, most would like to use the app again.

The data indicates that experienced carers see more value in CaReflect. They reacted more positive to CaReflect across all items (CL1, CL2, NPS). We see two possible reasons for this difference.

(a) Experienced carers have a different role in the care home that encourages thinking beyond their own work but about the general implemented processes.

(b) Experienced carers have more knowledge that can be used to relate to the captured data.

Note that it is definitely not the intention to give everyone the same amount of attention – on the contrary, the individual needs of each resident will be different, and judged accordingly. These vary from person to person, and from day to day. When a resident was dying, they received almost constant attention, whereas others were happy on their own for periods of time. The knowledge of these individual needs is used to evaluate the CaReflect data – this is where reflective learning occurs most clearly. For example, one carer noted that a particular resident with a sensor seemed to get more attention than usual, and responded by being more alert and brighter.

The strength of CaReflect is the simplicity of the concept. Therefore, it allowed the carers, without technical background or training, to adapt the system by placing sensors not only on residents but at places that are relevant for their work. Hence, they could not only capture data about their interaction with residents but time spent on documentation. In this study the number of available sensors was limited and not all ideas could be realized. Carers strongly requested

more sensors to equip every resident and more places. This behaviour provides insights for the design of capturing solutions. They should be easy to use and adapt. If users can adapt a solution to the needs of a workplace, they can build their "own" custom solution and are more engaged.

The testing of different MIRROR Apps in the RNHA test sites have shown that the app alone will not produce reflective learning – there needs to be a time for reflection ('Conduct the Reflection Session' in our model) to understand the app's output and the user needs to reflect on their actions in the app. This applies more to CaReflect than any of the other RNHA apps – particularly due to the complexity and richness of the data (and the need for initial interpretation) and the time to understand it, from many different perspectives. However, the graphics are good and can be readily understood.

The potential of CaReflect is clearly seen by senior carers and the management. Managers in all our studies (formative & summative: overall 4) asked for a commercial version. The flexibility of the proximity sensing technology triggers many ideas for the application of CaReflect. The short term usage is appreciated. Therefore, managers ask for renting and consulting options.

6.2 The WATCHiT and WATCHiT Procedure Trainer Evaluation at Regola

WATCHiT is a wearable computer embedded in a wristband for non-disruptive data capture (e.g. stress level or time-on-task) in a crisis scene. For this evaluation WATCHiT has been used with “WATCHiT Procedure Trainer”, a smartphone app that aims at promoting reflective learning using the data collected through WATCHiT.

6.2.1 Organisational context

Test bed organisation and the organisational unit

After the initial evaluation of WATCHiT (as described in D10.2), an improved version of WATCHiT was assessed by means of a summative evaluation. The evaluation was conducted during training events simulating emergency work. The evaluation was set up with support from Regola, but the event was organized by a separate organization as part of their training activities (external to MIRROR). Events of this type are a core part of crisis and emergency workers training and are designed to resemble as much as possible real situations.

In total, the evaluation involved teams from 10 different associations. These associations are part of the large body of voluntary organizations operating in Italy in the medical, social and charitable sector. These associations (and associations of associations) vary in size and geographical distribution. For example, ANPAS (http://www.anpasnazionale.org/) encompasses 869 Associations throughout Italy, 90.000 volunteers, 400.000 members, 1.000 youths engaged in National Civil Service, 3.100 professionals and 7.000 vehicles.

Test users and their job roles

The app was evaluated with volunteer workers specialized in medical care and transportation. This is a critical sector since, according to ANPAS sources, in Italy 70% of the medical emergency interventions outside hospitals are performed by volunteers. The work is performed on a voluntary basis, depending on the availability of the individuals. There are therefore no structured workdays, as in more regular working domains. Teams are also not fixed, but are based on availability.

In this evaluation we focused on ambulance personnel working with the procedure “Soccorso Trauma”. This procedure defines the steps to perform from the moment a rescue team gets on the emergency scene in order to load the victim on a spinal board and prepare him to be transported in an ambulance. This procedure is critical for the injured survival rate and it has to be performed as quickly as possible, but without errors. The procedure is normally executed by a team of three, one acting as leader and having the responsibility to overview the execution of the procedure and keeping the patient‘s head immobilized until the procedure is completed. For training purposes this procedure is performed on a dummy or an acting injured, trying to keep the highest degree of realism.

We decided to focus on this group of workers and work procedure because this was one of the possible scenarios of use identified by respondents during the formative evaluation.

Identified need and potential for reflective learning

The work domain addressed in this evaluation is rather complex. On one side, there are strict protocols and procedures to be followed, but they slightly vary from region to region. In addition, each procedure has to be situated since different factors, e.g. the type of injury, the available equipment, the environmental conditions, might restrict the space of possibility.

To add complexity, our users are volunteers who work only part time in the emergency associations. Therefore, they miss the everyday possibility to learn and share experiences. There is therefore a critical and well-recognized need to learn from experience. Debriefings are a regular part of any large event, but they tend to focus on the organizational level and they miss data from the field. The experiences of individuals and teams are often not explicitly addressed.

Potential organizational impact

If the app would be rolled out in a large part of the organization, the main Key Performance Indicator that we would expect to be influenced is reduced error rates. Also we could expect an improved employee satisfaction because they get a precise feedback on their performance and on their ability to act properly in the given situation.

6.2.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

WATCHiT is a wearable computer embedded in a wristband for non-disruptive data capture in a crisis scene. Data captured might include information from the individual (e.g. stress level or time-on-task) and the environment (e.g. temperature, noise, location) and can be user-controlled using RFID tokens. Data might be shared with other MIRROR apps. The main difference with respect to related sensor-based approaches is that users have control over the collection of data, which is activated by the use of tags. WATCHiT is also modular and it can be configured to collect different types of data and be used in different scenarios.

For this evaluation WATCHiT has been used with “WATCHiT Procedure Trainer”, a smartphone app that aims at promoting reflective learning using the data collected through WATCHiT. In this setting, workers use WATCHiT with (RFID) tokens to collect the time taken to complete each step of a rescue procedure and self-report their errors. Then the application promotes reflection guiding users through a set of steps: (i) visualization of data captured (completion time and errors for each step) for performance self- assessment and rating, (ii) comparison of each own performance with best practices provided by experts and previous performances by colleagues, (iii) collection of notes on possible improvements

Relation to MIRROR CSRL Model

WATCHiT focuses on the first stage of the CSRL model, promoting collection of data (monitor work). The WATCHiT Procedure Trainer app is then using the collected data to support a reflection session in the field, i.e. immediately after the procedure has been performed. For this reason it focuses on a quick reflection session with easy triggers. The reflection session supported by the app is highly structured and users have to go through a set of quick steps that are designed to make people revisit their experience (with the self-evaluation screen), and then compare to optimal performance. The app aims at triggering reflection by possible divergences among the perceived performance (self-evaluation), the time that was actually used to perform each step (as captured by WATCHiT), and the optimal time. As last step, the app asks to record outcomes.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

The work on which to reflect is collaborative, with activities being highly coupled. WATCHiT Procedure Trainer app can be used for individual reflection or, preferably, by the whole team sharing the mobile, promoting discussion outside the app.

Since WATCHiT Procedure Trainer uses the MIRROR Spaces Framework, these outcomes could be shared with other applications to be used by the team as input for richer reflection sessions, for example when they are out of duties, or by the organization, for example to reflect on the training needs of their workers.

6.2.3 Research approach

Design and procedure

The summative evaluation of WATCHiT and WATCHiT Procedure Trainer was performed in two separate studies:

1) Training sessions of a voluntary organization in the Piedmont region, these are small events run monthly by local emergency organizations
2) A large training event encompassing rescue operations after a simulated earthquake, held together with a national championship where 15 registered teams from across Italy performed different procedures

The first evaluation involved 8 participants; while in the second the system was evaluated with 9 teams of volunteers, all from different associations, for a total of 27 participants. During the first evaluation WATCHiT was worn by the team leader. In the second evaluation WATCHiT was worn by the team leader only for the first team, while in the other 8 teams we gave it to the other two members of the team because they have more freedom of movement. The

person wearing WATCHiT, which for this experiment was protected by an hard shell, also wore three tokens: one programmed for signaling completion of a procedural step with no error, one for signaling step completion with minor errors and one for reporting competition with critical errors. Tags were color-coded respectively in green, yellow and red for mnemonic aid.

Participants

The first experimental group included 8 participants (6 male, 2 female). Most (71%) of the participants were between 18 and 40 and only two participants over 50, one participant didn't report his age. Among them, 5 were expert and 2 novices.

The second group included 27 participants (16 male, 11 female). Most of participants (92%), except 2, were between 18 and 49 years old. 13 were experts and 13 were novices. Experience as volunteer varied significantly in both groups, ranging from one year to more than 20 years.

Summative evaluation methods used

The first evaluation was conducted in the context of training sessions. The app developers were not in place and the evaluation was conducted by one of the coordinator of the volunteer group. This person has an outstanding experience in the area and he has been providing feedbacks on the MIRROR apps since an early stage. In this perspective, we can identify him as a *champion*.

The second evaluation was performed over two days with teams from different organizations participating to the national championship. There was no pressure on them to participate, since there was no organizational demand, though we had full support from the organizers of the event. We invited all the teams that participated in the championship and we have been available throughout the two days in a nearby stand. One team at a time was performing the procedure and we had no pre-fixed schedule, but rather tried to accommodate the time constraints and availability of the teams.

In both cases the teams filled in a pre and post questionnaire. The use of the app was also observed and video-recorded during the second event.

The questionnaires in both cases included a subset of the evaluation toolbox, namely demographical questions, application specific questions (CA), the short reflection scale (SRS), and core question about learning outcomes (CL).

6.2.4 Results

This section aggregates results from both experiments.

6.2.4.1 Level 1: Reaction (Usage)

The work performed by volunteers in emergency situations is different than work in more conventional context (e.g. office settings). It was therefore not possible to use the app more than once. Each team has used the app for a period of approximately 20-30 minutes, including wearing WATCHiT, performing the rescue procedure, and going through the collected data. In addition all the teams have filled in a pre and post questionnaire. All the participants did it on a voluntary basis, without any organizational pressure.

The respondents were overall satisfied with the use of the system (SAT01, M=4.20, SD=0.58) and perceived it as a useful tool for training (SAT02, M=4.26, SD=0.61). The respondents also agreed that the system helped them to reflect on their work (CA2, 4.06, SD=0.59). The

information collected with WATCHiT was perceived as accurate (GAE17 M=3.97, SD= 0.57), relevant (GAE19, 4.11, SD=0.53), and collection of data was effortless (GAE18, 3.91, SD=0.56).

6.2.4.2 Level 2: Learning

Learning Process

Respondents agree that the system has provided them with support for reflection in terms of relevant contents, provided help in capturing and reconstructing work experiences and guidance throughout the reflection process (CAs, M=4.13, SD=0.42). The acceptance rate is higher in the population of participants with more than 5 years of experience (M=4.26, SD=0.41), than among novices (M=3.96, SD=0.37). A higher acceptance rate is found among experts across all the app-specific questions asked (see Table 6.2.1), the performed (two-tailed) t-test shows that the difference is statistically significant (p=0.036).

Table 6.2.1. Means and standard deviations for app specific questions

ID	Question	All (N=35)* M (SD)	At least 5 years of experience (N=18)	Less than 5 years of experience (N=15)
CA2	*WATCHiT helped me to reflect on experiences*	4.06 (0.59)	4.28 (0.46)	3.73 (0.59)
CA6	*WATCHiT helped me to reconstruct a work experience*	4.29 (0.52)	4.39 (0.50)	4.13 (0.52)
CA7	*WATCHiT helped me by capturing my reflection outcomes*	4.15 (0.56)	4.18 (0.53)	4.07 (0.59)
CA12	*WATCHiT helped me by providing accurate information about my work*	4.03 (0.51)	4.22 (0.55)	3.80 (0.41)
CA40	*WATCHiT provided relevant content for reflection*	4.29 (0.52)	4.44 (0.62)	4.07 (0.27)
CA41	*WATCHiT guided me through the reflection process*	3.91 (0.61)	4.00 (0.59)	3.80 (0.68)

**two users didn't report years of experience*

In both experiments we have added in the pre-questionnaire the short reflection scale (CRs). Results show that participants have a positive inclination towards reflection both on an individual level (M=4.38, SD=0.30) and as a team (M=4.22, SD=0.38). As the evaluations were both short term, the SRS was not repeated in the post-questionnaire. In this perspective, the scale has not been used to detect changes, as it was originally designed for, but rather to understand the attitude of the user group towards reflection.

Learning Outcomes

The quantitative results of the questionnaires indicate that respondents in general agree on the impact on learning. Respondents agree that after using the system they made a conscious decision about how to behave in the future (M_{CL01}=4.09, SD=0.61) and that they gained a deeper understanding of their work life (MCL2=4.09, SD=0.74). They were also motivated to actually change their behavior (M_{BI06}=4.17, SD=0.57).

Intention to change, as recorded with an open question in the questionnaire, included e.g. need for more attention at the individual level and better coordination within the team.

The system stimulated knowledge exchange within the team (M=4.11, SD=0.41). In particular, we observed that while going through the steps of the mobile apps some of the teams discussed their performance and continued afterwards.

6.2.4.3 Level 3: Behaviour

Respondents acknowledged the system from prompting them in improving their work practices. After the use of WATCHiT participants reported that the system motivated them in actually changing their work behaviours (BI06, M=4.17, SD=0.57). Furthermore, the system made them more confident in succeeding work tasks (WK12, M=3.89, SD=0.75); this belief is more evident among experienced workers (M=4.31, SD=0.48) than among novices (M=3.46, SD=0.78).

6.2.4.4 Level 4: Results

The app was tried out by 9 of the 15 teams present at the second event. This is a very good result considering that there was no organizational requirement to join and participation was completely voluntary. One of the first teams that performed the evaluation asked us to try again in order to check whether they could improve their performance. Though this was not possible due to time constraints, we perceive this as a positive result.

After the evaluation another team leader, asked us to discuss with the coordinator of his association to present the app and introduce it also in their training.

Out of 32 participants who answered to the loyalty metric, we have 15 supporters, 8 passives and 9 detractors with an overall Net Promoter Score of 19%.

It is worth here to observe that for the first group we have 5 promoters and no detractors. This might be due to a slightly different setting, a quieter environment, or to the presence of a coordinator who has been able of conveying the benefit of the app in a more effective way. In the second group, the ones directly wearing WATCHiT have a higher median (Md=8, N=17) on the loyalty metric than the ones who have not (Md=7, N=8*). *two didn't report the value

6.2.5 Conclusion & Discussion

Participants reported a high acceptance rate for WATCHiT, no one refused to wear the prototype after it was demonstrated, or reported to have felt impeded or intimidated by the use of technology. This is an important result considering that wearable computers are not yet pervasive for the public domain as, for example, smartphones. Moreover participants were not given any pecuniary rewards.

In the evaluation of WATCHiT we observed how the data could be used to trigger new reflection cycles involving e.g. other teams or instructors. For example, towards the end of the first day one of the coordinators of the event joined us while one of the teams was performing

the procedure. When the team started their reflection session, the coordinator joined with a coaching role. The pattern was the following: the team was first discussing to set they self-evaluation and then this was discussed with the coach. We can see this process as two nested reflection cycles, within the team and team and coach. The time for each step recorded by WATCHiT and visualized in the app was used by both team and coordinator as part of their revisiting of the experience.

This however requires sharing of data. Most of the participants stated that they would have no problem to share their data. This result might be influenced by the fact that the intended use of WATCHiT is in a training context, so this reduces issues of accountability. In this perspective, the result is difficult to generalize to other contexts. It seems however that there is no clear perception of the risks connected to sharing of collected data. This should be addressed in the design and in deployment. Ways for easy sharing under user control should be defined.

In the evaluation of WATCHiT sensors where used to collect the time to perform the steps of a rescue procedure. There is a risk that putting the focus on one aspect, the others get neglected, for example, time to perform a procedure rather than quality. It is therefore important to find the right set of data to collect, one shading light on the different perspectives that one should reflect on. There is however a careful tradeoff analysis to be performed. More data on which to reflect requires more effort for both collection and analysis, increasing complexity. In this perspective, this issue is confirming the importance of the MIRROR approach, based on supporting ecologies of reflection tools that together can support reflection on different aspects at different times.

6.3 The CLinIC – The Virtual Tutor serious game Evaluations at the University of Bergamo

This and the following two sections describe evaluations of the Virtual Tutor Serious Games (CLinIC, Think better CARE, Rescue League). The 3D serious games aim at supporting reflective learning, both during and after the virtual experience, by providing tools such as tutor and overall feedback, individual reflection sessions, or a learning diary.

To avoid redundancy all aspects which refer to more than one evaluation are described only once in this section 6.3.

6.3.1 Organisational context

Test bed organisation and the organisational unit

The University of Bergamo is a State University with about 16,000 students and more than 300 PhD students. It has 6 departments and research centres which are divided into 3 areas, namely Campus of Economics and Law, of Humanities and of Engineering, and which are strictly intertwined in the town life. Thanks to the staff and the increasing number of students, it provides a dynamic environment open to innovations.

As regards the nursing bachelor degree course, it is part of the University Bicocca situated in Milan and it qualifies students to work as nurses. Among the theoretical lectures, the course includes a period of training, so that students can start practicing their professional activity.

Test users and their job roles

The test was conducted with 16 students who were attending the second semester (first year) of the nursing bachelor degree course at the University of Bergamo. It has to be pointed out that they had not started their training experience yet, as it will start in April 2014, namely one month after the test conduction, so they didn't have any experience as nurses or in the hospital.

Identified need and potential for reflective learning

For the users of the 'CLinIC-The Virtual Tutor' serious game the potential for reflective learning lies in the possibility to live real situations in a safe virtual environment, which gives participants a trigger to think back to the corresponding real life working situations.

At the moment, students have the possibility to train their ability on the field with an internship of about 6 months in one hospital as a nurse. However they don't have the possibility to be practically prepared for this experience and to share their thoughts or experiences with other colleagues in a systematic way.

The use of digital technology in the form of 'app/game' to assist reflective learning, is readily appreciated by student, who are used to the practice of reflection in supervision, and in some cases, in formal and informal reflective sessions as groups. For the first time, the new generation of nurses is increasingly comfortable with digital technologies, primarily due to the widespread use of smartphones and social networking apps.

Potential organizational impact

If the serious game achieves its objective, its usage in a care home (or for CLinIC a hospital such as NBN) would have the potential to provide new carers (nurses), who will be well trained and able to deal with difficult and stressful situations with residents (patients). From an

organisational point of view this means that new carers (nurse) will be more prepared for their work and the care home (hospital) will maintain high quality of care with less expense. In turn, this should improve the satisfaction of employees (carers/nurses) and their residents (patients).

6.3.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

The goal of this short term evaluation was to find out if the serious game has the potential (i) to trigger reflective learning on an individual level, (ii) to increase the users' awareness about the topic difficult and stressful situations with patients and (iii) to increase the users' awareness about what could happen if they were already working as nurses.

*'*CLinIC - The Virtual Tutor' is a 3D serious game focused on difficult communication between nursing staff and patients. In order to support learning by reflection, both during and after the virtual experience, a number of specific tools were added to the game: a notebook (that allows users to collect notes about their feelings during the game), game score, tutor feedback, overall feedback, individual reflection session and a learning diary. These provide support, motivation, guidance, and the opportunity to compare the improvement of performance in patient satisfaction and time management over time.

The game is a stand-alone, single player application, and users at the end of their virtual experience can follow their game history, including all the notes collected during gameplay, from their learning diary, as a stimulus to reflect about their experiences. At the end of the game users have the possibility to discuss their game experience with the support of a supervisor, with colleagues or as a team, to reflect collaboratively about their real and virtual working experiences. Furthermore users have the possibility to add their own content or experiences to the game, through a tool called 'wizard tool' (for more information about this tool, see section 3.1 in D7.3).

Relation to MIRROR CSRL Model

The 3D serious game to train nurses has been developed in order to support almost all the steps and transitions of the CSRL model:

- the 'plan and do work' step is supported through the simulation of working normally, including being confronted with a variety of real-world challenges typical of the relationship nurses-patients, but in the safe virtual 3D environment;
- the 'data transition' is supported through the generation of different types of data during the game, e.g. scores, choices made by the player, and feedback from the Virtual Tutor that are made available after the game to be used for reflection.
- the 'initiate reflection' step is supported through the possibility to take notes, receive immediate feedback after each choice made in the game and add personal content to the game. The trigger to initiate will be typical challenges, a range of options as possible responses and the discrepancy between the chosen option and the feedback from the system;
- the 'frame' transition is supported thanks to the information about insights captured during the play, scores, notes etc. that help players to identify points requiring attention and therefore setting their objectives. However, objectives are not explicitly captured in the game.

