Collaborative Project

Integrated Project
Reflective Learning at Work

European Commission Seventh Framework Project (IST-257617)

Deliverable ***1.7***

Report on Summative Evaluations

Editor	*Bettina Renner, Gudrun Wesiak*
Work Package	*1*
Dissemination Level	*Public*
Status	*Final*
Date	*June 30, 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	25/05/2014	Gudrun Wesiak	First version of document (structure)
v0.2	10/06/2014	Gudrun Wesiak, Bettina Renner	Draft version for Internal Review
v1.0	30/06/2014	Gudrun Wesiak, Bettina Renner	Final version for Submission to the EC

Contributors

Name	Institution
Bettina Renner	KMRC
Gudrun Wesiak	KNOW
Roy Ackema	TQ
Michele Biole	REG
Marina Bratić	KNOW
Dominik Cavael	NBN
Monica Divitini	NTNU
Samia Drissi	INFOM
Nils Faltin	IMC
Angela Fessl	KNOW
Thomas Kleinert	DFKI
Ellen Leenarts	BT
Neil Maiden	CITY
Simone Mora	NTNU
Dalia Morosini	IMA
Lars Müller	FZI
Viktoria Pammer	KNOW
Verónica Rivera-Pelayo	FZI
Michael Prilla	RUB

Kevin Pudney	RNHA
Lisa Reinmann	NBN
Simon Schwantzer	IMC

Reviewer

Name	Institution
Roy Ackema	TQ
Monica Divitini	NTNU
Nils Faltin	IMC
Kevin Pudney	RNHA
Konstantinos Zachos	CITY

Executive Summary

This deliverable aims at summarizing the main results from the MIRROR project from two perspectives. Part 1, as the outcome of task 1.4, summarizes and integrates the results from summative evaluations conducted in WP10 and reported in D10.3. It is thus based on a variety of short and longer-term studies that took place in MIRROR's different testbeds to evaluate the effectiveness of MIRRORs solutions for reflection based learning. Part 2 of this deliverable summarizes the lessons learned from four years of research on reflective learning at work. It integrated the perspectives of the whole MIRROR consortium, i.e. scientific partners, app-developers, and application partners.

Part 1, representing the user side of the deliverable, is structured as follows: After the introduction (Chapter 1) it reports results from evaluation studies in two ways. First, the data of 20 different evaluations, covering 12 apps tested in 9 different testbeds is analysed on an aggregated level along the four levels of Kirkpatrick's evaluation model (Section 2.2), i.e. reaction, learning, behaviour, and results on an organisational level. Second, we analysed in detail how the MIRROR apps support the four stages of the Computer Supported Reflective Learning (CSRL) model developed in WP1 (Section 2.3). The underlying concept for Part 1 analyses is a triangulation of methods, i.e. we combined quantitative and qualitative data from log-files, questionnaires, interviews, focus groups, and organisation specific KPIs (Key Performance Indicators) of the evaluations reported in D10.3.

Main results:

- Participants were satisfied with the MIRROR apps and indicated that they are useful for professional competence development. Main barriers for using the apps were a lack of time and space for reflection, but generally participants tended to be in favour of continuing their app usage.
- Participants showed a high initial level of reflection, especially for individual reflection, perceived the app-specific support of reflective learning provided by the MIRROR apps as positive and also reported positive learning outcomes.
- With respect to the effect MIRROR apps had on the working behaviour of participants, the results indicate that users perceived some improvements in their behaviour at work, as well as increased satisfaction or confidence with the working tasks.
- On the results level we found only marginal changes and these should be interpreted carefully as changes on this level can be influenced by a variety of factors.
- Regarding the CSRL model the evaluated MIRROR apps successfully support reflection at the first three stages of the reflection process and also the transitions between these stages as well as the transition to the apply outcome stage are covered by the tools. Further work should elaborate more on the apply outcome stage.
- Considering the different settings of the conducted studies, we found effects of organisational sector (business vs. emergency vs. healthcare), reflection context (training vs. work), and work experience (low vs. high) on variables at all evaluation levels.