- the ‘conduct reflection session’ step is supported by game through the possibility to do an ‘individual reflection session’, through specific feedback from the Virtual Tutor during the game, and global feedback with game documentation and a learning diary available for consultation at the end of the game and after;
- the outcomes from reflection sessions (‘outcome’ transition) might be recorded as notes in the notebook provided in the game. These will be available together with the content about the playing sessions in the learning library, giving users the possibility to re-contextualize their notes if necessary.
- the ‘apply outcome’ step can be enabled by the game thanks to the availability of the game documentation used in reflection. However, this step really occurs only outside the app, when users are back in their real work environment and can apply what they have learnt with the game.
- as for the ‘apply outcome’ step, the ‘change’ transition is something that can be supported by the game (i.e. If users generate useful ideas while playing the game and collect them in the notebook or in the game through the wizard tool, they can decide to use these ideas in their work environment after the game experience and improve their work with them.) but can really occurs only outside the game when users are back in their real work environment.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

With the use of this serious game, individual and collaborative reflection levels are touched. In particular the game’s functions are able to trigger an individual reflection process, with the support of several tools (notes, feedback, score, learning diary). The collaborative reflection levels can be triggered both informally, e.g. between nurses discussing their games, or formally, e.g. during the de-briefing session after the game, or with the help of a supervisor. In this way a flexible transition between the individual and the collaborative levels is supported by the serious game.

Regarding the transfer model (Figure 10.3.1) it can be stated that some tools and challenges offered in the game serve as triggers for reflection (Step 1). Once users have finished the game they have the possibility to discuss in group (with a de-briefing session) the game experiences they had, which could lead to a consecutive recursive reflection process (Step 2). The gained insights or outcomes can also be stored within the game through the possibility to use the wizard tool in order to insert personal comments/stories/experiences/suggestions in the game. Applying these insights or outcomes during work (Step 3a and Step3b) is not directly supported by the game but leaves the task to the user.

Since the serious game was designed to support individual reflection, push or pull mechanisms to initiate communication and collaborative reflection do not apply.

6.3.3 Research approach

Design and procedure

The evaluation at the University of Bergamo was conducted at the beginning of March 2014 in a computer room that was offered by the University. The evaluation lasted about 3 hours and no control group was included in the evaluation design. A short presentation about the Mirror project and the serious game as a new tool for training was given to the students by one

researcher from Imaginary. Afterward, each student had the possibility to play the game and once finished they were asked to fulfil a post-app usage questionnaire.

Participants

The demographic of the participants who attended the 'CLinIC-The Virtual Tutor' serious game evaluation, was so characterized:

- 16 nursing students participated to the evaluation, 10 males and 6 females, 7 were aged lower than 19 and 9 were aged between 20 and 29, all of them were students from the second semester of the first year of a nursing bachelor degree course and none of them had a practical experience as a nurse.

Summative evaluation methods used

For the University of Bergamo-students evaluation the following procedure was followed:

- Demographic Information: the questionnaire contains all demographic questions from the toolbox, except for Team-ID, department and job scope because these data were not pertinent with this target.
- Level 1 Reaction: all usage data can be received from the logs of the serious game. Further some other questions were added here in order to better investigate general satisfaction and usefulness of the app (i.e. easy to use, satisfied, no advantage, useful for professional training, long-term advantage, use app continuously as part of work, satisfaction with the introduction, frequency of the use of note function).
- Level 2 Learning: the short reflection scale was presented in the post-questionnaires. Additionally, some app specific reflection questions, that fitted best to the evaluation approach, were added to the questionnaire. In particular these were differentiated among items that investigate concrete effect of the game on the users and items that investigate at the hypothetical level how the game would affect the users if they were already nurse. Finally the two Learning Outcome questions were added to the questionnaire.
- Level 3 Behaviour: the core behaviour question was adjusted to this evaluation by asking users, at an hypothetical level, how the game would improve their work performance if they were already a nurse.
- Level 4 Result: a Loyalty metric scale was used here to understand whether users will suggest the game to some other colleagues or friends working as nurses. Due to the characteristic of this evaluation (users played the game only once and fulfilled immediately after its usage the post questionnaire) it was not possible to measure the KPIs.

Additionally, some more questions about general comments to the game, and users' expectations about it were added to the questionnaire to get deeper insights about level 3 and level 4.

6.3.4 Results

6.3.4.1 Level 1: Reaction (Usage)

All the participants played the game only once for about 25 minutes (min. time 12.59 minutes, max. time 45.17 minutes). In general users used quite rarely the note function available in the game: in fact only 5 users used this function from 1 to 2 times during the game.

Participants were satisfied with the game (M=3.19, SD=1.38) that was evaluated as a usable tool (M=4.19, SD=0.66). Further they evaluated the game as a really useful tool for professional training of the medical staff (see Table 6.3.1).

Table 6.3.1. Level 1 'Reaction'

	University of Bergamo students
I think the app is useful for professional training and human resource development.	M=4.31; SD=0.70
I see the long-term advantage of using the app in the work of medical staff.	M=4.00; SD=0.73

Finally, introductions given to participants were evaluated adequate in order to play the game (see Table 6.3.2).

Table 6.3.2. Satisfaction with the introduction

	University of Bergamo students
After the introduction, I was able to use the app.	M=4.56; SD=0.73

6.3.4.2 Level 2: Learning

The goal of this evaluation was to investigate whether the 'CLinIC-The Virtual Tutor' serious game was able to increase the students' awareness about the topic 'difficult and stressful situations with patients' and if the game was able to increase the users' awareness about 'which kind of situations/problems they will be faced with once they will become nurses'.

In order to achieve these goals, questions used in the questionnaire to investigate level 2 'learning' were splitted in two different groups: one group with questions that investigated concrete effects of the game on the users and another group of questions that instead investigated potential effect of the game if all users were already working as nurses.

Data collected about this level (reported in the sections below) show how in general users acquired more confidence about their role of nurses thanks to the game.

Learning Process

Looking at the results of the reflection scale (see Table 6.3.3) it is clear how the participants of this evaluation were in general quite reflective. On average, individual reflection is higher than team reflection which is no surprise in this group as they were students and do not really work together as a team.

Table 6.3.3. Reflection scale

	University of Bergamo students
Mean Reflection scale.	M=3.90; SD=0.50
Mean Individual reflection	M=4.25; SD=0.44
Mean Collaborative reflection	M=3.55; SD=0.66

Regarding the app specific question, first of all participants evaluated the game as a useful tool to better understand the work of nurses. The scores slightly over average described in the table below (see Table 6.3.4) show how thanks to the game, participants increase their awareness about 'which are the main difficulties a nurse can be faced with'.

Table 6.3.4. App specific questions

	University of Bergamo students
The game helped me to better understand which are the main difficulties a nurse can be faced with	M=3.44; SD=1.26
Thanks to the game, I discovered something new about the role of a nurse within a hospital.	M=3.13; SD=0.96
Playing the game helps me to better imagine what the jobs as a nurse looks like.	M=3.88; SD=0.89

Further, about the reflection process, the game seemed to help participants to reflect by reminding them to reflect but also by the experiences made in the game (see Table 6.3.5). Interesting to see how in general users evaluated the simulation of events in care work really positive as a method to help them think about these events.

Table 6.3.5. App specific questions

	University of Bergamo students
The app helped me to reflect on experiences I made in the game.	M=3,63; SD=0,96
The app helped me by reminding me to reflect.	M=3,94; SD=0,44
The app helped me by providing information about similar experiences (for instance others' experience of the same or similar situation).	M=3,88; SD=0,89
The simulation of events in care work helped me to think about these events.	M=4,06; SD=0,57
The simulation of events in care work helped to consider whether I should behave differently.	M=3,94; SD=0,57
Being able to explore these 'virtual' experiences of caring was very helpful to me.	M=3,56; SD=0,96

The note function was not seen as really useful, which corresponds with the low usage of this function. (see section 6.3.4.1).

Table 6.3.6. App specific questions - notes function

	University of Bergamo students
Making notes in the game has helped me to build insights gained through the game for later use.	M=2.88; SD=0.96
Making notes in the game has helped me to reflect on my experiences in the game.	M=3.13; SD=1.15

Playing the game seems also to have helped participants to better imagine difficult situations during their work and to be prepared for them, furthermore students feel more confident now about performing their work successfully (see Table 6.3.7).

Table 6.3.7. App specific questions - results

	University of Bergamo students
By playing the game I feel better prepared for stressful situations in the work of a nurse.	M=3.25; SD=0.93
By playing the game I feel better prepared for difficult conversations with patients.	M=3.63; SD=0.72
The app helped me to imagine what kind of difficult situations I could be confronted with during my work.	M=4.06; SD=0.25
Using the app made me more confident of performing my work successfully.	M=3.81; SD=0.75

Finally from the perspective of a nurse, participants also tend to agree that the app would help them with their work and that the game would have a positive impact on their work as well (see Table 6.3.8).

Table 6.3.8. App specific questions - Hypothetical questions

If I was already working as a nurse…	**University of Bergamo students**
I think the app could help me to reflect on my work experiences.	M=3.88; SD=0.81
I think the app would make reflection sessions on my own more effective.	M=3.75; SD=0.93
I think the app would make collaborative reflection sessions more effective.	M=3.69; SD=0.87
I think the app would help me to gain a deeper understanding of my work.	M=3.50; SD=0.73

Learning Outcomes

Participants evaluated the learning effects due to the game quite reluctantly (see Table 6.3.9).

Table 6.3.9. Learning outcomes

	University of Bergamo students
After having played the game I would now behave differently in my work as a nurse than I had planned it before.	M=3.50; SD=0.73
As a result of playing the game I gained a deeper understanding of the work as a nurse.	M=2.94; SD=0.95

Interesting to analyse how the game was able to teach something to participants about how to behave in their work as a nurse (score slightly over average) while it was not really able to give users a better understanding of their work. Analyzing the qualitative data collected with some open questions in the questionnaires (see Table 6.3.10), it is possible to assume that the reasons of this score slightly below average are due to the fact that users whished more content and problem to be solved inside the game.

Table 6.3.10. General comments about the game

Discuss Topic	**Comment**
Game content	'I would like to see more procedurals challenges'; 'Really nice serious game but I think there is the need for more situations to be solved'; 'The game was easy and fun to use but there are still few content'; 'I see a great potential in this game and I think that with more problems and situations to be solved that game will be amazing'.

6.3.4.3 Level 3: Behaviour

Looking at the results described in the previous section, the serious game was evaluated positively with respect to its learning goals both on the concrete level and on the hypothetical level. At the behavioural level, instead, the effects on the game were investigated only on a hypothetical level because real impacts of it are expected only with users that are already working as nurses and that have the opportunity to actually change something in their work.

Table 6.3.11. Level 3 'Behaviour'

	University of Bergamo students

If I was already working as a nurse, I think the app would help me to improve my work performance.	M=3.75; SD=1.00
After having played the game I would now behave differently in my work as a nurse as I had planned before.	M=3.50; SD=0.73

Is it interesting to underline how the knowledge acquired with the game would be used by participants if they were already working as nurse in order to improve their work performance.

6.3.4.4 Level 4: Results

The first really interesting element about the level 4 results is that the site in which the evaluation was conducted is outside the consortium. That means that the University of Bergamo was so impressed by the serious game developed within the Mirror project for the training of new nurses that they decided to test it with a group of student enrolled in the second semester of the first year of the nursing faculty.

Secondly, really interesting are the results that have been collected with the loyalty metric. In particular it was asked to participants how likely (using a scale from 1 ‘not at all likely’ to 10 ‘extremely likely’) they would recommend the app to a friend or another person who is working as a nurse in a hospital as it is now and with more situations/problems to be solved.

The results described in the table below (see Table 6.3.12) show definitely the wish for more problems but also the general participants' satisfaction with the idea and the concept of the game.

Table 6.3.12. Loyalty metric

	Game as it is now	**Game with more problems in it**
Promoters (score from 9 to 10)	5	13
Passive (score from 7 to 8)	9	2
Detractor (score from 0 to 6)	2	1
NPS (promoters-detractors)	19%	75%

These data demonstrate how the game has been well accepted by participants who see in it a great new tool to support their training as a new nurse.

6.3.5 Conclusion & Discussion

Students in general perceived the game as a useful tool for professional training in the medical area and they stated that, if the game was extended with more content and situations to be solved, it would be great to use it as a part of the nurse training. Also, the positive results collected in the app specific questions show how the game was able to help participants to better imagine difficult situations they could be faced with during their work and to better understand how to deal with them. Further the game was evaluated as a useful tool to support their evaluation process. In particular the game supported well the first 3 steps of the reflection model thanks to its ability to simulate real work events and to offer users several tools able to initiate and conduct their reflection session. With respect to the last step of the reflection model

('apply outcome'), unfortunately it is not possible to know if participants would really change their behaviour thanks to the game because no data were collected about it. Despite that, the first collected results in level 3 are already promising and seem to indicate a good potential of the game also in supporting the 'apply outcome' step. Students also stated that the game helped them to feel more confident about how to perform their work successfully. According to these data it is possible to argue that the game is able not only to increase the users awareness about the topic 'difficult and stressful situations with patients' but it is also able to increase the user awareness about 'what could happen if they were already working as nurses'.

The positive results collected in the app specific questions, also demonstrate how participants judged the game as a useful tool to support their reflection process. In particular the game supported well the first 3 step of the reflection model thanks to its ability to simulate real work event and to offer users several tools able to initiate and conduct their reflection session. In fact participants used quite often the note function within the game stating that they really see a benefit in its usage. With respect to the last step of the reflection model ('apply outcome'), unfortunately it is not possible to know if participants would really change their behaviour thanks to the game because no data were collected about it. Despite that, the first collected results in level 3 are already promising and seem to indicate a good potential of the game also in supporting the 'apply outcome' step.

Finally the results collected in the loyalty metric seem to demonstrate how this serious game is a very good exploitable tool for the whole MIRROR project with regard to individual reflective learning at work. With more content and situations to be solved, the majority of the participants stated, in fact, they will recommend the game to other colleagues or friends that are working as nurses.

6.4 The Think better CARE – The Virtual Tutor Evaluation at RNHA

Think better CARE is a twin game to the CLinIC serious game described in the section above. It follows the same principles, the only difference is that is designed for care homes instead of hospitals.

6.4.1 Organisational context

Test bed organisation and the organisational unit

For a general description of RNHA we refer to section 3.6.

For the serious game evaluation, RNHA contacted a number of care homes from two groups – any of the 1200 RNHA members who had expressed an interest in being involved in the MIRROR research, and some 40 UK midlands care homes which were recipients of a government grant (Get Connected) for technology infrastructure in social care. From this group, 10 homes were chosen to be test sites – 7 in one group, Ashmere, and others in the midlands and in Suffolk. RNHA consultants visited each of these homes and spoke the manager or training manager, as appropriate and explained the proposed activity. In a second visit, a group of up to 10 staff were shown a demonstration, or given an explanation of the research task. If chosen/chose to participate, staff from each care home provided email addresses. These care staff – mixed volunteers – were to be the end users to provide sufficient quantitative data.

Test users and their job roles

Most staff in a care home is care staff that is direct givers of care to residents, providing for their personal care needs, and their recreational and therapeutic activities. There are small numbers of specialist auxiliary roles e.g. cook, handyman, cleaners, some of which can still be considered care staff because they have some contact with residents and they do receive training in a number of the caring areas.

Care staff, that is mainly ‘care assistants’, is typically managed by senior care assistants, who provide regular supervision and guidance in professional matters. Care staff often works in teams, led by a senior care assistant providing support to particular groups of residents, who may be organised by location e.g. a wing of a care home, or a specialist unit e.g. for residents with advanced dementia. Hierarchically, above Senior Care Assistants are Registered Nurses, who are medically qualified staff, and are statutorily required in sufficient numbers according to a ratio of residents to staff. A care home is led by a Registered Manager, who is often a Registered Nurse by background.

Although each care home has its own local policies and procedures, these are always within a national context of standards, regulation and inspection.

The users that participated in the serious game evaluation were mainly carers with no or few work experiences in the field. Further, some nurses and auxiliary staff participated in the evaluation as well.

Identified need and potential for reflective learning

Although often paid around the statutory minimum wage, a new care worker is expected to undertake an induction period and then training in some 13 or more mandatory areas of professional knowledge in his first two years of work, ranging from ‘manual handling’ to dementia care and ‘end of life’ care. Induction involves the shadowing of experienced carers,

as well as knowledge-based training. While e-learning is increasingly a part of this training, most training is still part of the traditional small group type with a specialist trainer presenting for a half-day or more. However, such general approaches cannot hope to cover all the variants of challenges likely to be faced by staff from their residents and their unique demands – that often requires some reflection, and some help, for example, asking experienced staff, or using creative thinking for solutions.

Typically culture, and opportunities to reflect formally and informally, is set by the care home manager. In the care homes used for the serious game evaluations, there is a culture of openness and a willingness to learn. In an ever-changing environment with unique demands there are always new challenges, and the care of people with dementia is still evolving.

The case of digital technology in the form of 'apps' to assist reflective learning is readily appreciated by staff, who are used to the practice of reflection in supervision, and in some cases, in formal and informal reflective sessions as groups of staff. And, for the first time, staff is increasingly comfortable with digital technologies, primarily due to the widespread use of smartphones and social networking apps.

Potential organizational impact

See section 6.3.

6.4.2 Theoretical assumptions

Selected approach for reflective learning and description of the app

The goal of this short term evaluation was to find out if the serious game has the potential (i) to trigger reflective learning on an individual level, (ii) to increase users' awareness about the topic 'difficult and stressful situations with residents'.

'Think Better Care - The Virtual Tutor' is a 3D serious game focused on difficult communication between care staff and residents. In order to support learning by reflection, both during and after the virtual experience, a number of specific tools were added to the game: notebook, game score, tutor feedback, overall feedback, individual reflection session and a learning diary. These provide support, motivation, guidance, and the opportunity to compare the improvement of performance in the emergency situations management over time.

After the initial serious game testing at Care Home H during Y3, in Y4 the game has been offered to several RNHA care homes to be included in the training process for new and inexperienced carers. The game is a stand-alone, single player application, and users at the end of their virtual experience can follow their game history, including all the notes collected during gameplay, from their learning diary, as a stimulus to reflect about their experiences. At the end of the game users have the possibility to discuss their game experience with the support of a supervisor and with colleagues to reflect collaboratively about their real and virtual working experiences.

Relation to MIRROR CSRL Model

See section 6.3.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

See section 6.3.

6.4.3 Research approach

Design and procedure

The serious game 'Think Better CARE-The Virtual Tutor' was introduced to 7 care homes as part of their week's induction process of new recruits, each month from November 2013 to March 2014. These groups were held on:

- 5th Nov.2013 at Codnor Park care home
- 2nd Dec. 2013 at Codnor Park care home
- 9th Jan.2014 at Smalley Hall care home
- 3rd Feb.2014 at Smalley Hall care home
- *4th March 2014 at Smalley Hall care home*

These groups, with between 4 to 6 care staff, were shown the game (by the RNHA consultants), or a description of the game was presented – depending on the equipment at the different locations. Due to time, cost and technical constraints in care homes, the game was made available remotely to the carers, allowing them to play it where and as many times as they preferred. To facilitate this process, each user received a personal username and password to play the game and URLs for:

- YouTube video as an introduction to the game;
- YouTube instruction to the keyboard navigation keys used in the game;
- a 'pre-use' on-line questionnaire (to be fulfilled before testing the game);
- a 'post-use' on-line questionnaire (to be fulfilled after testing the game).

No control group was included in this evaluation design.

Participants

Even if more than 30 participants were involved in the serious game evaluation, only 17 users fulfilled the pre and/or the post questionnaires. The reasons for this low participation are several and will be fully discussed in the next sections.

Notwithstanding this, the demographic of the participants who attended the 'Think better CARE-The Virtual Tutor' serious game evaluation, was so characterized:

- 17 participants; 1 male and 16 females; the median age was 30-39 (see Table 1); most of them were working as carers or nurses (see Table 6.4.1);15 were working full-time while 2 were working part-time; all of them were really newcomers to that position(i.e. Time in current job in years M=0.03 SD=0.05; Years in current department or team M=0.03 SD=0.05) but some of them have already some experiences from other care homes (Time doing this job -all places- in years M=4.14 SD=4.40)

Total number of participants: 17

Table 6.4.1. Demographics

gender		age				
male	female	under 20	20-29	30-39	40-49	50-59
1	16	1	6	3	5	2
job role		**working hours**				
full-time	part-time	carer	nurse	domestic	cook	kitchen assistant
15	2	10	4	1	1	1

Further, participants were in general really satisfied about their job role and they were confident to perform their job at their best also with respect to the relationship with residents (see Table 6.4.2).