Part 2, representing the project side of the deliverable (Chapter 3), summarizes the insights gained throughout as lessons learned according to the following aspects:

- **Potential** for reflective learning: What potential for reflection at work was identified during the (first years of the) project?
- **Forms** of reflection: What insights were gained w.r.t. different forms of reflection? These include different levels of reflection, formal vs. informal settings, real-life vs. virtual experiences, and reflection as campaign vs. continuous process.
- **How to successfully introduce reflective learning?** Several crucial factors have been learned along the following categories: Technical aspects, management support, testbed characteristics, introduction of reflective learning & apps, data capturing, and the long term process of reflection.
- **Effects** of apps in different testbeds: What effect of the MIRROR apps could be found in the different testbeds?
- **Evaluation** aspects: What experiences did we gain on the methodological side in evaluating MIRROR apps?

To conclude this deliverable, **Chapter 4 integrates insights from Parts 1 and 2, i.e. the user side with the project side**. Overall, the conjoint analysis of the data collected in a variety of individual evaluations and the integration of experiences gained from all partners (scientific, app developer, application) helped to derive valuable insights on a more aggregated level beyond the findings of the individual evaluations. It showed that the introduction of technology support for reflective learning at work is able (a) to trigger new reflection processes on individual, team, and organisational level, (b) to improve employees' working behavior w.r.t. different aspects, and (c) to foster the reflection process starting from capturing data to documenting reflection outcomes. Future work should concentrate on fostering the adoption of technology support by providing a socio-technical framework with a holistic approach to reflective learning at work. This might also include the combination of several apps to cover the entire reflective learning cycle.

Table of Contents

Table of Figures

List of tables

1 Introduction

This deliverable summarizes and integrates the results from the summative evaluations conducted in WP10. WP1 aims at providing a scientifically and methodologically common ground for the project partners. Together with D1.5 the D1.7 is the outcome of task T1.4. The first part of this task was to develop a methodology for evaluating the impact and effectiveness of MIRROR apps for reflection based learning in longer-term studies in MIRROR's different testbeds (D1.5), while the second part of summarizing and integrating the outcomes of the evaluations is addressed in this deliverable. During the four years of MIRROR, more than 20 apps have been developed, which all aim at supporting reflective learning at work. Due to the common methodology, which all evaluation studies followed, it was now possible to assess the learning effectiveness MIRROR apps had in the wild and the impact of the developed reflection methods. However, the apps address the common topic of work-integrated reflection on different levels regarding the scope of reflection within an organisation (individual, collaborative, vs. organisational reflective learning) as well as on different phases of the reflection process. The development process started in year 1 (Y1) with user studies to specify the AS IS status of reflection in the MIRROR testbeds as well as the needs of the potential users of MIRROR apps. This first phase was followed by 2 years of app development and continuous improvements based on the results obtained via formative evaluations as well as insights gained by accompanying conceptual work on reflective learning at the work-place. As an outcome of this work, which has been reported in detail in Y1, Y2, and Y3 deliverables, 15 apps have reached a status which allowed an integration into the work (or training) process of employees. The evaluation of those apps' (in one or more testbeds each) followed the criteria of a summative evaluation (as defined in D10.1, D10.2), which investigates the impact of the apps and their underlying approach to reflective learning not only regarding users' reaction to and learning from the apps, but also their effects on employees' actual working behavior and on organisational factors in terms of KPIs (Key Performance Indicators). The methodology for the summative evaluations was developed in Y2 and is described in D1.5. The results of these separate evaluations are reported in the form of individual evaluation reports in D10.3.

However, in order to get more general insights regarding the overall MIRROR approach to reflective learning and thus the project impact as a whole, the data obtained in the single studies was aggregated into one common data set and analyzed with respect to MIRROR's overarching research questions. This was only possible because of the common evaluation framework of D1.5 which was used in all evaluations. Additionally, the analyses took into consideration that the MIRROR testbeds are representing different organisational sectors (business, health, emergency), that the participating units tested the apps in different contexts (work vs. training), and that the evaluations differed regarding their duration (short- vs. long-term). The results of these overall analyses are reported in the first part of this deliverable, which represents the user side or user perspective of the MIRROR approach.