Table 6.4.2. Job satisfaction

	RNHA care homes
How satisfied are you with your job/role in general?	M=4.64; SD=0.63
How satisfied are your residents in general with your care?	M=4.54; SD=0.66
I am confident of performing my work as a carer successfully	M=4.78; SD=0.44
I know what situations to expect from my work as a carer.	M=4.74; SD=0.44

Summative evaluation methods used

For the RNHA care homes evaluation the following procedure was followed:

- Demographic Information: the questionnaire contains all demographic questions from the toolbox, except for Team-ID and department because these data were not pertinent with this target.
- Level 1 Reaction: it was not possible to collet usage data from the log of the serious game. Instead, some other questions were added here in order to better investigate general satisfaction and usefulness of the app (i.e. easy to use, fun to use, useful for professional training, enough problems situations in it and frequency of the use of note function).
- Level 2 Learning: the short reflection scale was presented in the 'pre-app-use' questionnaire. Additionally, some app specific reflection questions, that fitted best to the evaluation approach, were added to the questionnaire. Finally, the two Learning Outcome questions were added to the questionnaire.
- Level 3 Behaviour: the core behaviour question was adjusted to this evaluation by asking users if the game helped them to improve their work performance and to increase their confidence of performing their work successfully.

- Level 4 Result: a Loyalty metric scale was used here to understand whether users will suggest the game to some other colleagues or friends working as carers. Due to the characteristic of this evaluation (users played the game only once and fulfilled immediately after its usage the post questionnaire) it was not possible to measure the KPIs.

Additionally, some more general comments and feedback about the game were collected through an unstructured interview with one of the consultant of RNHA who presented the game to the care homes, in order to better understand which were the problems/constrains that prevented a wider evaluation of the game and which were the general users' impressions about it.

6.4.4 Results

6.4.4.1 Level 1: Reaction (Usage)

As reported before, more than 30 care staff was involved in the serious game evaluations. Unfortunately, only 17 among them answered the pre questionnaires and only 5 answered the pre and post questionnaires. Therefore the results described in the following sections should be interpreted quite carefully.

Regarding the qualitative data, the 5 users that fulfilled the post-app questionnaires evaluated the game as a useful and fun tool (see Table 6.4.3).

Table 6.4.3. Level 1 'Reaction': usability and fun

	RNHA care homes (N=5)
The app has worked well.	M=4.00; SD=1.22
The app was fun to use.	M=3.80; SD=0.84

More neutral comments were instead collected about the usefulness of this tool as a complement to the training of new staff (see Table 6.4.4).

Table 6.4.4. Level 1'Reaction': usefulness

	RNHA care homes (N=5)
I think the app can be used to complement professional training of carers.	M=3.00; SD=0.71
I think the app has enough problems/situations in it to be useful.	M=3.40; SD=0.55

Further, regarding the note function available within the whole game no one of the 5 participants who fulfilled both the pre and post questionnaires, used this function.

6.4.4.2 Level 2: Learning

The goal of this short term evaluation was to find out if the serious game has the potential (i) to trigger reflective learning on an individual level and (ii) to increase users' awareness about the topic difficult and stressful situations with residents.

In particular, as stated before, the game was introduced to several care homes, during their week's induction process of new recruits, to see whether the game can be a good tool to complement the training of new carers/new care staff.

Learning Process

Looking at the results of the reflection scale (see Table 6.4.5) it is clear how the participants (N=16) of this evaluation were in general quite reflective. On average, individual reflection is higher than team reflection which seems to be mainly because there is not so much collaborative reflection outside of team meetings ('Outside of meetings, I often talk with my colleagues about the quality of the relationship with the residents' M= 3.13 SD=1.06).

Table 6.4.5. Reflection scale

	RNHA care homes (N=16)
Mean Reflection scale.	M=4.09; SD=0.45
Mean Individual reflection	M=4.37; SD=0.53
Mean Collaborative reflection	M=3.81; SD=0.68

Further, about the reflection process, the game seemed to be a useful tool in the process of helping participants to reflect by reminding them to reflect (see Table 6.4.6).

Table 6.4.6. App specific questions

	RNHA care homes (N=5)
The app helped me to reflect on experiences I made in the game.	M=3.63; SD=0.96
The app helped me by reminding me to reflect.	M=3.94; SD=0.44

Interesting to see how in general users evaluated the simulation of events in care work a quite positive method to help them think about these events (see Table 6.4.7).

Table 6.4.7. App specific questions

	RNHA care homes (N=5)
The app helped me by providing information about similar experiences (for instance others' experience of the same or similar situation).	M=3.80; SD=1.30
The simulation of events in care work helped me to think about these events.	M=4.00; SD=1.22
The simulation of events in care work helped to consider whether I should behave differently.	M=3.20; SD=1.48
Being able to explore these 'virtual' experiences of caring was very helpful to me.	M=3.40; SD=1.14

The note function was not evaluated as very useful which corresponds to the low usage numbers (see section 6.4.4.1).

Table 6.4.8. App specific questions - notes function

	RNHA care homes (N=5)
Making notes in the game has helped me to build insights gained through the game for later use.	M=2.88; SD=0.96
Making notes in the game has helped me to reflect on my experiences in the game.	M=3.13; SD=1.15

Learning Outcomes

Participants (N=5) evaluated quite neutral the impact of the game on the learning level (see Table 6.4.9).

Table 6.4.9. Learning outcomes

	RNHA care homes (N=5)
As a result of playing the game I made a conscious decision to change something in how I do my work.	M=3.20; SD=0.84
As a result of playing the game I gained a deeper understanding of my work life.	M=3.40; SD=0.89

6.4.4.3 Level 3: Behaviour

At the behavioural level, with the short term evaluations, users do not have the opportunity to actually change their behaviour before answering the post-app questionnaire (that is immediately after the game usage). This is the reason why a follow up questionnaire was created in order to check one month after the game usage, if some changes in behaviour occurred. Unfortunately, in this game evaluation we received no data for the follow up questionnaires. Therefore the only data that can be analysed in this level, are the ones described in the table below (see Table 6.4.10) that are based on the users subjective estimation if they would now behave differently.

Table 6.4.10. Level 3 'Behaviour'

	RNHA care homes (N=5)
By playing the game I feel better prepared for stressful situations at work.	M=3.00; SD=1.41
By playing the game I feel better prepared for difficult conversations with residents.	M=3.00; SD=1.41
The app helped me to imagine what kind of difficult situations I could be confronted with during my work.	M=3.20; SD=1.10
Using the app made me more confident of performing my work successfully.	M=3.20; SD=1.10

As shown in the table above, it is clear how in general users evaluated in a neutral way the possible consequence of the game.

6.4.4.4 Level 4: Results

At the end of the game experience, it was asked to participants how likely (using a scale from 1 'not at all likely' to 10 'extremely likely') they would recommend the game to a friend or another person who is working as a care staff in a care homes as it is now and with more situations/problems to be solved.

The results described in the table below (see Table 6.4.11) indicate the wish for more problems but also the general participants' satisfaction with the idea and the concept of the game.

Table 6.4.11. Loyalty metric

	Game as it is now	**Game with more problems in it**
Promoters (score from 9 to 10)	2	3
Passive (score from 7 to 8)	1	2
Detractor (score from 0 to 6)	2	0
NPS (promoters-detractors)	0%	60%

6.4.4.5 Expert interview

During March 2014, an expert unstructured interview with one of the RNHA consultants was conducted in order to get more feedback and insights about the game evaluations. In particular attention was focused on the following topics: constrains of the game evaluation, user experience, and professional evaluation of content. Below, the most significant results about these topics are described.

Constrains of the game evaluation

As reported before, more than 30 care staff was involved in the serious game evaluations. Unfortunately, only 17 among them answered the pre questionnaires and only 5 the pre and post questionnaires. The reason of this low participation was mainly due to the time, resource and technical constrains of the care homes that oblige users to run the game from their home. Playing the game for almost one hour during own personal leisure time was not perceived as a priority. Further questionnaires was judged by the care staff too long and too complicated, posing a challenge to many with limited understanding of 'reflective learning'.

However, this did not mean that care staff didn't have opinions. Several did play the game and expressed their views by email or via the training managers. One result of this dialogue was the great potential of this tool for the training of new recruits to residential care – 'I'd wish I'd seen something like this when I started'.

User Experience

The 'Think better CARE-The Virtual Tutor' serious game is a PC/laptop based app, requiring some familiarity of browsers and the internet, requires a plug-in (which confuses users), and uses a keyboard for navigation within the game. Those users not used to IT (pre smartphones and tablets) were not easily persuaded to use this technology. Further, even if staff is increasingly comfortable with digital technologies, there are still many within the personal care sector who have little or no practical experience of using PCs.

However, the game was not perceived as difficult to play, and with some hands-on demonstration, most non-users could use it and became familiar with the controls relatively quickly.

Regarding the 'game access', if played within the care homes, there were wifi reception problems in many locations – and no wifi at all in many homes. This is still a general problem in the sector. Accordingly games were played at on the personal machines of the care staff as well. Some didn't possess PCs at home.

There were few technical problems received when using the game, in fact the game did what users expected, and worked without bugs or interruptions. Although keyboard navigation was seen as 'old-fashioned' by some of the younger users, it was seen as effective. Movement in the game was smooth and interaction with objects was 'forgiving'.

Finally, there were considerable differences in the time taken to play a game. Some younger staff, familiar with technology, could play a game in little over 20 minutes – others would still be playing after an hour.

Professional Evaluation of Content

The general evaluation of the game content much depended on the experience of the carer. For those with considerable experience in the care sector, the problems were not challenging enough. However, they accepted that the content could be made more challenging, and some

suggested new scenarios, posing more difficult, and randomised, dilemmas. It was suggested that there should be different levels of difficulty, and progression through the levels.

Many users commented on the limited number of problems. More rooms and problems would have encouraged multiple uses of the game. It was rare for users to play more than once. Accordingly the facility to use the notes for future reference was rarely used in practice – though it was seen as potentially useful in a longer game i.e. with levels of difficulty.

For the newer carers, or new recruits to the sector, the game was seen as a good way of seeing what the sort of problems they were likely to face. The game was evaluated as a good tool to help them reflect upon situations they haven't faced yet, and without feeling bad about making mistakes.

The 'realism' of the problems set, and possible resolutions, was considered good - carers sometimes could put names to the characters that looked or behaved like residents they knew. The graphical home was seen as attractive, if somewhat 'bare' – 'our home has lots more decorations than that'. Although each home is different, individual rooms for residents are standard now – so 'wards' for residents in the game are seen as 'how it used to be'. Further, training managers liked the idea of the safe environment for the new recruits to 'experiment on residents'.

6.4.5 Conclusion & Discussion

Due to the low number of collected questionnaires, it is impossible to argue if the game's goals were achieved from a statistically point of view. However, matching the qualitative and the quantitative data collected with the 'expert interview', it is possible to state that there was a positive response to this novel approach to communicate professional information in a relevant, realistic but safe environment. Carers did not have a problem in suggesting uses of the basic game 'template', to be used in a number of specialist ways – for example for the recruitment filtering and for new recruits.

The game evaluation was planned with more participants (as described in the section below). The time, resource and technical constrains of the care homes that hosted the evaluations can be considered the most significant barriers that prevented a wider use of the game and a wider fulfilment of the questionnaires.

The general neutral results that have been collected with the quantitative data can be explained first of all due to the demographical characteristics of the sample that participated in the research. Some of these participants had, in fact, already a lot of experiences in their work and this could be the reason why they already felt so confident in their job and why there was a rather low score on the items about consequences of the app.

However, the qualitative results seem to show how the staff in general appreciated the game as part of training, particularly induction training. They in fact underlined how the game could be useful for emphasising good practice, and highlighting poor practice – with more examples needed.

Further, the fact that the newer staff appreciated how the game help them reflect upon situations they haven't faced yet, and without feeling bad about making mistakes could be a good hint about the game ability to support and trigger the individual reflection.

In conclusion, it is not possible to assess the likely long term effect on an individual or an organisation's performance from consistent use of the game, because these conditions were

not simulated – the game was generally played once, which is insufficient to produce a change of behaviour.

However, looking at the collected results, a ‘market ready’ version, with a greater range of problems could be part of an induction package for new recruits, and is likely to be well received by most care staff, particularly new recruits and those familiar with new technologies.

6.5 The Rescue League serious game Evaluation at Regola (118 emergency associations)

'Rescue League-The Virtual Tutor', is a 3D serious game that aims to train volunteers to face anxiety, medical protocols and dramatic choices

6.5.1 Organisational context

Test bed organisation and the organisational unit

The organizations where the evaluations of the 'Rescue League-The Virtual Tutor' serious game took place, are two public voluntary associations, named Croce Bianca Fossano (based in Cuneo-Italy) and SOS Novate Milanese (based in Milan, Italy). In Italy Public Safety Authorities and their professionals are strongly supported by voluntary associations that lend their vehicles and personnel (volunteers) to respond to different activities: civil protection, emergency medical services, planned/unplanned health care services (non-emergency), etc. Therefore the voluntary associations have to provide professional and high quality services to the community and to Public Authorities; hence the concrete need to professionally train volunteers, as well as rescue workers.

Test users and their job roles

The target users of this specific serious game are first of all volunteers, who work for the associations detailed above. Therefore they are volunteers engaged in the emergency medical services, in the civil protection, and many other applicable domains will come. Their formal activities are: responding to huge disasters coordinated by professionals, responding to medical emergencies on ambulances, transporting non-urgent patients from/to hospitals/medical facilities, etc.

The evaluation affected their two most important tasks which are responding to huge disasters coordinated by professionals and to medical emergencies. Their job always consists in checking the equipment on board and in the rescue bags, and to be prepared to apply rescue manoeuvres and procedures anywhere and in any condition, individually and in team. Different types of professionals can take part in managing their tasks, depending on the situation that the volunteers are facing.

Identified need and potential for reflective learning

New volunteers are usually trained in different steps: first they follow in class theoretical courses to learn how and when to apply rescue manoeuvres and protocols, then they spend other weeks to put in practice what they learnt previously. After passing one or more exams, they still have to spend many months placed side by side with experienced teams, watching real situations and hearing past experiences. It is clear that to be prepared effectively, many years of training are required. Coaching is sporadic and it is made through occasional updating courses and emergency simulations with dozens of volunteers.

Usually in these events managers or coordinators spend an amount of time in debriefing sessions (they also do briefing ones). The sessions at the end are useful to share opinions, to notify mistakes and to evaluate the training. However volunteers feel stressed and tired, they are put together in the global discussion with many others and the managers don't have neither time nor tools to talk about good/bad performances of each volunteer. So the experience is slowly forgotten in time by any attendee, especially among volunteers.

In most cases it is difficult to find an open culture (by managers / coordinators) towards the use of creative innovative solutions, but thanks to an experienced Coordinator that operates in one of the associations described before, open to new technologies and willing to discuss, take and test Mirror Apps in their domains, great opportunities have opened up.

The potential of reflective learning is very high, if this serious game could be implemented at different levels, for individual training and for collaborative sessions, targeting new volunteers, skilled and experienced members, managers and coordinators, with the purpose of improving the quality of each task at each level.

Potential organizational impact

If the serious game achieved its objective, its usage at Croce Bianca Fossano and at SOS Novate Milanese would have the potential to support the training of new volunteers and to increase their awareness about the difficult and stressful situation that could occur during an emergency situation. From an organisational point of view this means, that volunteers would be more prepared once a real disaster will occur and they will spend less money in organising real simulations. Hence, the volunteers' satisfaction as well as the manager satisfaction will be improved.

6.5.2 Theoretical assumption

Selected approach for reflective learning and description of the app

The goal of these evaluations was to find out, if the serious game has the potential to (i) trigger reflective learning on an individual level and (ii) to increase users' awareness about the difficult and stressful situation that could occur during an emergency situation.

'Rescue League-The Virtual Tutor', is a 3D serious game that aims to train volunteers to face anxiety, medical protocols and dramatic choices. In order to support learning by reflection, both during and after the virtual experience, a number of specific tools were added to the game: notebook, game score, tutor feedback, overall feedback, individual reflection session and a learning diary. These provide support, motivation, guidance, and the opportunity to compare the improvement of performance in the emergency situations management over time.

The game is a stand-alone, single player application, and users at the end of their virtual experience can follow their game history, including all the notes collected during gameplay, from their learning diary, as a stimulus to reflect about their experiences. At the end of the game users have the possibility to discuss their game experience with the support of a supervisor and with colleagues to reflect collaboratively about their real and virtual working experiences. Furthermore users have the possibility to add their own content to the game, through a tool called 'wizard tool' (for more information about this tool, see section 3.1 in D7.3).

Relation to MIRROR CSRL Model

See section 6.3.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

See section 6.3.

6.5.3 Research approach

Design and procedure

The evaluation at Croce Bianca Fossano was conducted at the end of October 2013 while the one at SOS Novate Milanese was conducted in April 2014. Both evaluations took place in one of the offices offered by the associations; since the participants were all volunteers, the evaluations were conducted in the evening in order to allow people who are working during the day to attend it. Further, in both cases, participants were asked to bring with them a personal laptop in order to play the game, because both associations did not have PCs in their offices. Since that in the SOS Novate Milanese evaluation not all the participants were able to bring with them a laptop, some of them played together the game.

Both evaluations lasted about 3 hours and no control group or data as baseline measure were collected.

Participants were asked to fulfil a consent form and a pre-app usage questionnaire. Then the game was presented to the participants by one consultant of Regola with the support of an expert Manager that has already collaborated with Mirror in the case of the Croce Bianca Fossano evaluation and by one researcher of Imaginary in the case of the SOS Novate Milanese evaluation. Afterward, participants were allowed to play the game and once finished they were asked to fill out a post-app usage questionnaire.

Participants

The demographic of the participants who attended the serious game evaluations was composed as follows:

- Croce Bianca Fossano: 19 volunteers, 10 females and 9 males, 7 were aged from 20-29, 7 were aged from 30-39 and 5 were aged from 40-49, 17 were new volunteers with few months of experiences in the field and only two were expert volunteers with 5 and 12 years of experiences.
- SOS Novate Milanese: 14 volunteers, 5 females and 9 males, 2 were aged lower than 20, 9 were aged from 20-29, 2 were aged from 30-39 and 1 was aged from 40-49; 12 were new volunteers with few months of experiences in the field and only two were more expert volunteers with 1 years and 1 years and 6 months of experience.

Both samples were in general really satisfied about their job role and they were quite confident to perform their job successfully and to know what kind of situations they will be faced with during their work as volunteer (see Table 6.5.1). Further, participants in both groups had a medium IT attitude (see Table 6.5.2).

Table 6.5.1. Job satisfaction

	Croce Bianca evaluation	**SOS Novate evaluation**
How satisfied are you with your job/role as volunteer in general?	M=4,; SD=1.1	M=4.07, SD=0.83
I am confident of performing my work as a volunteer successfully	M=3.95; SD=0.7	M=3.64; SD=0.93
I know what situations to expect from my work as a volunteer.	M=4.47; SD=0.69	M=4, SD=1.11

Table 6.5.2. IT attitudes

	Croce Bianca evaluation	SOS Novate evaluation
IT attitudes	M=3.1; SD=0.46	M=3.14, SD=0.41

Summative evaluation methods used

For both evaluations the following materials were used:

- Demographic Information: the pre-app-use questionnaire contains all demographic questions from the toolbox, except for Team-ID and department because these data were not pertinent with this target.
- Level 1 Reaction: all usage data can be received from the logs of the serious game. Further some other questions were added to the post-app-use questionnaire in order to better investigate general satisfaction and usefulness of the app (i.e. easy to use, fun to use, useful for professional training, enough problems situations in it and frequency of the use of note function).
- Level 2 Learning: the short reflection scale was presented in the 'pre-app-use' questionnaire. Additionally, some app specific reflection questions, that fitted best to the evaluation approach, were added to the post-app-use questionnaire. Finally, the two Learning Outcome questions were added to the post-app-use questionnaire.
- Level 3 Behaviour: the core behaviour question was included in a follow-up questionnaire and adjusted to this evaluation by asking users if the game helped them to improve their work performance and to increase their confidence of performing their work successfully.
- Level 4 Result: A Loyalty metric scale was included in the post-app-use questionnaire to understand whether users will suggest the game to some other colleagues or friends working as volunteers. Due to the characteristic of this evaluation (users played the game only once and fulfilled immediately after its usage the post questionnaire) it was not possible to measure the KPIs.