The second main part of the deliverable is concerned with the project side, i.e. the experiences gained by the MIRROR partners during their four years of researching technology-supported reflective learning. The insights gained from the MIRROR partners consider the perspective of scientific partners, app developers, as well as application/testbed partners. This way it was possible to gain comprehensive insights on reflective learning at the workplace and to examine the topic exhaustively w.r.t. the potential of reflective learning, different forms of reflection, technical, organisational, and context-related aspects for the successful introduction of reflective learning, effects of the apps in organisations, and also methodological aspects to be considered in such a variety of evaluation studies.

1.1 Structure of this document

This deliverable summarizes the overall project results from two perspectives. On the user side the results gained from summative evaluations in Y4 of the project are integrated and on the project side the insights of the project partners are brought together. Thus, this report is split into two main parts: Results from overall analysis of Y4 summative evaluations (Chapter 2: user side) and lessons learned by the project partners (Chapter 3: project side).

After a short review of the objectives and research questions underlying this work, Chapter 2 starts out with a short description of the used methodology and the data set used for the overall analyses (Section 2.1). The obtained results are presented in two ways, namely along the evaluation model by Kirkpatrick (Section 2.2) and along the CSRL model (Section 2.3). The latter analysis gives insights to which extent the apps actually cover the four stages of the model and how this support of reflective learning at different phases of the reflection process is perceived by the users. Chapter 3 summarizes the lessons learned by the project partners and structures them along the most relevant topics found in the reports provided by each partner (Sections 3.1 – 3.5). Both Chapters 2 and 3 are concluded by short discussions, whereas an overall conclusion is provided in Chapter 4. There, we try to integrate the user and project perspective by connecting the results from data analysis with the insights provided by the scientific and application partners.

1.2 Objectives: Integration and reporting of results

This deliverable is the outcome of task 1.4 – it summarizes and integrates the results from the summative evaluations conducted in WP10 and reported in detail in D10.3. The aim was to evaluate the effectiveness of MIRROR solutions for reflective learning in longer-term studies at different MIRROR testbeds.

The integration of results is possible due to the summative evaluation framework reported in D1.5, which was the basis for all summative evaluations conducted by WP 3-8. Thus, in spite of different user groups and different applications used, the results can be compared via the common toolbox, which provides a set of core questions used in almost all evaluations as well as a set of additional questions selected by the evaluation partners according to their needs.

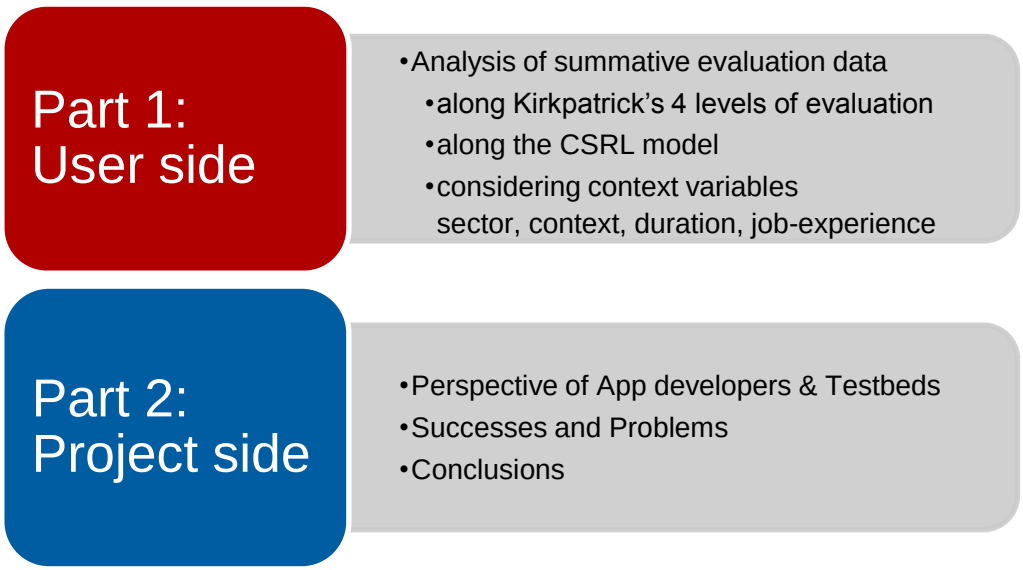

Figure 1. Overview of data analysis dimensions (user side) and structure of project insights (project side)

Additionally, the insights gained by the project partners throughout the four years of MIRROR have been collected and integrated in order to obtain some overall view of the goals reached, the perspectives changed, and the issues still open after four years of research. In order to gain a more holistic picture of the project's current status in terms of what has been learned, how reflective learning at the workplace could be advanced, and what is still open for future research, the insights include the app developers' as well as testbed partners' points of view. Figure 1 gives an overview of the content provided in the two parts of this report.