Additionally, with the pre-questionnaire the IT attitude was measured, and two questions about general satisfaction with the work were asked. Further, both in the pre and post questionnaires, 3 questions about procedures were added to see if the game was able to teach something about procedures. Finally two more general questions about comments and suggestions about the game were added to the post questionnaire.

6.5.4 Results

6.5.4.1 Level 1: Reaction (Usage)

All the participants played the game only once: the Croce Bianca participants played the game for about 11,38 minutes (min. time 4,04 minutes, max. time 32,47 minutes) while the SOS Novate Milanese participants played the game for about 41 minutes (min. time 15 minutes, max.time 50 minutes). In general participants of the Croce Bianca evaluation used often, compared to the other group and to other game evaluations, the note function during the game: in fact each participant wrote more than one note on average (Croce Bianca M=1,31; SD=2,38;

SOS Novate Milanese M=0.57; SD=0.85). Participants were satisfied with the game that was evaluated as a usable and really fun tool (see Table 6.5.3).

Table 6.5.3. Level 1 'Reaction': usability and fun

	Croce Bianca evaluation	**SOS Novate evaluation**
The app has worked well.	M=3,68; SD=1,06	M=3.64; SD=0.74
The app was fun to use.	M=4,21; SD=0,7	M=4.00; SD=0.55

Further the Croce Bianca participants evaluated the game as a really useful tool to complement the professional training of new volunteer, even if both groups, and in particular the SOS Novate Milanese participants, underlined the necessity of more content and more situations to be solved inside the game (see Table 6.5.4).

Table 6.5.4. Level 1 'Reaction': usefulness

	Croce Bianca evaluation	**SOS Novate evaluation**
I think the app can be used to complement professional training of new volunteers.	M=4,47; SD=0,7	M=3.57; SD=0.85
I think the app needs more content to be useful.	M=3,53; SD=1,12	M=4.07; SD=1.07

6.5.4.2 Level 2: Learning

The goal of this short term evaluation was to find out, if the serious game has the potential to (i) trigger reflective learning on an individual level and (ii) to increase users' awareness about the difficult and stressful situation that could occur during an emergency situation.

In particular, as stated before, the game was introduced to several new volunteers with no or few experience on the field to see whether the game can be a good tool to complement their training.

Learning Process

Looking at the results of the reflection scale (see Table 6.5.5) it is clear how the participants of these evaluations were in general quite reflective. On average, individual reflection is higher than collaborative reflection in both groups. Especially informal collaborative reflection outside of team meetings for the Croce Bianca participants does not seem to be quite common (Outside of team meetings, I often talk with my colleagues about the management of the maxi emergency situations. M=2.53, SD=1.17).

Table 6.5.5. Reflection scale

	Croce Bianca evaluation	SOS Novate evaluation
Mean Reflection scale	M=3.87; SD=0.48	M=3.86; SD=0.44
Mean Individual reflection	M=4.23 ; SD=0.57	M=4.01; SD=0.61
Mean Collaborative reflection	M=3.56 ; SD=0.64	M=3.71; SD=0.47

Further, about the reflection process, the game seemed to be a good tool to support participants in their reflection process especially for the Croce Bianca participants (see Table 6.5.6).

Table 6.5.6. App specific questions

	Croce Bianca evaluation	SOS Novate evaluation
The serious game helped me to reflect on experiences from work.	M=4.16; SD=0.76	M=3.43; SD=1.02
The serious game helped me by reminding me to reflect.	M=4.21; SD=0.85	M=3.57; SD=0.85
The serious game provided relevant content for reflection	M=4.26; SD=0.65	M=3.86; SD=0.66

Interesting also to see how in general users evaluated the serious game as a positive method to show them which situations they could be faced with during their work (see Table 6.5.7).

Table 6.5.7. App specific questions

	Croce Bianca evaluation	SOS Novate evaluation
The app helped me by providing accurate information about my work.	M=3.84; SD=0.83	M=3.64; SD=0.43
The serious game helped me by providing information about related experiences.	M=3.58; SD=1.17	M=3.36; SD=0.84
The serious game helped me by providing access to data relevant to the experience.	M=3.95; SD=0.7	M=3.57; SD=0.94
The app helped me by simulating the work process.	M=3.95; SD=0.7	M=3.79; SD=0.58
The serious game helped me by providing me with virtual experience in my work domain.	M=3.83; SD=0.98	M=3.57; SD=0.94

Finally, about the note function, in general the Croce Bianca participants evaluated slightly positive the benefit of using this tool during their virtual experience (see Table 5.5.9). The SOS

Novate Milanese participants were instead more neutral about it which corresponds to the low usage number (see section 6.5.4.1).

Table 6.5.8. App specific questions - notes function

	Croce Bianca evaluation	**SOS Novate evaluation**
Using the note-function has helped me to keep insights through the game for later use.	M=3.56; SD=0.85	M=3.29; SD=1.20
Using the note-function has helped me to reflect on my experiences in the game.	M=3.39; SD=1.03	M=3.00; SD=0.88

Learning Outcomes

Croce Bianca participants evaluated more positively than the SOS Novate Milanese participants the learning outcomes of the game, stating that the game help them to better understand their work (see Table 6.5.9).

Table 6.5.9. Learning outcomes

	Croce Bianca evaluation	**SOS Novate evaluation**
After playing the game I made a conscious decision to change something in my work behavior.	M=3.68; SD=1.0	M=2.71; SD=0.73
After playing the game I gained a deeper understanding of my work life.	M=3.63; SD=1.06	M=3.14; SD=0.66

Further, some questions about practical procedures were asked to participants both in the pre-app-usage and in the post-app-usage questionnaires. Is it interesting to see, as described in table 10, how an improvement on the learning questions occurred due to the virtual experiences (in particular for question 1 and question 3). Looking at the results collected with respect to question 2 it is possible to state that instead this question was obviously too easy or at least almost all knew the answer already beforehand.

Table 6.5.10. Learning questions

	Croce Bianca evaluation **Right answers/pre-app**	**Croce Bianca evaluation** **Right answers/post-app**	**SOS Novate Milanese evaluation** **Right answers/pre-app**	**SOS Novate Milanese evaluation** **Right answers/post-app**
Q1	N=12	N=18	N=6	N=9
Q2	N=16	N=17	N=10	N=10
Q3	N=9	N=16	N=11	N=14

6.5.4.3 Level 3: Behaviour

At the behavioural level, with the short term evaluations, users do not have the opportunity to actually change their behaviour before answering the post-app questionnaire (that is immediately after the game usage). This is the reason why a follow up questionnaire was created in order to check one month after the game usage, if some changes in behaviour occurred. Unfortunately, in these game evaluations no data could be collected with the follow up questionnaires. The only data that can be analysed at this level, are the ones described in the table below (see Table 6.5.11) that are based on the users subjective estimation of the game effects on their behaviour.

Table 6.5.11. Level 3 'Behaviour'

	Croce Bianca evaluation	**SOS Novate evaluation**
By playing the game I feel better prepared to deal with maxi emergency situations.	M=3.74; SD=0.99	M=3.43, SD=0.94
The app helped me to imagine what kind of difficult situations I could be confronted with during my work.	M=3.84; SD=1.01	M=3.57, SD=0.94
Using the app made me more confident that I can succeed in my work-tasks.	M=3.79; SD=0.79	M=3.57, SD=0.76

As shown in the table above, it is clear how in general users evaluated slightly positive the possible consequence of the game.

6.5.4.4 Level 4: Results

At the end of the game experience, it was asked to participants how likely (using a scale from 1 'not at all likely' to 10 'extremely likely') they would recommend the game to a friend or another person who is working as a volunteer as it is now and with more situations/problems to be solved.

The results described in the table below (see Table 6.5.12) show the general Croce Bianca participants' satisfaction with the idea and the concept of the game. No significant differences have been notice between the loyalty metric scale of the 'game as it is now' and the 'game with more problems in it'.

Table 6.5.12. Loyalty metric (Croce Bianca participants)

	Game as it is now	**Game with more problems in it**
Promoters (score from 9 to 10)	12	14
Passive (score from 7 to 8)	5	4
Detractor (score from 0 to 6)	2	1
NPS (promoters-detractors)	53%	68%

The results described in the table below (see Table 6.5.13) instead show the SOS Novate participants' willingness to more problems and situations to be solved inside the game.

Table 6.5.13. Loyalty metric (SOS Novate participants)

	Game as it is now	Game with more problems in it
Promoters (score from 9 to 10)	0	7
Passive (score from 7 to 8)	8	4
Detractor (score from 0 to 6)	6	3
NPS (promoters-detractors)	-43%	29%

6.5.4.5 Additional questions

Several open questions were added in the post-app-usage questionnaire in order to better investigate the general users' expectation and comment about this tool. Below, the most interesting collected feedback is described (see Table 6.5.14).

Table 6.5.14. General comments about the game

Discuss Topic	Croce Bianca Comment	SOS Novate Comment
Expectations about the game	'I was expected to see a new method to learn something and my expectations were totally met!'; 'I expected a tool that helped me to increase my ability about the procedures, and I can say that after the game experience my expectations were enough met' 'I was expecting nothing different from what I saw in the game…congratulations, well done!'	'I did not have any expectation about the game. It is the first time that I see a tool like this and I think it should be really useful for the training of new volunteers (of course with more content and situations to be solved)'.
Comment/suggestions about the game	'Compliments! This is a really good idea to train in a fun way new volunteers'; 'Useful! Spectacular!'; 'I like it very much even if I had some technical problems'.	'Good tool! I think some more improvements are necessary in order to have an excellent and really useful serious game'.

As displayed in the Table 6.5.14, only invariable positive feedback were collected. This is due to the fact that only participants that appreciated the game decide to leave a comment.

6.5.5 Conclusion & Discussion

Looking at the results of the 'Rescue League-The Virtual Tutor' serious game evaluations, described in the chapter above, it is possible to argue that the game was generally well

evaluated and well accepted by the new volunteers even if they were not so confident and familiar with this kind of technology (as demonstrated by the medium score in the IT attitude scale).

The positive results collected in the app specific questions, especially for the Croce Bianca evaluation, also demonstrate how participants judged the game as a useful tool to support their reflection process. In particular the game supported well the first 3 step of the reflection model thanks to its ability to simulate real work event and to offer users several tools able to initiate and conduct their reflection session. In fact participants that evaluated more positively the game used quite often the note function within it stating that they really see a benefit in its usage. With respect to the last step of the reflection model ('apply outcome'), unfortunately it is not possible to know if participants would really change their behaviour thanks to the game because no data were collected about it. Despite that, the first collected results in level 3 are already promising and seem to indicate a good potential of the game also in supporting the 'apply outcome' step. The game seems to have had an impact also on the learning of participants: thank to the virtual experience, some participants increase their knowledge about the right procedures that need to be used during an emergency situation.

Finally the results collected in the loyalty metric seem to demonstrate how this serious game is a very good exploitable tool for the whole MIRROR project with regard to individual reflective learning at work. With more content and situations to be solved, the majority of the participants stated, in fact, they will recommend the game to other colleagues or friends that are working as volunteers.

7 Interim evaluation report

This section contains the report of the current state of the Yammer app evaluation. It is designed as a summative evaluation and therefore reported in this deliverable but as the current state does not allow to report results from all four levels of the Kirkpatrick Model we present it here as a progress report with the so far received results. The evaluation was set up for eight months. It started in November 2013 with a baseline condition and is expected to last until the end of June 2014 (end of second intervention). Results that are still outstanding at the time of completing this report, will be presented at the final review. Work carried out after the project ends, will be done on CITY's and RNHA's own budget.

7.1 The Yammer Evaluation at RNHA (Nightingale)

The Yammer app aims at improving person-centred care of care-home residents by providing support for capturing, recording, and sharing care records and other information about residents.

7.1.1 Organisational context

Test bed organisation and the organisational unit

The summative evaluation took place in Nightingale House in South West London, one of the Jewish community's leading providers of care for older people. Nightingale House prides itself in providing high quality holistic care in a safe and stimulating environment. It provides residential, nursing and dementia care, and accommodates 200 residents. It offers a home for life and it cares for residents from the moment they arrive and for the rest of their lives, no matter how much their physical or mental health may deteriorate, thereby giving security to both residents and their families. It has a dedicated, well-trained team of staff and volunteers to ensure that individuals living in both of our homes receive attention all day, every day. The carers and residents that took place in the evaluation lived and worked in the dementia wing of Nightingale House. All of the 40 residents had some form of dementia.

Test users and their job roles

The test users were 20 carers who worked on the dementia wing of Nightingale House. The carers had been working in the residential home for a minimum of 2 weeks, and a maximum of 18 years with an average of 6 years working for the care home. Of these 29 participants, 24 are female. They range from 26 years to 65 years of age, with an average age of 43.

Identified need and potential for reflective learning

The residential home had identified the potential for reflective learning about individual residents as part of the approach to deliver more effective person-centred care to residents in the home. The decision was made to pilot the evaluation in one wing of the home that housed people living with dementia, who have the potential to benefit more from enhanced person-centred care. That said, although the need and potential for reflective learning was identified, the home identified a more important and basic need to be resolved with the technology – more complete and accurate care records to be recorded, which could then enable and improve more effective reflective learning and person-centred care.

Potential organizational impact

The primary potential impact of the summative evaluation was improved person-centred care of the residents who were living in the home. The intention was to improve person-centred care through the introduction of new technologies that would improve the capture and recording of care records and other information about residents, that would improve the sharing of this information between carers and other staff in the home, and that would support more creative thinking and reflective learning about residents. As such, reflective learning is one of a number of strategies to improve person-centred care through the use of the Mirror apps.

7.1.2 Theoretical Assumptions

Selected approach for reflective learning and description of the app

Yammer is an enterprise micro-blogging app that we installed on mobile iPod Touch devices. Carers were expected to record care notes about each resident in situ during care shifts and communicate these notes with other carers using a private and encrypted network in real-time. Functions on the mobile client with which to post messages, tag them to enable viewing, searching, alerting, and adding comments to previous posts were used to capture, share and reflect on care notes. Figure 7.7.1 depicts some typical uses of the Yammer app to document and reflect on recorded care notes. The left-hand side shows how a carer can enter a care note into the app using the standard device keyboard. The middle of the Figure demonstrates the use of a search function to retrieve all resident care notes with the term '*fluid'*. And the right-hand side depicts a stream of sequential care notes around a resident, with replies from other carers, with which to reconstruct the care experience and articulate meaning to the care of one resident.

Figure 7.1.1: Uses of the Yammer app adapted for care note recording and reflective learning

Relation to MIRROR CSRL Model

Planned use of the app to support most of the activities described in the computer-supported model of reflective learning is described in Table 7.1.1. Not all of the activities were supported

explicitly by app features. Different work events, such as starting or ending a care shift, or encountering a challenging behaviour by a resident, might trigger a decision to reflect using the app, and explicit criteria with which to re-evaluate an experience were assumed to be part of the care processes adopted in the residential home.

Table 7.1.1. Designed use of Yammer app features, and related work redesigns, to support reflective learning activities described in the model of computer-supported reflective learning.

Reflection model activity	**Yammer feature**
Monitor work	Record care notes in situ at the time that care delivered
Decide to reflect	*No explicit support*
Frame the reflection session	Define search queries to retrieve previous care experiences documented in notes about residents, then select the one or more notes to reflect about
Make related experience available	Directly browse all previous care notes about the resident from all carers
Reconstruct experience	Read back one or more care notes to reconstruct the past care experience. Notify other carers of selected care notes
Understand meaning	Each care note is presented in context of the feed of related care notes about the resident
Articulate meaning	Reply to a care note to add more meaning to that note
Frame the re-evaluation	*No explicit support*
Critique experience	*No explicit support*
Reach a resolution	Document or more new care notes that record the resolution to be applied for the resident

The Evaluated Apps

Two related Mirror apps were rolled out for use in the summative evaluation:

1. A version of the Yammer app extended to support individual reflective learning and care note recording about residents in situ during care work. The app was delivered on iPod Touch mobile devices provided to carers for use on their shifts. As well as enter care notes about individual residents via the device's keyboard and speech recognition, each carer was able to browse all care notes entered by all of the carers about each resident, search for care records based on keywords, and respond directly to entries about residents about other carers;
2. A web-based Care Reflection app that supports individuals and teams of carers to select individual or sets of care notes about a single resident for reflection, then guides structured reflection about each resident, resulting in new, more reflective care notes being recorded about the resident – care notes that could be accessed via the Yammer app in-situ.

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

Explicit support for reflection was provided at two levels:

1. Individual reflection about one resident at a time by single carers and other staff through use of the Yammer app;
2. Collaborative reflection about one resident at a time by groups of carers in structured reflection sessions through use of the Care Reflection app, under the guidance of a more senior member of staff.

There was no planned organisational reflection. The object of reflection at any time was an individual resident in the home, and organisation-wide reflection about residents did not take place in the home.

7.1.3 Research approach

Design and procedure

All carers on the dementia wing were required to take part in the evaluation, and all started to use the selected Mirror apps at the same time on instruction from the managers on the wing. The apps fully replaced the previous manual process used for care note recording and analysis for the duration of the evaluation. All 20 carers received face-to-face training in both apps at the point of their introduction. To provide training to use the Yammer app, one Mirror researcher who was on-site delivered training in the app, procedures to use the app, the use of the iPod Touches that gave access to the app, and how to reconnect the device to the residential home's wireless networks in case of network outage. The same researcher, at a later date, provided training about the Care Reflection app to carers in groups. The Care Reflection app was accompanied by use of the Hazel Court app to train these carers in more creative and reflective thinking for person-centred care.

The summative evaluation was designed to collect and analyse care data from a baseline condition and from two intervention conditions:

1. The baseline condition was the current use of paper-based care note recording about selected residents living in the dementia wing of the residential home;
2. The first intervention condition was the introduction of the Yammer app to all carers working on the dementia wing to replace the paper-based care note recording and reviewing. Care notes were accessible both online via the Yammer app and in digital printed form in regular care notes;
3. The second intervention condition was the additional introduction of the web-based reflection learning application on top of use of the Yammer app in structured reflective learning sessions led by senior carers on the wing.

The evaluation was set up to run for the last 8 months of the project to maximise the data and participants in the evaluation, across the following dates:

1. The baseline condition was set up and data collection from it took place from 1st November 2013 to 16th December 2013;
2. The first intervention was set up and data collection from it took place from 17th December 2013 to 31st March 2014;
3. The second intervention was started on 1st April and rolled out incrementally from that date, and data collection from it is taking place from 1st April 2014 to 30th June 2014.

In this version of the evaluation report, we summarise findings from the first intervention. Data collection related to the second intervention is currently ongoing.

Participants

All participating carers had been working in the residential home for a minimum of 2 weeks, and a maximum of 18 years with an average of 6 years working for the care home. 24 are participants are female, 5 are male. They range from 26 years to 65 years of age, with an average age of 43.

Summative Evaluation Methods used

At the start of both research interventions, all participating carers were given training in the necessary hardware and software technologies, and provided with online access to an office-hours helpdesk. As the summative evaluation was conducted in situ over a long period, access to research data was limited. The following data was collected and analysed in the summative evaluation:

- All care notes recorded about residents living in the dementia wing by the participating carers;
- Reports of software and hardware problems and errors from carers and managers in the residential home over the period;
- Periodic interviews undertaken by one researcher with participating carers and their managers in the period. Questions asked included:

Question	Purpose
How has the Yammer trial been so far?	Opens up the discussion, gives an opportunity to discuss some of the niggling issues like the network
Tell me about your experience of using Yammer?	How have they experienced using Yammer as part of their daily work. Looking for answers beyond just the practical and exploring any positives or negatives voiced by users.
When do you use it?	Looking for particular time of day, point in shift, or if they are really using it as intended which would be whenever something happens
Where do you use it?	Are users seeking out somewhere private or are they able to use it while looking after residents
Has the introduction of Yammer changed the way you make notes	Early on the unit manager reported that the note making has improved. This question is looking to see if staff feel the same
Has anything changed in terms of the way you do your job based on using yammer?	Has Yammer had an impact on other areas of work, e.g. changes in work practice, changes in care, both positive and negative.
Do you make use of/read the notes recorded by other carers? How? Why? Can you give me examples?	Aiming to have a discussion about reflecting on care notes made by others based on the response to this question

Did you use to read the notes made by others when you used paper notes?	Exploring if the introduction of digital notes has made the users feel that they are more likely to read notes made by others.
Tell me about the devices – how are they working for you?	A lot of the problems previously noted has been in relation to the devices and to the network. Exploring if there are any other issues.