2 Overall Analysis of Summative Evaluations: Part 1 - User Perspective on Reflective Learning

In year 4 of the project, partners carried out final summative evaluations of the apps developed in MIRROR. The detailed analysis of these results provides a sound basis for best practices and for transferring the concept and systems to other domains. To reach this goal, plans for all five testbed scenarios were created and each testbed partner used MIRROR apps for a period of several weeks up to two months. During this period, the measurement instruments that were developed in WP1 (see D1.1 and D1.5 for detailed descriptions of MIRROR's research methodology and tooling) have been used to gather data on the processes and outcomes related to working with MIRROR apps. Whenever possible, the effects MIRROR has on the users and their environment have been assessed by comparing the situations before and after MIRROR apps were introduced to the workplace. This way, we could identify the strengths and weaknesses of the MIRROR approach across the experiences with different apps and different testbeds.

The obtained results will be presented from two perspectives: First, data are aggregated according to the four levels of evaluation by Kirkpatrick (reaction, learning, behaviour, organisation/results), which builds the basic framework for summative evaluations in Mirror. Second, data are presented along MIRROR's CSRL model to show how and to which degree the single stages are supported by the developed apps.

2.1 Methods

The overall analysis is based on the raw data obtained in the summative evaluations carried out during year 4. In order to meet the challenge of presenting an integrative summary with general results and to avoid conclusions that are based on experiences with a single app at a specific testbed, the following criteria have been applied for the selection of data and measurement units (variables):

- All data have to stem from summative evaluation settings, i.e. the evaluation study had to cover all four levels of Kirkpatrick's evaluation model and the sample had to represent the target group specified for using the respective application.
- Analysis results are only used as indicators for a specific measure, if the data stems from evaluations of at least three different apps (multiple evaluations of the same app in different settings are counted only once).
- Raw data from individual users are only included if the participant filled out (parts of) the post-questionnaires.

The variables used as indicators for measuring the effects of MIRROR apps on the different levels of evaluation (reaction, learning, …) are for the main part elements of the MIRROR toolbox described in D1.5. This includes objective usage data (logs-data captured by the apps), questionnaire data based on core questions and additional questions from the toolbox, and objective as well as subjective KPI measures. For the latter, the above mentioned criteria do not apply, since they are individual measures specified for each group of users, separately. Core questions have been intended to be used in all summative evaluations, whereas the additional questions have been selected according to the needs of the respective study setups. Except for questions with an open answer format, most questionnaire data is based on items with 5-point Likert scales as response format (with 1 – strongly disagree, 2 – disagree, 3 – neutral, 4 – agree, and 5 – strongly agree). Exceptions

from this scale are mentioned together with the respective results. Additionally, the coding scheme developed by WP1 and WP6 (see Appendix 5.3) is used to report qualitative data gathered from the content of reflection notes that have been entered by the participant of several evaluations.

For a better readability, the single variables and questionnaire items are presented together with the respective results in the subsequent subsections.

2.1.1 Classification of evaluation studies

The summative evaluations this deliverable is based on can be classified according to (a) their duration, (b) the context of evaluation, and (c) the organisational background of the testbeds (job descriptions of users). Regarding the duration, we differentiate between short-term (single use up to several days) and long-term evaluations (lasting between 2 and 11 weeks) and the context of the evaluation is either work-integrated or in a training setting. However, it has to be noted, that with training we do not refer to a traditional class room setting. In the context of the Serious Games the training consists of the simulation of work situations, the Medical Quiz aims at linking work relevant knowledge to real work experiences and the WATCHit evaluation tried to recreate realistic work situations which is also a kind of simulation. The organisational background of the testbeds varies between the business sector (Infoman, BT, or IMC), the health sector (RNHA, NBN), and the emergency sector (Regola). Table 1 gives an overview of the evaluations and their respective settings.

Table 1. Classification of summative evaluations for overall data analyses

App	Testbed	Organisational Sector	Duration	Context
KnowSelf	Infoman	business	long-term	work
KnowSelf, ARA	IMC	business	long-term	work
MoodMap App	BT, Regola[a]	business	long-term	work
TalkReflect	RNHA, NBN, RBKC[b]	healthcare, business	long-term	work
DoWeKnow	Infoman	business	long-term	work
IAA, IMA	BT, NBN	business, healthcare	long-term	work
MedicalQuiz	NBN	healthcare	long-term	training
CaReflect	RNHA	healthcare	short-term	work
WATCHiT	Regola (Cuneo[c])	emergency	short-term	training
The Virtual Tutor Serious Games	RNHA, external	healthcare	short-term	training
Rescue League Serious Game	Regola, external	emergency	short-term	training

Note. [a] users were employees at Regola's administration, not emergency volunteers (as in the WATCHiT and Serious Games evaluations; [b] RBKC: London Royal Boroughs of Kensington and Chelsea; [c] Cuneo: big emergency training event in Italy;

2.1.2 Data sources

The overall analysis of summative evaluations considers data from 20 different evaluations, covering 12 apps tested in 9 different testbeds. The testbeds include the five application partners, one scientific partner (IMC) as well as three external sites that were interested in trying out some of the MIRROR apps.

Table 2 shows the sample sizes by apps and testbeds. Shown are the numbers of participants using the respective apps and filling out the questionnaires provided within the evaluations studies (as well as the number of participants using the apps at least once). In some testbeds the same persons have been involved in testing different apps. However, if not indicated otherwise, these evaluation studies have been conducted independent of each other (concerning set up of evaluations and times of testing) and as this was the case only for a few individual participants we did not consider this aspect furthermore in the overall analysis. We therefore count these participants separately, i.e. as two cases.

Table 2. Sample sizes of summative evaluations per app and testbed.

App/ Testbed	BT	RNHA	NBN	IMC	Info-man	Regola	Uni Berga-mo	RBKC	Emerg Milan
MMA	**39***[1] (58)					**34** (35)			
CaReflect		**40** (44)							
Medical Quiz			**21** (24)						
KnowSelf				**10** (10)	**10** (12)				
ARA				**10** (10)*[2]					
Virtual Tutor SG (CLiniC, CARE)		**5**					**16** (16)		
Rescue League						**19** (19)			**14** (14)
Talk Reflect		**5** (9)	**10***[3] (9)					**7** (12)	
Watchit						**35** (35)			
IAA	**24***[1]		**11**						
DoWeKnow					**10** (10)				

Note. Bold numbers denote participants who used the app and answered the post–questionnaires, numbers in brackets denote all participants who used the app at least once (with or without answering questionnaires). *[1] 3(MMA) and 6 (IAA) participants answered the questionnaire without using the app *[2] Same participants as KnowSelf at IMC, *[3] Evaluation of 2 years, partly same participants.

Altogether 347 participants tested the MIRROR apps, 310 also filled out the questionnaires (ranging between 5 and 40 per app and testbed). As compared to participant data reported for individual evaluations in D10.3, the sample sizes reported in this deliverable are sometimes smaller, because of the criteria we used for the overall analysis (see above). Additionally, data from two formative evaluations at NBN (TalkReflect and IAA) are included in this analysis, which are reported in D10.2, but are still long-term evaluations conducted with the target group and considering the core questions of the toolbox. Data from the WP5 Yammer evaluation could not be included as the evaluation was still ongoing at the time this data analysis was conducted.

Because the different studies focused on different aspects of evaluation, most items have not been included in all of the studies. Furthermore, participation in the studies (using the apps) as well as answering the questionnaires has been performed on a voluntary basis, because of which some participants did not fill out the questionnaires at all while others left out some items. For these reasons sample numbers in the result section vary greatly and are thus reported with each individual result.