- Periodic focus groups undertaken by one researcher with participating carers and their managers in the period. One month into the trial the Unit Manager, a Team Leader and the Person Centred Care Facilitator participated in a focus group to explore the impact of the introduction of Yammer. Topics explored included choices around changes in work practices; general user experience from both staff and manager; and overall confidence in using the app. Periodic discussions with the manager and Person Centred Care Facilitator were held to explore progress and continued user experience with the app.

Due to university research ethics protocols, one researcher who worked for the Registered Nursing Home Association analysed the care notes recorded in the period. The researcher reviewed each care record, and counted the number of frequencies of each occurrence of each different type of content in the note, based on the following content analysis scheme:

1. A description of resident's observed regular behaviour, for example the resident sits to watch television in his room, or eating most of her lunch in the dining room;
2. A description of a resident's regular states based on direct observations of the resident, for example calm, happy, and withdrawn. These tended to be descriptions of the resident's emotional status, as inferred from their behaviour, including *slept well*, *ate well*, *was fine*;
3. A description of a resident's observed challenging behaviour, that the physical safety of the person or others is likely to be placed in serious jeopardy, or behaviour which is likely to seriously limit use of, or result in the person being denied access to, ordinary community facilities, for example refusing to take medication, and shouting violently at other residents;
4. A description of one or more observed and/or reported carer responses to observed challenging behaviours, for example asked another carer to provide the medication, and moved the resident into another room away from the other residents. In practice, these responses were both physical and verbal, and a pre-emptive action by staff following a trigger;
5. A description of one or more observed and/or reported carer interventions into a resident's regular care, e.g. helped to eat lunch. These included descriptions of monitoring, encouraging, reminding and asking;
6. Observed and/or reported relevant carer behaviour that does not include interactions with residents., for example consulting with the supervisor and talking to the relatives of the resident;
7. A description of an attribution to meaning of resident's behaviour, state or condition that could not be observed, for example suspect personal insecurities underlie this behaviour and appears not to enjoy gardening activities;
8. A description of a proposed resolution to a monitoring situation or encountered challenging behaviour, for example recommend the removal of the reasons for

personal insecurities during lunchtimes and suggest that the resident walks in the garden rather than undertake gardening activities;

9. A description of an explicit inference made about the resident, for example the resident is not always asleep when she appears to be and I believe that she has an allergic reaction to this foodstuff.

In the summative evaluation, we investigated data collected from the residential home to answer the following 3 research questions:

RQ1 Does the use of the Yammer app increase the volume of resident care note content that will provide the topic of reflection about residents?

RQ2 Does the use of the Yammer app increase the volume of documented reflection about residents?

RQ3 Does the use of the Care Reflection application increase the volume of documented reflection about residents?

In this deliverable, we report answers to the first two research questions.

7.1.4 Results

7.1.4.1 Level 1: Reaction

This version of the summative evaluation reports findings from 13 residents chosen for scrutiny to cover the spectrum of residents in the dementia wing. All handwritten care notes recorded for these 13 residents from the 1st Nov 2013 to 16th Dec. 2013 were analysed. A total of 1,058 care notes, with an average entry of 34 words, totalling 35,996 words were analysed. Of these, 59 care note records, or just over 5%, were partly or wholly illegible. The digital care notes recording using the Yammer app and mobile devices between the 17th Dec 2013 and 31st March 2014 were also analysed. The carers recorded a total of 2,920 care notes for the 13 residents over the period analysed, with an average word count of 39 (114,000+ words in total). All of the digitally recorded care notes were legible. The Table 7.1.2 reports a summary of the quantitative results from the first two phases of the evaluation.

Table 7.1.2. Quantitative results from first two phases of the evaluation

	Paper-based care notes			Digital care notes		
Res. No.	Total Care Notes	Average Words per Note	Average Scores	Total Care Notes	Average Words per Note	Average Scores
1	69	31	2.9	220	37	2.1
2	94	31	1.6	226	34	2.4
3	84	34	2.7	207	35	2.4
4	84	30	2.7	213	30	2.6
5	92	38	2.8	230	43	2.6
6	89	28	2.6	213	32	2.7
7	86	46	3.1	239	45	2.9
8	20	50	3.2	246	61	2.9
9	93	30	2.1	209	35	2.7
10	72	28	2.7	229	31	2.8
11	88	36	2.9	243	51	2.8
12	89	38	2.9	217	37	2.8
13	98	34	2.5	225	34	2.8
All	**1058**	**34**	**2.6**	**2917**	**39**	**2.7**

The average number of words per care note increased by 14.7%, from 34 to 39, after the introduction of the digital care note recording. A unpaired t-test for unequal variance revealed that the increase in the number of words recorded in the digital care notes was significant (p<.001). This result suggests that the introduction of the digital technology increased the volume of content per care note that was available to carers to reflect on. Moreover, the average number of care notes that were recorded per day per resident increased from 1.66 to 2.05 after the introduction of the digital technologies. A paired t-test on the averages per resident over the periods also revealed the difference to be significant (p<0.01). Again, this result suggests that the introduction of the digital technology increased the volume of content that was available to carers to reflect on.

7.1.4.2 Level 2: Learning

The content of each recorded care note was also analysed for evidence of reflection about residents and their care. Table 7.1.3 shows the quantitative overview of the occurrences of content in care notes recorded during the baseline period and after the first intervention. We

sought documented evidence of reflection in the form of descriptions of attribution to meaning of resident's behaviour, state or condition that could not be observed.

Table 7.1.3. Totals of different types of content recorded in paper-based and digital care notes in the summative evaluation period

	Paper-based care notes		Digital care notes	
Type of content	Number	%age of total	Number	%age of total
Description of resident's observed regular behaviour	956	90.4	2775	95.1
Description of resident's regular states based on direct observations of the resident	803	75.9	2203	75.5
Description of resident's observed challenging behaviour	15	1.4	91	3.1
Description of one or more observed/reported carer responses to observed challenging behaviours	6	0.6	58	2.0
Description of one or more observed and/or reported carer interventions	952	90.0	2543	87.2
Observed and/or reported relevant carer behaviour that does not include interactions with residents	43	4.1	114	3.9
Description of attribution to meaning of resident's behaviour, state or condition that could not be observed	1	0.1	5	0.2
A description of a proposed resolution to a monitoring situation or encountered challenging behaviour	0	0	5	0.2
A description of an explicit inference made about the resident	0	0	0	0

The data revealed little evidence of documented reflection in both the baseline condition and after the introduction of the Yammer app. In contrast, most recorded care notes described resident's observed regular behaviour, and regular states based on direct observations of the resident – content to trigger and become the subject of reflection.

To understand more about the content that was recorded by carers in the care notes, we computed the frequencies of the different types of content in each single care record. Quantitative results are reported in Table 7.1.4, and reveal two key differences after the introduction of the Yammer app. First, the introduction of the app substantially reduced the number of care notes that contained no meaningful information about the resident. In short, the use of the app appeared to provide more structure and focus for the care note recording. Second, the use of the app did lead to an increase in the number of complex care notes with 5 or more different types of contents, albeit in small numbers.

Table 7.1.4. Totals of occurrences of each type of content in a single care note record

	Paper-based care notes		Digital care notes	
Number of different types of content in each care note	Number	%age of total	Number	%age of total
0	59	5.58	6	0.21
1	43	4.06	222	7.61
2	175	16.54	677	23.21
3	747	70.6	1889	64.76
4	28	2.65	72	2.47
5	6	0.57	44	1.51
6	0	0	6	0.21
7	0	0	1	0.03

7.1.4.3 Level-3: Behaviour

Due to the nature of the summative evaluation, changes in carer and resident behaviour could not be observed. Therefore, we explored the impact of app use through the periodic interviews with carers and their managers. These interviews revealed that use of the app led to reductions in time spent on administrative tasks, allowing carers to spend more time delivering care to the residents. One carer reported that: "*I think this is much better {using Yammer} rather than writing on the care plan. It saves us from carrying loads of care plans, and taking them (the care plans) in the lounge, and it saves time as well for us*". This view was reported by the care home manager, who reported that: "*One of the positive things for me is that the staff are able to be in the vicinity with the residents, even though they're still completing the daily notes their presence is there. They don't have to be in the office and writing the daily notes*".

Another emerging change in carer behaviour was easier access to resident information. Access to this information is a pre-requisite to reflection, as it is an important instance of making related experiences available. The care home manager reported that: "*one of the staff*

mentioned that even if they were on their break, and they remembered something, they could immediately type it in instead of waiting to finish their break and then you totally forget". Another stated that: "*It's also good that all staff have access to all residents, because then the other don't have to wait to write something {in the care plan} because I'm still writing*." The manager, talking for one member of staff, said: "*It's easier instead of reading four pages of staff notes to pick out this one staff, what she wrote, you just log in under her staff {username in Yammer} and everything, you can show where the patterns are, because for every resident she was doing it {not addressing the resident by name}*". The manager was able to review the note taking with the staff member to help improve her work – a form of learning: "*With the handwritten notes, every week they archive it with <the Yammer app> I can go back months and see what they input*".

More importantly, the interviews revealed evidence that the app use was already enhancing person-centred care in the residential home – one carer reported that: "*because we are concentrating on person centred care, every day these residents they have different care needs so we need to cater for each and every resident every day on a daily basis. Now although there are some that have their routine, with the addition of yammer app it just makes easier for us to document and communicate or interact about the resident*"

7.1.4.4 Level-4: Results

At the time of writing, this summative evaluation has delivered an interim result, and we are still collected evidence of data to indicate evidence of results at this level.

7.1.5 Conclusion & Discussion

Information collected during the summative evaluation has enabled us to answer the first two research questions:

RQ1 The use of the Yammer app did increase the volume of resident care note content that will provide the topic of reflection about residents

RQ2 In contrast, the use of the Yammer app did not increase the volume of documented reflection about residents.

As such, the evaluation did reveal an impact from the introduction of the apps, however the support for reflective learning appears not to have been as effective as planned. In particular, individual reflection with an app that does not provide explicit support for all reflective learning activities described in the Mirror CSRL model was not an effective an anticipated. Moreover, this interim result suggests that a socio-technical solution for reflective learning, which provides management and procedural support, might be needed for effective reflection in residential care settings – the research question that we are exploring in the remainder of the summative evaluation.

The impact of the app introduction is best appreciated by the decision of Nightingale to continue to use the Mirror apps and technologies after the end of the summative evaluation, from July 1st onwards. We are currently working with the residential home to determine the most effective way to make this happen.

8 Conclusion and Outlook

In each of the reported evaluations valuable insights about the effects of introducing technology supported reflective learning at the workplace have been gained. These are reported in the respective sections of Chapters 5 through 7 and are therefore not repeated here.

By conducting summative evaluations, data has been collected, analysed, and reported across a wide range of research perspectives. Starting with the rather low-level aspects of usage and barriers for pure usage of apps up to high-level effects on KPIs for the relevant teams of an organisation. Thus, also the insights have been gained with respect to all four levels of the Kirkpatrick model.

One vital prerequisite for any evaluation study is that participants are motivated to use the tools they are supposed to assess. Across the different testbeds, we found some common barriers that prevented participants from using the MIRROR apps. Examples for how to foster app usage are to make sure that users understand the benefits they gain from using the app, that the management supports the evaluation and provides the time and space to participate, and that all technical problems are sorted out before testing. Studies reporting technical problems or a lack of management support have mostly rather negative outcomes. Additionally, a thorough introduction to the apps and the concept of reflective learning, as well as help with the interpretation of data and the process of reflection are crucial for a successful implementation of new technologies. Generally, the different evaluation settings showed that the more face to face contact there is between researchers or app developers and users, the more positively do users respond to the apps. With regard to the concrete results obtained on the level of reaction, we found for the most part a positive reception of the apps from the users. The app-specific support of reflective learning provided by the MIRROR apps was mostly perceived positive by the participants and several participants also reported a positive learning outcome. The results of several evaluations also indicate that users perceived improvements in their behaviour at work, as well increased satisfaction or confidence with the working tasks. A more comprehensive view on these data and the mentioned insights is given in D1.7 which analyses the data across all summative evaluations.

Looking at the highest level of results, in every evaluation, partners looked at specific KPIs on which the approaches and apps evaluated may have an effect. Depending on the evaluation setting and organisational context, the specified KPIs concerned either objective measures that were already installed in the organisational assessment processes (e.g. number of calls or customer ratings in MMA and IAA/IMA evaluations at BT), user satisfaction via the loyality metric or net promoter score (e.g. KnowSelf, CaReflect, WATCHiT, or serious games), or more individualized measures assessing relevant factors on either personal or team level, as for example employee, customer or patient satisfaction (e.g. evaluations of TalkReflect or DoWeKnow) or quality of work (e.g. medical quiz). Due to evaluation durations of maximal 2.5 months and mostly restricted number of participants, it was difficult to prove a direct impact of the reflection approaches on these KPIs. However, there are positive indications that reflective learning could increase quality of work, better individual work performance of employees, increased employee satisfaction, and increased client satisfaction.

Besides the methodological difficulties we faced with measuring objective KPIs, there are some other points to be noted regarding the used methodology. In order to obtain a comprehensive picture of the impact MIRROR apps have on reflective learning at work in different organisational settings, it was necessary to follow a common methodology in all

evaluations. This included guidelines for the general procedure as well as a toolbox with core questions to be asked in all evaluations and a set of additional questions selected by for each study individually. Qualitative data was also obtained by means of different app-specific questions integrated as open answers in questionnaires, as structured interviews or as workshop-topics. What we have learned during the evaluations was that participants who use a new technology during their normal working hours are rather positive about spending time in group discussions or interviews, but at the same time are very reluctant in filling out questionnaires. Thus the number of users providing a full feedback are often very low, irrespective of country or organisational sector. Since reflection in general and testing MIRROR apps in particular are not part of our users' primary work processes, also giving feedback to the used apps had rarely priority. Thus, in hindsight we know that getting valid sample sizes of at least 30 participants who volunteer to actively use the apps over a longer period of time and are prepared to give thoughtful feedback is a great challenge. From this perspective also the conclusions drawn from the evaluations are to be viewed as tentative. More conclusive interpretations can be drawn from the overall analyses reported in D1.7, as the samples including users from different studies are naturally larger. However, these overall conclusions do not cover such a variety of aspects related to reflective learning at work as the individual studies do.

One way to overcome the challenge of interpreting qualitative reflection data from different applications in a comparable way, was the development of a new measurement instrument.

With the reflection coding scheme for the analysis of content created in the apps such an instrument was developed which provided us with valuable insights about how reflection takes place in the apps. The content analysis showed that most text entries provided by the participants are descriptions of and reflection on experiences. In most evaluations where the content was coded we found also – although to a notably smaller part – higher levels of reflection in terms of learning or change resulting from reflection. It should be noted that this only relates to user generated content in the apps and does not allow making conclusions on the learning that took place in these evaluations in general. More aggregated insights about the application of the coding scheme are described in D1.7.

Finally, all apps tested in the evaluations reported here were scrutinized regarding their support in different phases of the reflection process. As each app has its focus on different stages of the CSRL model, app-specific support was evaluated under consideration of the functionalities provided by each app. The results show that users perceive the app-specific support positively across the different stages of the CSRL model. By this wide range starting from apps capturing data and initiating reflection over apps supporting the conduction of reflection session to apps which support the follow-up of reflection outcomes we also got insights about the whole process of reflection. Deeper analysis of these insights can be found in deliverables WP1.6, which focuses on the CSRL model as well WP 3-8 in which the apps developed in the respective WP are explicitly related to the CSRL model.

Overall, we can conclude that the sum of evaluations lead to valuable insights about how reflective learning can be introduced in organisations and what factors may affect the success. These factors for success also include a careful selection of the target users, as we found out that some approaches might differ in their value for different target groups. Whereas some apps support especially newcomers in their jobs as e.g., the virtual tutors, the data of other apps might be more useful for more experienced workers (e.g. CaReflect or Watchit). Similar, some apps are better introduced as kind of campaign, that is for only a short duration in which reflective processes about specific topics are triggered (e.g. KnowSelf or Serious Games),

whereas others are most supportive when implemented for continuous usage (e.g. Issue Articulation and Issue Management, MoodMapApp, or TalkReflect).

In this deliverable the goal was to report the results on the level of individual evaluations. An aggregation of the data and insights is the focus of the other summative evaluation deliverable, D1.7. For D1.7 the data of the evaluations reported here as well as the insights and experiences of the individual partners of the project have been considered, integrated, and summarized.

9 Annex 1: Evaluation Toolbox

In this section the Core questions of the summative evaluation toolbox according to D1.5 are described. For additional question we refer to D1.5.

9.1 Demographic Information

Some data about the participants is necessary to connect the participant data across different implementations of an app (or different apps). Each question in this and the following sections is labelled with question identifiers to allow for data integration. Here, **CD** stands for **C**ore Question **D**emographic items.

CD1 Participant ID

The Participant ID consists of the first letter of the participant's place of birth, the first letter of the participant's father's first name, the first letter of the participant's mother's first name, and the participant's own day of birth (two digits). If any of these elements are unknown, the participant uses a placeholder (X for the first unknown, Y for the second, Z for the third).

The following text should be used on every questionnaire administered:

Please write down your Participant Code. Your code consists of:

1. The first letter of your place of birth
2. Your own day of birth (two digits)
3. The first letter of your father's first name
4. The first letter of your mother's first name

Example: A person born in London on the 7th of July, with parents named Jake and Sue and born would enter: "L" (for London) in the 1st blank, and "07" (for a birthday on the 7th) in the 2nd blank, "J" (for Jack) in the 3rd blank, and "S" (for Sue) in the 4th blank. So, this person's code would be: **L 07 J S**.

If you don't know any of these, use the letter below the correct box shown below.

	1st letter of your place of birth	Your day of birth	1st letter of your father's first name	1st letter of your mother's first name
Your Code				
If unknown, use:	*X*	*00*	*Y*	*Z*

CD2 Team-ID

Only if the app is used in teams: Assign a common ID if the app is used within teams in test beds (e.g., "team A", "team B").

CD3 Current Date (dd-mm-yyyy)

CD4 **Gender** (1 = Male, 2 = Female), might not be allowed in all test beds

CD5 **Age** (range: 1 = ≤19, 2 = 20-29, 3 = 30-39, 4 = 40-49, 5 = 50-59, 6 = ≥60)

CD6 **Job Scope** (1 = Full-time, 2 = Part-time)

CD7 **Department**

CD8 **Position**

CD9 **Years in current position**

CD10 **Years in current team** (if applicable)

CD11 **Years in similar positions** (e.g., at another company)

9.2 Level 1: Reaction (Usage)

Usage data may be received either from log files or from self-report questions.

Core Question Log **F**ile (CF)

CF1 **Log File Data: Number of times used**

CF2 **Log File Data: Total time (minutes) used**

CF3 **Log File Data: Average time (minutes) used**

CF4 **Log File Data: Number of times each key function of app is used**

Core Question **U**sage (CU)

CU1 **Self-Report: Number of times used**

How many times have you used [the app]?

CU2 **Self-Report: Total time (minutes) used**

How many minutes did you spend using [the app] in total?

CU3 **Self-Report: Average time (minutes) used**

How many minutes did you spend using [the app] on average?