2.1.3 Participants

Across all evaluations, log-data and responses from 321 participants were analysed. Demographic data from pre-questionnaires is available for 283 individuals, of which 44% are male, 56% female. Data on participants' age was collected by means of 6 age groups (1: 19 years and younger, 2: 20-29, 3: 30-39, 4: 40-49, 5: 50-59, 6: 60 years and older), resulting in a median age group of 30-39 years. The distribution of age among the participants is depicted in Figure 2. Participants reported an average of 4.8 years (*SD* = 5.1, *N* = 243) to be in their current position (see Figure 2), with 62% working there for less than five years (*M* = 1.8y, *SD* = 1.2, *N* = 150) and 38% 5 years and more (*M* = 9.7y, *SD* = 5.2, *N* = 93). The most frequently reported jobs are those of an advisor (54 persons), carer (47), and volunteer (27).

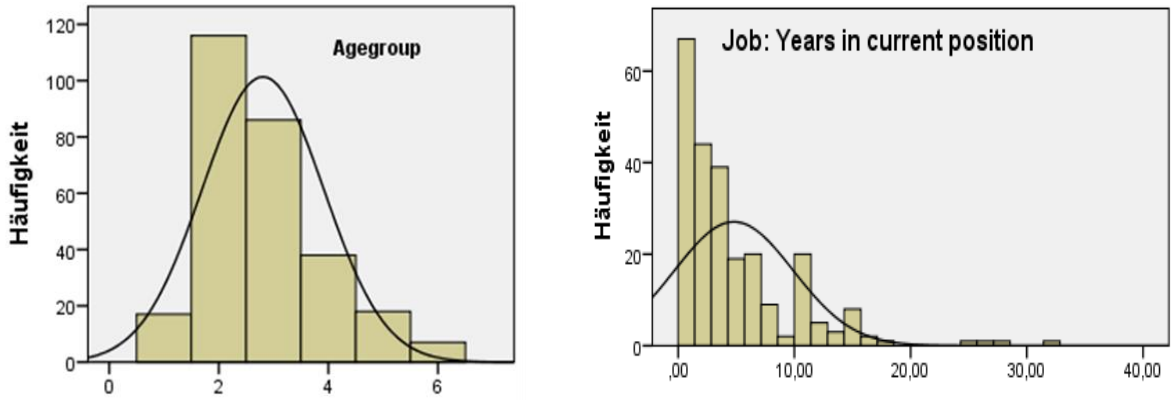

Figure 2. Demographics: distribution of age (<20, 20-29,..., 50-59, >59) and years in current position.

With regard to the Evaluation context, out of the 321 participants, 60% took part in a long-term evaluation (40% short term), 66% used the apps as work-integrated implementation (34% in a training-context), and 44% worked in the business sector, 39% in the health sector, and 17% in the emergency sector. However, it needs to be pointed out that these classification variables are confounded, which needs to be considered when interpreting the results of the following data analysis. More exactly, in the business sector all evaluations have been long-term and work-integrated, whereas the emergency sector evaluations have been conducted short-term in a training context. Only the health sector participated in long- and short-term as well as work and training context evaluations.

2.2 Data Analysis along Kirkpatrick's Evaluation Model

The development of the MIRROR evaluation methodology (as described in detail in D1.5) is based on the one hand on Kirkpatrick's model for summative evaluations (see Figure 9 in D1.5 for an overview of evaluation levels and criteria modified to the context of MIRROR's informal, reflective learning situation), and on the other hand on the *i** model, which is described in Detail in D1.5, Section 2.2). The combination of these two models (Kirkpatrick's levels with *i** criteria) results in a detailed and concrete framework, which specifies the effects expected after app usage for each evaluation level. The proposed model combination,

resulting rationale for summative evaluations, and derived evaluation criteria are given in Table 3 (extended version of Table 2 in D1.5). It describes for each of the modified Kirkpatrick levels the criteria of the *i** model, in order to add detail and concreteness. Additionally the research questions for each level which we wanted to answer are named.