CU4 **Self-Report: Number of times app is used completely**

How many times did you use [specific function] of the app? (repeat for each key function)

9.3 Level 2: Learning

9.3.1 App specific reflection questions

Core Question **A**pp-Specific Reflection Question (CA)

ID	Question	strongly disagree	disagree	neutral	agree	strongly agree
CA1	[The app] helped me to collect information relevant to reconstructing experiences from work.	☐	☐	☐	☐	☐
CA2	[The app] helped me to reflect on experiences from work.	☐	☐	☐	☐	☐
CA3	[The app] helped me to collect data on behaviour before the reflection session.	☐	☐	☐	☐	☐
CA4	[The app] helped me to collect data on behaviour after the reflection session.	☐	☐	☐	☐	☐
CA5	[The app] helped me to collect information that could help me decide when to	☐	☐	☐	☐	☐
CA6	[The app] helped me to reconstruct a work experience.	☐	☐	☐	☐	☐
CA7	[The app] helped me by capturing my reflection outcomes.	☐	☐	☐	☐	☐
CA8	[The app] helped me by making reflection outcomes available for later use	☐	☐	☐	☐	☐
CA9	[The app] helped me by capturing information for evaluation of learning/reflection.	☐	☐	☐	☐	☐
CA10	[The app] helped me by reminding me to reflect.	☐	☐	☐	☐	☐
CA11	[The app] helped me by providing information relevant for the decision to reflect.	☐	☐	☐	☐	☐
CA12	[The app] helped me by providing accurate information about my work.	☐	☐	☐	☐	☐
CA13	[The app] helped me by providing information relevant for the framing of reflection.	☐	☐	☐	☐	☐
CA14	[The app] helped me by showing the availability of resources needed for reflecting.	☐	☐	☐	☐	☐
CA15	[The app] helped me to allocate or structure the resources needed for reflection.	☐	☐	☐	☐	☐
CA16	[The app] helped me by providing information about related experiences.	☐	☐	☐	☐	☐
CA17	[The app] helped me to remember and reconstruct the experience/situation.	☐	☐	☐	☐	☐
CA18	[The app] helped me by providing access to data (e.g., simulations) relevant to the re-evaluation of experience.	☐	☐	☐	☐	☐
CA19	[The app] helped me by providing access to data relevant to the experience	☐	☐	☐	☐	☐

CA20	[The app] helped me by providing access to resources resulting from reflection sessions.	☐	☐	☐	☐	☐
CA21	[The app] guided me in capturing information about my work experiences.	☐	☐	☐	☐	☐
CA22	[The app] guided me in deciding whether/when to reflect.	☐	☐	☐	☐	☐
CA23	[The app] guided me in finding the resources needed for reflection.	☐	☐	☐	☐	☐
CA24	[The app] guided me in allocating/structuring the resources needed for reflection.	☐	☐	☐	☐	☐
CA25	[The app] helped me by supporting sharing of experiences.	☐	☐	☐	☐	☐
CA26	[The app] guided me in sharing experiences with others.	☐	☐	☐	☐	☐
CA27	[The app] guided me in reconstructing and remembering the experience/situation.	☐	☐	☐	☐	☐
CA28	[The app] guided me in articulating the meaning of an experience.	☐	☐	☐	☐	☐
CA29	[The app] guided us in negotiating the meaning of an experience.	☐	☐	☐	☐	☐
CA30	[The app] guided us in documenting different viewpoints on the experience.	☐	☐	☐	☐	☐
CA31	[The app] guided me in re-evaluating an experience.	☐	☐	☐	☐	☐
CA32	[The app] guided me in reaching a resolution.	☐	☐	☐	☐	☐
CA33	[The app] guided me in making the reflection outcome applicable to my work.	☐	☐	☐	☐	☐
CA34	[The app] guided me in making the reflection outcome applicable to further reflection.	☐	☐	☐	☐	☐
CA35	[The app] guided me in considering constraints of the reflection outcome.	☐	☐	☐	☐	☐
CA36	[The app] guided me in considering the option of not applying the reflection outcome.	☐	☐	☐	☐	☐
CA37	[The app] guided me in describing work scenarios that could lead to desired results.	☐	☐	☐	☐	☐
CA38	[The app] guided me in describing both “good practice” and “bad practice” work scenarios.	☐	☐	☐	☐	☐
CA39	[The app] provided help with collaboration.	☐	☐	☐	☐	☐
CA40	[The app] provided relevant content for reflection.	☐	☐	☐	☐	☐
CA41	[The app] guided me through the reflection process.	☐	☐	☐	☐	☐
CA42	[The app] helped me by simulating the work process.	☐	☐	☐	☐	☐
CA43	[The app] helped me by providing me with virtual experience in my work domain.	☐	☐	☐	☐	☐

9.3.2 Short Reflection Scale

Core Question Short **R**eflection Scale (CR)[21]

ID	Question	strongly disagree	disagree	neutral	agree	strongly agree
CR1	I often reflect on my work in order to improve it.	☐	☐	☐	☐	☐
CR2	We as a team often reflect on our work in order to improve it.	☐	☐	☐	☐	☐
CR3	I think it is important to try to improve [specific work task].	☐	☐	☐	☐	☐
CR4	I frequently reflect on [specific work task].	☐	☐	☐	☐	☐
CR5	Reflecting on [specific work task] helps me to improve [the task].	☐	☐	☐	☐	☐
CR6	In team meetings we frequently talk about how we can improve [specific work task].	☐	☐	☐	☐	☐
CR7	Outside of meetings, I often talk with my colleagues about [specific work task].	☐	☐	☐	☐	☐
CR8	It is important to me to discuss frequently with others about [specific work task].	☐	☐	☐	☐	☐
CR9	Conversations with colleagues help me to improve [specific work task].	☐	☐	☐	☐	☐
CR10	Even a few days later, I can remember the [specific work task/event] well when I reflect on it by myself or with others.	☐	☐	☐	☐	☐

Subscale individual reflection: CR1, 3, 4, 5, 10

Subscale team reflection: CR 2, 6, 7, 8, 9

[21] The phrases in square brackets have been replaced with the test bed's relevant work task(s).

9.3.3 Learning Outcomes

Core Question **L**earning (CL)

ID	Question	strongly disagree	disagree	neutral	agree	strongly agree
CL1	I made a conscious decision about how to behave in the future.	☐	☐	☐	☐	☐
CL2	I gained a deeper understanding of my work life.	☐	☐	☐	☐	☐

9.4 Level 3: Behaviour

Core Question **B**ehaviour (CB)

ID	Question	strongly disagree	disagree	neutral	agree	strongly agree
CB1	The app helped me improve my [work performance].	☐	☐	☐	☐	☐

9.5 Level 4: Results

Change over time in testbed-relevant organizational-level KPIs should be assessed. The KPIs should be measured starting at the first implementation until the end of (or even shortly after) all implementations. Care should be taken that only the relevant unit(s)/personnel who used the apps are assessed, in order to more accurately detect changes. It is important to note that context factors that cannot be controlled or influenced by the apps during the project can dampen changes that could be observed; for this reason, precision in measurement is key.

As the KPIs depend highly on the testbed researched we do not list any KPIs here. The observed KPIs are therefore described in every individual evaluation report.

10 Annex 2: Outline of document structure

Section 7.1 contains the template that was used to format the results of summative evaluations as reported in section 5 of this document. It is based on the structure of D10.1 and D10.2 but adapted to the requirements of the report of summative evaluations.

10.1 The <Application Partner> Evaluation of the <App name> App

Please leave the headers (all three levels) in this template intact and touch in your text all the questions that are applicable and relevant for your evaluation.

Please define the type of evaluation: <short-term summative evaluation** (interventions of one or a few days but summative methodology) vs. **long-term summative evaluation** (interventions over several weeks or months with summative methodology) **>

10.2 Organisational context

Test bed organisation and the organisational unit

Describe the department, institute, ward etc. where the evaluation was performed. What is the business function/services it offers to customers? Where is it located, how many employees, etc.? What is the larger organization it belongs to?

Test users and their job roles

What kind of users used the MIRROR apps for reflective learning? What are their formal job roles? What part of the job roles was affected by the evaluation? How is their workday structured, how are they managed etc.?

Identified need and potential for reflective learning

How are new hires trained currently? What role does on the job learning play? Is there some kind of coaching? How are problems at work handled and solutions sought and approved? Is there an open culture encouraging employees to create innovative solutions, to discuss them and to present them to management? Is management open for change initiated by employees? Are there already examples of such change proposals that have been implemented on an organizational level? Are there team sessions with open discussion and functioning decision processes? Are there obvious needs for more learning by reflection voiced by the employees or their managers? What is the potential for reflective learning (based on the voiced needs and other potential identified by the researchers / consultants)?

Potential organizational impact

What organizational knowledge artefacts were planned to be created or updated? What role was organizational decision makers expected to play? If this reflective learning setup would be rolled out in a large part of the organisation, what positive changes did you hope for? What organisational Key Performance Indicators might be influenced (like employee satisfaction, productivity, reduced error rates, innovation, customer satisfaction, sales turnover).

10.3 Theoretical assumptions

Please try to answer all questions within each subsection.

Selected approach for reflective learning and description of the app

Which approach did you use for reflective learning? Which potential does it address? What automatic recording and what data entry/note taking were introduced? How was it embedded in work routines? How was data aggregated and presented/ visualized? How was data interpreted and discussed? How were conclusions created and documented? Was the reflection done individually or collaboratively? What does the app do?

Relation to MIRROR CSRL Model

How did the chosen reflective learning approach relate to the Computer Supported Reflective Learning model? What reflection cycle phases were performed, were they supported by software (apps)? Which transitions were supported? How did you evaluate if the tool helped the user to conduct these phases? You may find it helpful here to use the one-page overview of types of tool use vs. the reflection cycle phases (v1.2.1 version of the CSRL model) at:

https://drive.google.com/?tab=wo&authuser=0#folders/0B96dI87ecMbFRDBSSW1STldWOEE

Reflection levels (Individual, Collaborative, Organizational) and expected organizational effects of app usage

Which levels of learning by reflection are supported by the App? [Please describe which level of reflection the App supports (individual, collaborative, organizational)]

Transition model:

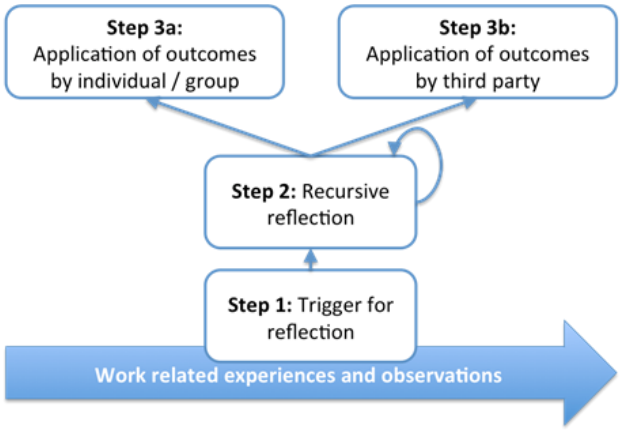

Figure 10.3.1: Transition model. One salient work experience triggers (step 1) a reflection process (either individual or collaborative reflection). The reflection outcomes may lead to consecutive reflection processes (recursion into step 2). The outcomes can either be applied by the reflection participants (step 3a) or by third parties (step 3b). Source: Prilla, Pammer, & Balzert (2012).

Considering the transition model in Figure 10.3.1, which of the depicted steps are supported by the App and how?

Regarding the transfer of information, reflection outcomes, etc. between individuals, groups, or organizational layers, does the App follow a pull or push mechanism[22] and how?

10.4 Research approach

Design and procedure

How was the evaluation designed? Did you collect data as a baseline measure? Did you have a control group? When was the evaluation done? When was the data collected? How long was the app used/was the intervention? Who did introduce the app to the participants and how?

Participants

How many participants were in the experimental group (and in the control group)? Describe shortly the demographics of your participants (mean age, number of men and women, job, mean years in that position).

Summative evaluation methods used

What summative evaluation methods did you use? Please refer to the toolbox (section 3) and describe if and how you adapted it for your evaluation.

10.5 **Results**

Please describe here the results according to the 4 levels of the evaluation framework. If you report on quantitative data please indicate mean and standard deviation (and/or median). For a better readability you may use tables and graphs. You may use the Guidelines Analysis of Results for suggestions of presentation. Please contact us if you need help with inferential statistics.

10.5.1 Level 1: Reaction (Usage)

To what degree did participants react favourably to the app?

Please describe the main indicators of usage: How often did the participants use the app (mean, standard deviation)? How long did they use the app in total (mean, standard deviation)? How long was the average usage (mean, deviation)? How often were different functions of the app used (mean, standard deviation)? In which situations and for which purposes was the app used?

[22] **"Push" mechanism:** if those individuals or groups initiate the necessary communication (reflection session) that have the impression that more people within the organization need to be involved in order to effectively solve a problem. Push mechanisms fail in changing things on a group or organizational level (barrier between the knowledge created and those enabled to apply it).
"**Pull"-mechanism:** could shift the burden of communication and enable individuals or groups to easily share triggers, salient work experiences that could be the starting point for reflective learning or any outcome of their reflection. Initiation of follow-up activities is shifted to third parties, who are responsible for making the shared knowledge applicable.
Both mechanisms can – and sometimes must - complement each other, e.g. when knowledge is promoted to a certain group level via a push mechanism and then has to be transferred into a pull mechanism to be recognized by the third party responsible or enabled to apply it.

If you have additional data about how motivated participants were to use the app, how satisfied they were with the app and data about their inclination of long-term usage, please describe that also here.

What kind of barriers occurred? What did participants hinder to use the app (e.g. motivation, manager restrictions, technical problems...)? Whatever kind of data you have (quantitative, qualitative, e.g. answers from interviews) – please report that here.

10.5.2 Level 2: Learning

App-Specific Reflection Questions, Short Reflection Scale, Learning Outcomes

To what degree did participants acquire knowledge, skills, attitudes, confidence, and commitment due to app usage?

10.5.2.1 Learning Process

Did the app help to support certain aspects of the reflection model? Which aspects did it support, which not? Report here results of the app-specific questions. Report also results on the control items if you used some.

Did the usage of the app result in a change of reflection (individual and team)? Report here about the pre-post comparison of the reflection scale if possible. Are there differences between team or/and individual reflection? How did reflection change in the experimental group compared to the control group? Is there a main effect on reflection due to the app? Are the differences between certain groups of users? You could report here on correlations between the short reflection scale or app-specific questions on the one side and participant characteristics (e.g., personality, attitudes, years in job) or usage indicators on the other side. You can also report means for different groups of users (e.g., age group).

10.5.2.2 Learning Outcomes

How did participants respond to the questions regarding learning outcomes? Report here the results of the corresponding core questions (CL) and of additional data you collected regarding learning and learning outcomes like the qualitative analysis of reflection notes and notes of reflections outcomes.

10.5.3 Level 3: Behaviour

To what degree do people apply what they learn?

If learning occurred (Level 2), were the new knowledge and skills put to use? Report here the results regarding the corresponding core question (CB). Do you have additional data regarding the behaviour of the participants? Report here also on employee-relevant test bed KPIs on an individual level if possible. If the application of reflection outcomes was somehow documented report on that. (e.g. time spent on task). The results for ranking questions regarding the application of reflection outcomes should also be reported in this section.

10.5.4 Level 4: Results

To what degree do targeted outcomes occur as a result of MIRROR?

If possible report here on the development in organizational-level KPIs for the group (ward/team) you tested over the time your app was used. Report on differences between this group and other groups in the organization which didn't use the app.

You could also report here on the potential uptake of the community. E.g. Requests from other organizations to use the apps, percentage of people willing to use the MIRROR apps (of those we offered to use them.) If you used a loyalty metric you can also report the results here: How many users would recommend your app to others? You can report here the percentage of promoters (number of answers 9-10), passives (score 7-8) and detractors (score 0-6). Report also the NPS (net promoter score): This is the percentage of promoters minus the percentage of distractors.

10.6 Conclusion & Discussion

Did the evaluation reveal what you expected? Where did you not find the expected results? Do you have an explanation for the results? What are the strengths and challenges of the app in this test bed? If and how has the situation at the test bed changed due to the MIRROR intervention?

Summary table (to be reported for every evaluation; if you report about evaluations in several test beds/groups together please add as many columns as you need)

	Experimental group	**(control group)**
Duration of intervention	Time in days/weeks	Time in days/weeks
Number of participants	Number	Number
Numbers of time used	Number	-
Total time (minutes) used	Time in minutes	-
Average time (minutes) used	Time in minutes	-
Number of times each key function of app is used	Number	-
App specific reflection questions – mean over all CA questions asked	M (SD)	-
Single means for each app specific reflection question (each in one row)	M (SD)	-
App specific control questions – mean over all control questions	M (SD)	-
Baseline (pre) Short reflection scale – mean	M (SD)	M (SD)
Baseline (pre) Short reflection scale – subscale individual reflection	M (SD)	M (SD)
Baseline (pre) Short reflection scale – subscale team reflection	M (SD)	M (SD)
Post Short reflection scale – mean	M (SD)	M (SD)
Post Short reflection scale – subscale individual reflection	M (SD)	M (SD)
Post Short reflection scale – subscale team reflection	M (SD)	M (SD)
Learning outcomes (CL1)	M (SD)	-
Learning outcomes (CL2)	M (SD)	-
Behavior (CB1)	M (SD)	-
Baseline KPI measurements		
Post KPI measurements		

11 Annex 3: Data overview tables

This section contains means and standard deviations for all evaluations for the core questions of the summative evaluation toolbox (D1.5) if available.

11.1 The KnowSelf and ARA Evaluations

	KnowSelf Evaluation at Infoman	***Knowself/ARA App Evaluation at IMC***
Duration of intervention	*30 days/6 weeks*	*6 weeks (30 work days)*
Number of participants	*12 (Details: 12 for pre-questionnaire,10 for post-questionnaire, 7 participated in interviews and sent their log data*	*10 (all 10 used the ARA App; 5 of them used the KnowSelf App as a tracking tool, 5 ManicTime)*
Numbers of time used	*215*	*-*
Total time (minutes) used	*270 minutes (4 hours 30 minutes)*	*-*
Average time (minutes) used	*Per person: 38.54 minutes* *Average 6.42 minutes per week* *Per app usage: 1.26 minutes*	*-*
Number of times each key function of app is used	*Total Number of diary entries: 143* *Total Number of Activities recorded: 9*	*-*
App specific reflection questions – mean over all CA questions asked	*M=3.28 (SD=0.66)*	***KnowSelf / ManicTime*** *M=3.28/ 2.68 (SD=0.81/0.67)* ***ARA**: M=3.30 (SD=0.83)*
Single means for each app specific reflection question (each in one row)	*CA1: M=3.40 (SD=0.97)* *C7: M=3.20 (SD=0.92)* *CA10: M=3.40 (SD=1.08)* *CA12: M=3.00 (SD=1.16)* *CA40: M=3.40 (SD=0.70)*	***KnowSelf / ManicTime*** *CA1: M=3.60/3.20 (SD=0.89/1.09)* *C7: M=3.00/2.00 (SD=1.23/1.23)* *CA10: M=3.20/2.20 (SD=1.30/0.84)* *CA12: M=3.60/3.60 (SD=1.14/0.89)* *CA40: M=3.00/2.40 (SD=1.00/1.52)* ***ARA:*** *CA2: M=3.70 (SD=0.82)* *CA6: M=3.00 (SD=1.16)* *CA7: M=3.50 (SD=1.18)* *CA8: M=3.00 (SD=1.56)*

	KnowSelf Evaluation at Infoman	***Knowself/ARA App Evaluation at IMC***
App specific control questions – mean over all control questions	*CA16: M=2.10 (SD=0.88)*	***KnowSelf/ManicTime*** *CA16: M=2.20/1.00 (SD=1.30/0.00)* ***ARA:*** *CA42: M=1.50 (SD=0.85)*
Baseline (pre) Short reflection scale – mean	*M=3.26 (SD=0.65) α=0.769*	*M=2.97 (SD=0.34)*
Baseline (pre) Short reflection scale – subscale individual reflection	*M=3.72 (SD=0.85) α=0.77*	*M=3.50 (SD=0.69); α=0.80*
Baseline (pre) Short reflection scale – subscale team reflection	*M=2.80 (SD=0.79) α=0.77*	*M=2.44 (SD=0.64); α=0.65*
Post Short reflection scale – mean	*M=3.37 (SD=0.92) α=0.924*	*M=3.11 (SD=0.57); α=0.824*
Post Short reflection scale – subscale individual reflection	*M=3.72 (SD=1.12) α=0.946*	*M=3.78 (SD=0.53); α=0.588*
Post Short reflection scale – subscale team reflection	*M=3.02 (SD=0.88) α=0.825*	*M=2.44 (SD=0.73); α=0.817*
Learning outcomes (CL1)	*M=3.30 (SD=1.06)*	*M=4.30 (SD=0.68)*
Learning outcomes (CL2)	*M=3.50 (SD=0.71)*	*M=2.80 (SD=1.03)*
Behavior (CB1)	*M=3.40 (SD=1.08)*	*M=4.10 (SD=0.74)*
Baseline KPI measurements	*Time management scale (12 items)* *M=3-60 (SD=0.61) α=0.78* *Time wasters scale (12 items)* *M=1.89 (SD=0.33) α=0.67*	*Time management scale (12 items)* *M=3.05 (SD=0.45); α=0.67*
Post KPI measurements	*Time management scale (12 items)* *M=3.58 (SD=0.40) α=0.607* *Time wasters scale (12 items)* *M=2.10 (SD=0.45) α=0.812*	*Time management scale (12 items)* *M=3.48 (SD=0.39); α=0.622*

11.2 The MoodMap App evaluations

	MoodMap App evaluation at BT	***MoodMap App evaluation at Regola***
Duration of intervention	*31 days*	*7 weeks (34 working days)*
Number of participants	*67/53 (total/app)*	*35*
Numbers of time used	*1914 interactions*	1767 *interactions*
Total time (minutes) used	*13582 minutes (with interruptions)*	*22774 minutes (with interruptions)*
Average time (minutes) used	*8 h. 42 min. (all users in a day)*	*10 h. 32 min. (all users in a day)*
Number of times each key function of app is used	*991 moods, 991 contexts, 946 notes*	2250 *moods, 31 contexts, 226 notes*
App specific reflection questions – mean over all CA questions asked	*M* = 3,23 (*SD* = 0.9)	*M = 2.87 (SD = 0.80)*
Single means for each app specific reflection question (each in one row)	CA01: *M* = 3.34 (*SD* = 0.97) CA02: *M* = 3.46 (*SD* = 1.04) CA05: *M* = 3.06 (*SD* = 1.00) CA07: *M* = 3.14 (*SD* = 1.03) CA08: *M* = 3.09 (*SD* = 1.09) CA17: *M* = 3.23 (*SD* = 1.09) CA21: *M* = 3.29 (*SD* = 1.05)	CA01: *M* = 2.91 (*SD* = 0.78) CA02: *M* = 2.94 (*SD* = 0.92) CA05: *M* = 2.82 (*SD* = 1.00) CA07: *M* = 3.06 (*SD* = 1.07) CA08: *M* = 2.62 (*SD* = 0.89) CA17: *M* = 2.85 (*SD* = 1.02) CA21: *M* = 2.91 (*SD* = 0.93) CA41: *M* = 2.74 (*SD* = 0.99)
App specific control questions – mean over all control questions	*M* = 2.54 (*SD* = 1.15)	*M* = 2.82 (*SD* = 0.97)
Baseline (pre) Short reflection scale – mean	*M* = 3.98 (*SD* = 0.49)	*M* = 3.78 (*SD* = 0.40)
Baseline (pre) Short reflection scale – subscale individual reflection	*M* = 4.01 (*SD* = 0.56)	*M* = 4.10 (*SD* = 0.37)
Baseline (pre) Short reflection scale – subscale team reflection	*M* = 3.94 (*SD* = 0.52)	*M* = 3.47 (*SD* = 0.58)
Post Short reflection scale – mean	*M* = 3.89 (*SD* = 0.53)	*M* = 3.72 (*SD* = 0.46)
Post Short reflection scale – subscale individual reflection	*M* = 3.96 (*SD* = 0.61)	*M* = 4.00 (*SD* = 0.45)

	MoodMap App evaluation at BT	***MoodMap App evaluation at Regola***
Post Short reflection scale – subscale team reflection	*M* = 3.81 (*SD* = 0.57)	*M* = 3.45 *(SD* = 0.61*)*
Learning outcomes (CL1)	*M* = 2.91 (*SD* = 1.09)	*M* = 2.62 *(SD* = 1.04*)*
Learning outcomes (CL2)	*M* = 2.86 (*SD* = 1.14)	*M* = 2.71 *(SD* = 0.97*)*
Behavior (CB1)	*M* = 2.77 *(SD* = 1.27*)*	*M* = 2.53 *(SD* = 1.02*)*
Baseline KPI measurements .	*KPI1a: Team GMa Volume* *M* = 10.68 (*SD* = 4.08) *KPI2a: Team GMa Average rating.* *M* = 82.79 (*SD* = 7.98) *KPI1b: Team STh Volume.* *M* = 27.65 (*SD* = 17.00) *KPI2b: Team STh Average rating.* *M* = 82.82 (*SD* = 8.24) KPI3-5 (GMa/STh): *KPI3: NPI.:* 25 / 30 *KPI4: Advisor Sat.:* 84 / 84 *KPI5: Recap.:* 85 / 79	*job satisfaction* *M* = 3.31 *(SD* = 1.34*)* *improve_indiv_work* *M* = 3.52 *(SD* = 0.74*)* *improve_team_performance* *M* = 3.76 *(SD* = 0.64*)*
Post KPI measurements	*KPI1a: Team GMa Volume.* *M* = 10.05 (*SD* = 3.95) *KPI2a: Team GMa Average rating.* *M* = 89.58 (*SD* = 5.81) *KPI1b: Team STh Volume.* *M* = 24.41 (*SD* = 17.91) *KPI2b: Team STh Average rating.* *M* = 83.00 (*SD* = 11.86) ***KPI3-5 (GMa/STh):*** *KPI3: NPI.:* 35 / 35 *KPI4: Advisor Sat.:* 90 / 85 *KPI5: Recap.:* 89 / 82	*job satisfaction* *M* = 3.79 *(SD* = 1.08*)* *improve_indiv_work* *M* = 3.28 *(SD* = 0.92*)* *improve_team_performance* *M* = 3.14 *(SD* = 0.83*)*

11.3 The Talk Reflect Evaluations at RNHA, NBN and RBCK

Talk Reflect Evaluations	***E1 (NBN)***	***E2 (Care)***	***E3 (Interns)***	***E4 (RBKC)***
Duration of intervention	*42 days*	*50 days*	*51 days*	*80 days*
Number of participants: Quest/logdata	*5 / 9*	*5 (0 post) / 9*	*8 (0) / 18*	*12 / 12*
Numbers of time used	*See Table 5.5.6*	*See Table 5.5.6*	*See Table 5.5.6*	*See Table 5.5.6*
Total time (minutes) used (summed up over all participants)	*N/A*[23]	*N/A*	*N/A*	*N/A*
Average time (minutes) used *per person* *per quiz*	*N/A*	*N/A*	*N/A*	*N/A*
Number of times each key function of app is used	*See Table 5.5.6*	*See Table 5.5.6*	*See Table 5.5.6*	*See Table 5.5.6*
App specific reflection questions – mean over all CA questions asked	*M=2.9 (SD=0.42)*	*-*	*-*	*M=2.95 (active 3.15)* *SD=0.15 (0.2)*
Single means for each app specific reflection question (each in one row)	*CA1: M=3.60 (SD=0.49)* *CA2 ind: M=2.60 (SD=0.80)* *CA2 coll: M=2.60 (SD=0.49)* *CA8: M=2.60 (SD=0.80)* *CA33: M=2.80 (SD=0.75)*	*-*	*-*	*CA1: M=3.13 (active 2.5); SD=0.99 (1.29)* *CA2 ind: M=3 (3.25); SD=0.93 (1.26)* *CA2 coll: M=2.88 (3); SD=0.83 (1.15)* *CA8: M=2.88 (3); SD=0.99 (1.41)* *CA33: M=2.88 (3.5); SD=1.05 (1.0)*
App specific control questions – mean over all control questions	*-*	*-*	*-*	*-*

[23] We could not log the time the app was used in, as this can only be estimated: for normal usage there is no clear end of using the given or implicated. For example, even after minutes of no activity to be logged by the system it could be the case that a user is- reading an experience report.

Talk Reflect Evaluations	***E1 (NBN)***	***E2 (Care)***	***E3 (Interns)***	***E4 (RBKC)***
Baseline (pre) Short reflection scale – mean	M=3.50 (SD=0.58)	*M=4.19 (SD=0.43)*	*M=3.58 (SD=0.92)*	*M=4.0 (SD=0.45)*
Baseline (pre) Short reflection scale – subscale individual reflection	M=4.05 (SD=0.19)	*M=4.46 (SD=0.23)*	*M=4.19 (SD=0.38)*	*M=4.14 (SD=0.39)*
Baseline (pre) Short reflection scale – subscale team reflection	M=3.0 (SD=0.41)	*M=3.96 (SD=0.49)*	*M=2.93 (SD=0.86)*	*M=3.83 (SD=0.53)*
Post Short reflection scale – mean	*M=3.34 (SD=0.61)*	-	-	M=3.62 (SD=0.32)
Post Short reflection scale – subscale individual reflection	*M=3.80 (SD=0.49)*	-	-	M=3.9 (SD=0.23)
Post Short reflection scale – subscale team reflection	*M=2.88 (SD=0.36)*	-	-	M=3.4 (SD=0.25)
Learning outcomes (CL1)	*M=2.00 (SD=0.63)*	-	-	M=2.4 (active 2.75); SD=0.88 (active 0.96)
Learning outcomes (CL2a)	*M=2.00 (SD=0.63)*	-	-	M=2.78 (3.25); SD=1.05 (0.96)
Learning outcomes (CL2b)	-	-	-	M=3.11 (3.75); SD=1.05 (0.5)
Behavior (CB1)	*After using the app I can explain things better to relatives M=2.4 (SD=1.02)*	-	-	*M=2.67 (3.0); SD=0.87 (0.82)*
Baseline KPI measurements	-	-	-	-
Post KPI measurements	*By using the app, we could better coordinate the communication with relatives among us.* M=2.8 (SD=0.40) *By using the app, we could identify problems and good practice, by which we can*	-	-	*My superiors have been more satisfied with my work since I have been using the app. M=2.33 (active 2.5); SD=0.87 (0.58)* *I can better deal with situations and problems at*

Talk Reflect Evaluations	*E1 (NBN)*	*E2 (Care)*	*E3 (Interns)*	*E4 (RBKC)*
	communicate better with relatives. M=2.8 (SD=0.40) *By using the app, we could identify problems and good practice that are also relevant for other colleagues (outside the ward).* M=2.20 (SD=0.75) *After using the app for the last 4 weeks, I have a better structure for conversations with relatives.* M=2.60 (SD=1.02) *After using the app for the last 4 weeks, relatives have more faith in me when I talk to them.* M=2.00 (SD=0.63) *After using the app for the last 4 weeks, I can guide relatives better in conversations (finding the right words, reducing complaints)* M=2.40 (SD=1.03)			*work since I have been using the app.* *M=2.56 (3); SD=1.01 (0.82)* *I can better work together with colleagues since I have been using the app.* *M=2.67 (active 3.25); SD=1.12 (0.96)*

11.4 The DoWeKnow Evaluation at Infoman

DoWeKnow Evaluation	Infoman
Duration of intervention	*10 weeks*
Number of participants	*10*
Numbers of time used	*n.a.*
Total time (minutes) used	*n.a.*
Average time (minutes) used	*n.a.*
Number of times each key function of app is used	*57 Comments* *55 Ratings*
App specific reflection questions – mean over all CA questions asked	*4.05 (0.11)*
Single means for each app specific reflection question (each in one row)	*CA2a: 4.00 (1.00) [.. to think about slides in a constructive way]* *CA2b: 4.00 (1.10) [.. to think about slide presentations in a constructive way]* *CA29a: 4.27 (0.79) [.. to discuss slides with colleagues]* *CA29b: 4.00 (0.89) [.. to discuss slide presentations with colleagues]* *CA7: 4.00 (0.89) [… to capture discussion results]*
App specific control questions – mean over all control questions	*n.a.*
Baseline (pre) Short reflection scale – mean	*3.71 (0.67)*
Baseline (pre) Short reflection scale – subscale individual reflection	*n.a.*
Baseline (pre) Short reflection scale – subscale team reflection	*3.71 (0.67)*
Post Short reflection scale – mean	*3.97 (0.58)*
Post Short reflection scale – subscale individual reflection	*n.a.*
Post Short reflection scale – subscale team reflection	*3.97 (0.58)*
Learning outcomes (CL1)	*4.18 (0.12)*
Learning outcomes (CL2)	*3.73 (0.11)*

DoWeKnow Evaluation	Infoman
Behavior (CB1)	*4.04 (0.06)*
Baseline KPI measurements	Marketing: 4.00 (0.38) Sales: 3.19 (0.30)
Post KPI measurements	Marketing 3.89 (0.69) Sales: 4.25 (0.22)

11.5 The Issue Articulation and Management App Evaluation at BT

IAA/IMA at BT	***Experimental group***	***(control group)***
Duration of intervention	*6 weeks?*	-
Number of participants	*57*	*10*
Numbers of time used	*3655 actions*	-
Total time (minutes) used	-	-
Average time (minutes) used	*1.875 minutes*	-
Number of times each key function of app is used	*Save Coaching Need:49* *Show Coaching Observation Form: 217* *List Coaching Needs: 178* *Show specific issue: 189*	-
App specific reflection questions – mean over all CA questions asked	*M (SD)*	-
Single means for each app specific reflection question (each in one row)	*CA2: 3.533 (0.884)* *CA7:3.333 (0.869)* *CA10: 3.6 (0.8)* *CA13: 3.266 (0.929)* *CA17: 3.4 (0.712)* *CA26: 3.466 (0.618)* *CA31: 3.666 (0.596)* *CA32: 3.466 (1.024)* *CA33: 3.333 (0.869)* *CA34: 3.429 (0.821)* *CA37: 3.266 (0.928)* *CA38: 3.2 (0.98)* *CA41: 3.333 (0.943)*	-
App specific control questions – mean over all control questions	*2.733 (0.772)*	-
Baseline (pre) Short reflection scale – mean	*4.162 (0.814)*	*Only 2 questionnaires*
Baseline (pre) Short reflection scale – subscale individual reflection	-	-

IAA/IMA at BT	***Experimental group***	***(control group)***
Baseline (pre) Short reflection scale – subscale team reflection	*-*	*-*
Post Short reflection scale – mean	*3.475 (1.013)*	*Only 1 questionnaire*
Post Short reflection scale – subscale individual reflection	*-*	*-*
Post Short reflection scale – subscale team reflection	*-*	*-*
Learning outcomes (CL1)	*3.333 (1.011)*	*-*
Learning outcomes (CL2)	*3.214 (1.081)*	*-*
Behavior (CB1)	*3.214 (1.013)*	*-*
Baseline KPI measurements	*-*	*-*
Post KPI measurements	*-*	*-*

11.6 The Medical Quiz Evaluation at NBN (Workshop and Stroke Unit)

Medical Quiz Evaluation at NBN	***Qualification Program***	***Stroke Unit***
Duration of intervention	*8 weeks*	*6 – 7 weeks*
Number of participants: Quest/logdata	*21/ 18*	*3*
Numbers of time used	*411 (finished quiz attempts)*	*11 (finished quiz attempts)*
Total time (minutes) used (summed up over all participants)	*61:04 h*	*01:16 h*
Average time (minutes) used *per person* *per quiz*	 *3:24 h* *06:57 min*	 *08:54 min* *07:48 min*
Number of times each key function of app is used	*411 (finished quiz attempts)*	*11 (finished quiz attempts)*
App specific reflection questions – mean over all CA questions asked	*M* = 3.51 (*SD* = 0.42)	*M* = 2.67 (SD = 0.68)
Single means for each app specific reflection question (each in one row)	CA6: *M* = 3.33 (*SD* = 0.84) CA8: *M* = 3.71 (*SD* = 0.71) CA9: *M* = 4.11 (*SD* = 0.58) CA10: *M* = 3.28 (*SD* = 0.96) CA11: *M* = 3.17 (*SD* = 0.51) CA12: *M* = 4.00 (*SD* = 0.69) CA22: *M* = 3.50 (*SD* = 0.62)	CA6: *M* = 2.67 (*SD* = 1.53) CA8: *M* = 2.67 (*SD* = 0.58) CA9: *M* = 3.00 (*SD* = 1.00) CA10: *M* = 2.67 (*SD* = 1.15) CA11: *M* = 2.67 (*SD* = 0.58) CA12: *M* = 2.67 (*SD* = 0.58) CA22: *M* = 2.33 (*SD* = 0.58)
App specific control questions – mean over all control questions	CA26: *M* = 3.22 (*SD* = 0.81)	CA26: *M* = 3.00 (*SD* = 1.00)
Baseline (pre) Short reflection scale – mean	*M* = 3.65 (*SD* = 0.54)	*M* = 3.53 (*SD* = 0.38)
Baseline (pre) Short reflection scale – subscale individual reflection	*M* = 4.09 (*SD* = 0.54)	*M* = 3.56 (*SD* = 0.63)
Baseline (pre) Short reflection scale – subscale team reflection	*M* = 3.22 (*SD* = 0.66)	*M* = 3.47 (*SD* = 0.23)
Post Short reflection scale – mean	*M* = 3.18 (*SD* = 0.45)	*M* = 3.20 (*SD* = 0.66)
Post Short reflection scale – subscale individual reflection	*M* = 3.57 (*SD* = 0.45)	*M* = 3.60 (*SD* = 0.53)
Post Short reflection scale – subscale team reflection	*M* = 2.79 (*SD* = 0.59)	*M* = 2.80 (*SD* = 0.80)

Medical Quiz Evaluation at NBN	***Qualification Program***	***Stroke Unit***
Learning outcomes (CL1)	*M* = 3.06 (*SD* = 0.80)	*M* = 2.67 (*SD* = 1.53)
Learning outcomes (CL2)	*M* = 3.22 (SD = 0.88)	*M* = 2.67 (*SD* = 1.53)
Behavior (CB1)	M = 3.69 (SD = 0.75)	*M* = 3.00 (SD = 1.00)
Baseline KPI measurements	-	
Post KPI measurements *KPI1: Das Quiz hat meine Arbeitszufriedenheit verbessert.* *KPI2: Nachdem ich das Quiz in den letzten Monaten gespielt habe, hat sich die Betreuung meiner Patienten verbessert.* *KPI3: Nachdem ich das Quiz in den letzten Wochen gespielt habe, sind weniger Probleme oder negative Situationen aufgetreten.*	*M = 3.45 (SD = 0.91)* *M = 2.89 (SD = 0.83)* *M = 2.67 (SD = 0.84)*	*M* = 2.33 (*SD* = 0.58) *M* = 2.67 (*SD* = 0.58) *M* = 2.33 (*SD* = 0.58)

11.7 The CaReflect App Evaluation at RNHA

CaReflect App Evaluation	RNHA
Duration of intervention	*4 days*
Number of participants: sensor usage/post shift questionnaires / concluding interview and questionnaire	*44/40/17*
Numbers of time used	*Once after each shift*
Total time (minutes) used (summed up over all participants)	*1256 Hours of data was captured*
Average time (minutes) used : *Number of contacts captured:*	*45769*
Number of times each key function of app is used	
App specific reflection questions – mean over all CA questions asked	*M=3.77 (SD=0.93)*
Single means for each app specific reflection question	*CA12: M=3.77 (SD=0.93)*
App specific control questions – mean over all control questions	-
Baseline (pre) Short reflection scale – mean	-
Baseline (pre) Short reflection scale – subscale individual reflection	-
Baseline (pre) Short reflection scale – subscale team reflection	-
Post Short reflection scale – mean	-
Post Short reflection scale – subscale individual reflection	-
Post Short reflection scale – subscale team reflection	-
Learning outcomes (CL1)	*M=3.66 (SD=0.85)*
Learning outcomes (CL2)	*M=4.03 (SD=0.55)*
Behavior (CB1)	-
Baseline KPI measurements	-
Post KPI measurements	-

11.8 The WATCHiT and WATCHiT Procedure Trainer Evaluation at Regola

WATCHiT and WATCHiT Procedure Trainer at Regola	***1st experimental group***	***2nd experimental group***
Duration of intervention	2 days	2 days
Number of participants	8	27
App specific reflection questions – mean over all CA questions asked	4.38 (0.32)	4.05 (0.42)
CA2	4.24 (0.46)	4.00 (0.62)
CA6	4.75 (0.46)	4.15 (0.62)
CA7	4.50 (0.53)	4.04 (0.53)
CA12	4.25 (0.46)	3.96 (0.52)
CA40	4.50 (0.53)	4.23 (0.51)
CA41	4.00 (0.53)	3.89 (0.64)
Baseline (pre) Short reflection scale – mean	4.28 (0.30)	4.31 (0.32)
Baseline (pre) Short reflection scale – subscale individual reflection	4.35 (0.38)	4.41 (0.30)
Baseline (pre) Short reflection scale – subscale team reflection	4.30 (0.30)	4.20 (0.38)
Learning outcomes (CL1)	4.13 (0.64)	4.07 (0.62)
Learning outcomes (CL2)	4.50 (0.76)	3.96 (0.71)
Behavior (BI06)	4.38 (0.74)	4.11 (0.51)

11.9 The CLinIC – The Virtual Tutor serious game and the Think better CARE Evaluations

Serious Games evaluations	***The CLinIC – The Virtual Tutor serious game Evaluations at the University of Bergamo***	***The Think better CARE – The Virtual Tutor Evaluation at RNHA***
Duration of intervention	*1 days*	*Game presentation in several care homes from November 2013 to March 2014.* *Users played the game usually only once.*
Number of participants	*16*	*17*
Numbers of time used	*16*	-
Total time (minutes) used	*402.45 minutes*	-
Average time (minutes) used	*25 minutes*	-
Number of times each key function of app is used	*Note function=7 times*	*Note function: 0 time used*
App specific reflection questions – mean over all CA questions asked	M=3.83 (SD=0.73)	M= 3.56 (SD=1.29)
Single means for each app specific reflection question (each in one row)	CA02: M=3.63 (SD=0.96) CA10: M=3.94 (SD=0.44) CA16: M=3.88 (SD=0.89) CA42a: M=4.06 (SD=0.57) CA42b: M=3.94 (SD=0.57) CA43a: M=3.56 (SD=0.96)	CA02: M=3.20 (SD=1.30) CA10: M=3.80 (SD=1.30) CA16: M=3.80 (SD=1.30) CA42a: M=4.00 (SD=1.22) CA42b: M=3.20 (SD=1.48) CA43a: M=3.40 (SD=1.14)
App specific control questions – mean over all control questions	-	-
Baseline (pre) Short reflection scale – mean	-	M=4.09 (SD=0.45)
Baseline (pre) Short reflection scale – subscale individual reflection	-	M=4.37 (SD=0.53)
Baseline (pre) Short reflection scale – subscale team reflection	-	M=3.81 (SD=0.62)
Post Short reflection scale – mean	M=3.90 (SD=0.50)	-

Serious Games evaluations	***The CLinIC – The Virtual Tutor serious game Evaluations at the University of Bergamo***	***The Think better CARE – The Virtual Tutor Evaluation at RNHA***
Post Short reflection scale – subscale individual reflection	M=4.25 (SD=0.44)	-
Post Short reflection scale – subscale team reflection	M=3.55 (SD=0.66)	-
Learning outcomes (CL1)	M=3.50 (SD=0.73)	M=3.20 (SD=0.84)
Learning outcomes (CL2)	M=2.94 (SD=0.85)	M=3.40 (SD=0.89)
Behavior (CB1)	M=3.75 (SD=1.00)	N=0
Baseline KPI measurements	-	-
Post KPI measurements	-	-

11.10 The Rescue League serious game Evaluation at Regola (118 emergency associations)

Recue League SG at Regola	***Croce Bianca evaluation***	***SOS Novate evaluation***
Duration of intervention	*1*	*1*
Number of participants	*19*	*14*
Numbers of time used	*19*	*14*
Total time (minutes) used	-	
Average time (minutes) used	*11.38 minutes*	*41 minutes*
Number of times each key function of app is used	*Note function: 21 times used*	*Note function: 8 times*
App specific reflection questions – mean over all CA questions asked	*M=3.97 (SD=0.57)*	*M=3.60 (SD=0.44)*
Single means for each app specific reflection question (each in one row)	CA2: M= 4.16 (SD=0.76) CA10: M= 4.21 (SD=0.85) CA12: M= 3.84 (SD=0.70) CA16: M= 3.58 (SD=1.17) CA19: M= 3.95 (SD=0.70) CA40: M= 4.26 (SD=0.65) CA42: M= 3.95 (SD=0.7) CA43a: M= 3.83 (SD=0.99)	CA2: M= 3.43 (SD=1.02) CA10: M=3.57 (SD=0.85) CA12: M= 3.64 (SD=0.63) CA16: M= 3.36 (SD=0.84) CA19: M= 3.57 (SD=0.94) CA40: M= 3.86 (SD=0.66) CA42: M= 3.79 (SD=0.58) CA43a: M= 3.570(SD=0.94)
App specific control questions – mean over all control questions	-	-
Baseline (pre) Short reflection scale – mean	M=3.9 (SD=0.48)	M=3.86 (SD=0.84)
Baseline (pre) Short reflection scale – subscale individual reflection	M=4.23 (SD=0.57)	M=4.01 (SD=0.61)
Baseline (pre) Short reflection scale – subscale team reflection	M=3.56 (SD=0.64)	M=3.71 (SD=0.47)
Post Short reflection scale – mean	-	-
Post Short reflection scale – subscale individual reflection	-	-

Recue League SG at Regola	***Croce Bianca evaluation***	***SOS Novate evaluation***
Post Short reflection scale – subscale team reflection	-	-
Learning outcomes (CL1)	M=3.63 (SD=1.1)	M=2.71 (SD=0.73)
Learning outcomes (CL2)	M=3.68 (SD=1.09	M=3.14 (SD=0.66)
Behavior (CB1)	N=0	N=0
Baseline KPI measurements	-	-
Post KPI measurements	-	-

12 Annex 4: Further material to individual evaluations

This section contains further material about the evaluations, namely extracts of interviews from the MMA evaluations. Because these extracts are rather comprehensive, they have not been included in the main reports. However, due to their value regarding the obtained qualitative data we included them here in order to not withhold the included information with very interesting insights into the personal views of users.

12.1 The MoodMap App Evaluation at BT

Interviews with Managers and Advisors

The summary of the interviews conducted with two managers and one advisor can be found below:

The manager of one team described the usage of the MoodMap App as a really good experience: *"I am a manager and I manage a team of 21 people. And the feedback I got from them and also myself, when I was using it, was very user-friendly, very easy and quick for me to go in, at different periods of the day to see how everyone was feeling"*. The manager also reported having discussed how everyone was feeling at a certain point of the day, to see if he could provide any support. Additionally they recognized that there were certain points in a day where the mood dropped, mid-afternoon and after lunch. So they were looking to see if there is anything they could put in place in order to change that.

The manager also indicated having received positive feedback from the coaches, although we could not gather any feedback in this direction through the questionnaires.

Not only problems were detected, but also positive experiences were highlighted by the manager: *"it is nice to see when people are feeling good after they have come of a call with a customer or they have mastered a deal with a customer they also like to put they have suffered with that on there, so sometimes it's nice for me to go and say well done, if I have known that they have done it."* Possible changes he was considering thanks to the MoodMap App were, if the advisors come to a huddle and their mood is low, he will change the time or maybe even the organization of the huddles themselves.

The manager also gave us some insights for uptake in other teams: *"I think it definitely improved my insight into my team. And I think it would also be useful for managers that have like a team that work away from the offices, that work at home so you can't always see them, but you can see of what kind of moods they are in - and how they're feeling and stuff. You know, if I wasn't in the office, I can still see how my team is feeling. […] It is definitely an improvement for me".*

In the second interview with the manager, he described the usage of the MoodMap App also as positive: *"It was very quickly, it made both me and advisors aware of how our mood and energy directly impacted the relationships within our team and the impressions of our customers".* Regarding the insights gained by the manager, he stated that *"As the moods were easily visible, this allows the managers to incorporate their coaching style to reflect the current mood*" and this insights made him also change his behaviour at work: *"when my energy level is low, I leave the desk and do something else until my level is up again. Then I return to my desk"*.

Finally, to the question if they would like to integrate the MoodMap App in their daily work, we received definitive affirmative answers from both managers. They also stated that the MoodMap App could be easily integrated in the advisors' work practices.

An interview with an advisor was also conducted. He discussed with the researchers some of the insights and benefits that he and his team had by using the MoodMap App. He and his team considered the experience as positive, and he remarked that: *"if you have a bad experience, a bad call, an experience with the department just you can vent how you feel on the app and then it's just drop, put it there and it is gone, you know you move on"*.

Some insights about when he used to capture his mood were also mentioned, although he admitted not having a relationship between coaching sessions and the application: *"I haven't come back of a coaching session and used it in even I had not think about it, to be honest. I used to fill it in at the end of a call or like a break or I have never achieved it after a coaching session after once a day I would fill it in probably. I have got a one-to-one today as well with my manager so I probably used it on that occasion as well".*

To the question if he had learned something by reflecting on his mood, he highlighted the fact that he could see the mood of his colleagues and compare himself to them: *"Yeah, I like the way you can see the team members as well, you can see where they are, or you sort of wonder yourself why are they there, or why are they are up there and I am down here or vice versa. So you sort of wonder and I ask these things to myself if they have to had a really bad day or have to had a bad call or just generally feeling unavailable […] So I think it is quite a good thing to look at, and I always compare myself to others"*. He also mentioned that reflecting on a certain call helped him move on to the next customer feeling better and not being affected by past negative experiences.

The advisor admitted not having used it in his coaching sessions, but being willing to use it: "*I think we should use them more in coaching, I think we should incorporate it in peer-coaching, but make it positive, don't reflect on a negative thing, […] So everybody likes positive feedback, nobody likes negative feedback, so I think that by using it in coaching you should always make it a positive…*".

One of the interesting insights we could gain in this interview is that the interviewee thought that social media was too intrusive, and therefore he did not use any of the extended social media portals. However, in the case of the MoodMap App that was not a problem at all and all his feedback was positive and considered it as being part of her work. For him it was not a problem to express himself, however he admitted having fear that other colleagues may be more hesitant to express their emotions and sharing them with managers and coaches.

12.2 The MoodMap App Evaluation at Regola

Interviews with participants

Seven interviews were conducted with the following participants: the manager of the Department Am, the manager and one of two staff members (who coordinated this evaluation from Regola's side) from the Department Pm, two staff members from the Department Co, one staff member from the Department Am, and another staff member from the Department Sv.

Generally, the feedback from both managers was very positive, as it is shown for example in the following statement "…*interesting experience as a user because in a way it is my company that becomes interested in my work. They try to be aware of my goals during my work activities and I think that it is a really positive thing …*".

The major problem of the company mentioned in the interviews is that they have all a very stressful job, they have to meet hard deadlines with customers, they are very busy at the moment and they have nearly no time to do something extra during work. One of the managers stated that "*we can't really make a business meeting to speak about this [the usage of the MoodMap App], this would be a nice thing but it was not possible, they didn't allow that we do it".* The lack of in-between meetings, where all the participants might have discussed more about the mood capturing and also where the project manager of the evaluation might have provided more insights into the MIRROR project might have led to more understanding in the company, as one manager admitted that "*they probably don't understand the positive intention of the project".*

The lack of time was also the reason for one manager to only capture his own mood, on the one hand regularly three to four times a day and when a problem occurred during work. On the other hand he did not even have had a look at other visualisations to see how the mood of his team was nor had he the time to reflect about his own mood which resulted in that "*I did not see a benefit for me".* The other manager also captured his mood regularly in the morning, before or after lunch and before leaving the office in the evening. Additionally he mentioned to state a mood, when he was much stressed. This led him to the insight that *"…when I am too much stressed maybe it is better if I stop for a moment, I take a breath and then continue to work*".

On the question if they have learned anything specific or gained any insights by having a look at and reflecting on their own mood or the mood of their teams the managers provided the following answers. One manager mentioned that his mood did not change during the day very often, but when it changed, then it was a significant change. The other manager found out for himself, that he is often in a very positive mood but had a low energy level which showed him to be often very tired during the day.

Both managers mentioned that they have seen the pop-ups, the reflection amplifiers and reflection interventions, but they did not really use them. Some cases in which they added notes to their moods were for example "*I often write a motivation reason, the reason why I changed it." or "because I was really stressed for a particular reason, in this case I wrote a comment on the mood."* Regarding the visualisations, in general one manager did not use any other visualisations at all for time reasons, while the other mentioned to prefer in general the CompareMe View, where he can see his mood and the mood of his team at a glance. He also used this view as a kind of confirmation of what he already knows about his colleagues "*I saw the mood and it was not a surprise, we work together and probably I look at the face of my colleague and I know that she is really busy or worried".* In order to really profit from the

MoodMap App or to get a clear benefit out of it, one of the managers mentioned *"… I think that we need more time to analyse the mood and to make meetings or others to speak about the mood. To speak about the moods and what is changed".* Furthermore both managers mentioned they should pay more attention to the colleagues who did not use it and explain it better and to dispel any fears in order to show them that "*it is an instrument that should help them"*.

An interview with one of the three persons who were responsible for introducing the MoodMap App at Regola gave us also very important insights regarding the preparation and conduction of such an evaluation. For him it was very difficult to convince the employees to use the MoodMap App on a regular base, but he still sees a clear benefit the MoodMap App could have for a company like Regola: "*Managers press you and they want you to finish in time. Usually they don't ask you if you are stressed, you are happy or in a good mood and low energy. I must say it is very positive that maybe someone now, I hope that our managers thinks about that also and not about only to finish the work and so. Because I think that if the staff is happy and working well the work is finished in time…".* He also tried to actively talk to the managers of the company about this trial, but "*they are very busy and it is very difficult to speak with them about something that is not the project*". He also mentioned that for the managers the most important thing is to close projects and to release projects and not to find out how their staff members are feeling. He also mentioned that his team members did not use it, because they think that nothing would happen in the end. Nevertheless he stated that "*I tried to encourage them to use it because I think it is a good idea, I told them"*. Unfortunately he himself is not a manager and therefore he could not change anything for his colleagues. He also captured his mood three times a day and when he was very stressed or sad, but he never added some notes. Regarding the pop-ups (reflection amplifiers or reflection interventions), he also never used them, but mentioned that *"but I think that it was a pity, because maybe it should be a good thing to do that, to focus on why I was feeling happier or worse, or compared to the others or compared to myself five minutes before".* Additionally he suggested that the pop-ups should not only ask for input but also make aware of interesting features of the MoodMap App or point to a view where the user can see where something is happening in the MoodMap App. Regarding the visualisations, for him "*the best one is the comparison bars that can make you understand in few seconds yourself and your situation compared to the others*". He also admitted that this comparison of moods allowed him to understand how other colleagues were feeling and he could act accordingly, e.g. by interrupting a colleague and talking about the problems when something was not going well.

Reflecting about the introduction phase at Regola, he mentioned that it would have been very useful to set some meetings during the evaluation, but for time reasons he was not able to do it. From the managers they also did not receive much feedback for two reasons. First, "*Managers effectively don't have time to reflect about the data and so they did not give feedback to the staff for some reason*" and secondly *"managers maybe looked at the data of the mood of the people, but did not decided to go and do something, but just follow their procedures,..".* This also happened, because the managers might have perceived this again as an additional working task, when reflecting and reacting on the data captured within the MoodMap App.

The feedback from the staff that participated in the interviews was also mostly positive. They described the experience positively, for example "*Yes it is an experience positive … I think it was a useful tool to access this state of mind at any given time of the day and particularly during work and activity*" or "*I see that the MoodMap App was very interesting used during a meeting or moments where I and the team begin to talk about work or begin a meeting to plan*

other projects" and "I thought that the usage of the MMA could help and support the entire company and team work". One participant mentioned that *"The MMA for me was a sort of memo for negative situations during the day".* Also the wish to integrate the application in their internal system was mentioned in order to not forget to state the mood during the day. A major concern regarding reflective learning came also up during the interview "*I think it was very simple to look at others moods but I did not see the real concrete benefit on how to trigger my individual information, to reflect on how to improve my work …But I think that the initial step is capturing as much information as possible, and contextualizing is important but I could not see the trigger to do the next step".* One of the participants stated that "*we were happy to use that, because we know the potential of the app, it could be interesting if you talk directly to our manager - he was also very exciting of testing such a tool - we are very positive about the MMA".* They also mentioned that their manager realized the importance and that he understood the benefit of such a tool, but the MoodMap App was not integrated in the daily work routine and therefore he missed to use the MoodMap App and missed to get an overview of the team an did not make next considerations.

The interviewees stated that they captured their mood regularly in the morning, during or after lunch and before they left the office, on average two to three times a day. Additionally they mentioned that they captured their mood in stressful situations or when they were angry "*I was very angry because I had a trouble with a customer or my colleague, I don't remember well the situation, and reported it into the MMA.*"

Regarding the different available views, all participants mostly liked the comparison bars because they were very intuitive and that "*is important to understand the motivation of my colleagues and their engagement".* Only one of the four interviewees mentioned to not have used the pop-ups (reflection amplifiers, reflection interventions) at all. The others perceived them sometimes as useful and sometimes as annoying. "*Sometimes I feel the pop-up was very useful to remember me to write something about my mood. In other case I felt it was very annoying" or "I used the pop-ups to annotate and give context, because my natural is to reflect about my work. I used this tool and I find it useful but just to formalize a way to work, in my usual way of working".* Only one of the four staff interviewees mentioned having reviewed the reports regularly.

Some of the participants also perceived a clear benefit or insight for themselves e.g. "*to express my feelings and in general and in the job/work /working day - it is possible that the MoodMap App helped me in this regard for activities with my colleagues or my team activities in general."* or "*I am a person who is very angry after a meeting. I used to go happy to the meeting and feel angry afterwards. The MMA allowed me to have a look - how can I switch my mood directly after the meeting - an app that can help me to improve - before meeting mood high - after meeting mood low".* On the other hand there was also a lack mentioned of how to get something relevant out of the captured data "*I collected a lot of data information and my trend. I had learned something more on my approach to the work. But then… or maybe I did not realise it or I am still missing of something that triggers and makes me shine on new perspectives."*

One wish which was also mentioned twice was to make the application available on a smartphone. Many employees of Regola are often out of offices and it would therefore be great to have it available on the mobile device. One participant mentioned that "*We use in Regola the MMA in working situation and I see that my colleagues use the MMA only to store the bad feeling of the day - there is something worth thinking about*." Another opinion was that "*I would spend some work in trying to make it… to trigger… to make more tangible, more evident how*

to reflect". Although the two statements above need some further considerations, we also received very positive statements "*I think it is a very welcome project very positive project for managers and also the staff.*" And "*I enjoyed it. I feel happy to have used the application".*

References

Bachmann, T., Jansen, A., & Mäthner, E. (2004). Check-the-Coach: Fragebogen zur Evaluation von Coaching, Goldenstedt, Christopher Rauen, 2004.

Boland, R. J. & Tenkasi, R. V. (1995). Perspective making and perspective taking in communities of knowing. Organization Science, 6(4), 350–372.

Hansen, K. (2001). Zeit- und Selbstmanagement, 1. Aufl., Berlin, Cornelsen.

Hatton N. & Smith D. (1995). Reflection in teacher education: Towards definition and implementation. Teaching and Teacher Education, 11, 33-49.

Krogstie B., Prilla M., & Pammer V. (2013). Understanding and Supporting Reflective Learning Processes in the Workplace: The RL@Work Model. Proceedings of the Eights European Conference on Technology Enhanced Learning (EC-TEL 2013).

Prilla M, Herrmann T, Degeling M (2013) Collaborative Reflection for Learning at the Healthcare Workplace. CSCL@Work: Case Studies of Collaborative Learning at Work

Prilla M, Pammer V, Balzert S (2012). The Push and Pull of Reflection in WorkplaceLearning: Designing to Support Transitions Between Individual, Collaborative and Organisational Learning. Proc. Seventh Eur. Conf. Technol. Enhanc. Learn. pp 278–291

Prilla M, Pammer V, Krogstie B (2013). Fostering Collaborative Redesign of WorkPractice: Challenges for Tools Supporting Reflection at Work. Proc. Eur. Conf. Comput. Support. Coop. Work ECSCW 2013

Prilla M, Renner B (2014) Supporting Collaborative Reflection at Work: A Comparative Case Analysis. Proceedings of ACM Conference on Group Work (GROUP 2014)

Seiwert, L. J. (2002). Das 1x1 des Zeitmanagements, München, Verlag Gräfe & Unzer.

MIRROR Deliverables

D1.4b Model of Computer Supported Reflective Learning –version 1 updated

D1.5 Specification of Evaluation Methodology and Research Tooling

D1.6 Model of Computer Supported Reflective Learning - version 2

D1.7 Report on Summative Evaluation

D4.2 Individual reflection Apps - version 1

D4.3 Individual reflection Apps - version 2

D4.4 Individual reflection Apps - version 3

D6.1 Design studies and Specifications

D6.2 Prototypes of Annotation and Scaffolding - version 1

D6.3 Enhanced prototypes of Annotation and Scaffolding (version 2); prototype for synergizing - version 1

D6.4 Prototypes of annotation and scaffolding in version 3, prototypes for synergizing in version 2

D7.3 Gaming apps - version 3

D8.2 Prototype for organisational learning intelligence - version 1

D8.3 Prototype for organisational learning intelligence - version 2

D9.5 Report on Integration testing, User Experience and privacy evaluations

D10.1 Initialization and preparation of test beds for use and evaluation of MIRROR tools and methods

D10.2 Formative evaluation of MIRROR Appsphere usage and effectiveness at test beds