Table 3: Four levels of evaluation, i criteria, rationale, and derived evaluation criteria*

Level	***i* Criteria***	***Rationale & Evaluation Criteria (research questions RQ)***
1 Reaction	General Criteria	The *i** general criteria are concerned with the motivation and opportunity to reflect; in other words, whether participants are motivated to use the app. This inclination is affected by how well participants like the app. RQ: Did participants use the app? Did they like the app?
2 Learning	Process Criteria	The *i** process criteria shed light onto whether and how participants engage in the process necessary for learning: in our project this is the reflection process. RQ: Do people reflect more after app usage? Which processes in the model are supported by the apps
	Outcome Criteria	Outcome criteria of *i** are mainly related to learning and include change in knowledge and behavioural intentions. RQ: Did people learn something by using the apps? What elements of reflection can be found in notes?
3 Behaviour	Work-Related Criteria	Work-related *i** criteria concern concrete action or they are related to behaviour (e.g., self-efficacy, work mastery, and employee satisfaction concern subjective evaluation of performance at work). RQ: Did participants change something in their work behaviour after app usage? Did they improve their work?
4 Results	Business Impact	Evaluation criteria for organisational learning are related to the high-level (business) impact of MIRROR. RQ: Is there evidence for benefits (increased satisfaction, improved work quality) of employees' app usage on an organisational level (KPIs)? Would participants recommend the app?

2.2.1 **Level 1: Reaction**

On the first level of evaluation, objective and subjective app usage have been considered, as well as barriers for usage, intended long-term usage, satisfaction with and usefulness of the app.

Objective Usage was measured in terms of usage time in minutes (item CF2) over the course of the evaluation period. Data from 124 participants using 6 different apps could be extracted from log files and yielded an average usage time of 97.8min (SD = 205.8) per person. Dividing the data according to the evaluations' duration, results in *M* = 115.9min (*SD* = 228.2, *N* = 98) for long-term and *M* = 30min (*SD* = 11.56, *N* = 26) for short-term usage. Data for the latter stems from serious games evaluations for which the playtime ranged from 9 - 40 minutes, usage in long-term evaluation varied between 2 and 1358 minutes.

Subjective Usage was assessed via self-reports in the post-questionnaires of long-term evaluations. Responding to an open answer format (CU1 – How many times have you used [the app]), participants indicated to have used the apps from ‚never' or ‚sometimes' to ‚several times a day for several weeks'. In order to differentiate between responses from

active users and those who just tried out the apps, the sample was divided into two groups for all further analysis on the reaction level. The groups are composed of 41 persons with no or limited use (tried app max. 2 times, but filled out the post-questionnaire) and 125 active users (used app at least 3 times).

Table 4 lists the **Additional Questions** that have been used for the evaluation of at least three different apps and are therefore included in this overall analysis.

Table 4. Items used to evaluate the apps' on Level 1 - reaction

Topic	Item Code	Item text
Usage Barriers	USE01	I did not have the time to use the App
	USE02	I did not have the place to use the App
	USE03	I did see no advantage in using the App
	USE04	I was not motivated to use the App
	USE05	I could find out how the app worked*
	USE06	I need more formal training with the app
Satisfaction, professional training	SAT01	I am satisfied with the App
	SAT02	I think the App is useful for professional training
	SAT03	I think the App can be used to complement professional training
Long-term usage	LT01	I see the long-term advantage of using the app in my work-life
	LT02	I would like to use the App continuously as part of my work-life
	LT03	It is practical for me to continue using the App in my work-life

*Note: * Item was recoded as barrier for analyses*

An analysis of the **barriers** (Figure 3a) that hindered participants from using the app, showed that compared to active users, the inactive users indicated more often, that the lack of time or space (USE01,02) constituted a barrier for them ($t_{(59)}$ = 2.16, p =.035). We found no difference regarding motivation or not seeing an advantage (USE03, 04) and usability (USE05, 06). There was also no significant difference between the three organisational sectors (emergency, business, health), which indicates that the barriers for usage, esp. lack of time or space, are not related to participants' working background. As far as job experience is concerned, participants with less experience (less than 5 years) perceived a lack of time as barrier (M = 3.24, SD = 1.16, N = 32), which was not the case for experiences employees (M = 2.48, SD = 1.18, N = 29). An independent samples t-test proved this difference to be significant ($t_{(59)}$=2.55, p =.013). Job experience did not influence the remaining types of barriers.

Additional qualitative results on barriers that we gained via open questions and interviews also show that time constraints are an important issue. However, participants reported also a lack of motivation in terms of not seeing enough benefit in using apps. Another point were technical constraints, which was especially the case with the Serious Games in the care sector as internet access is limited in the care homes. While most participants did not seem to have a problem with privacy issues some felt uncomfortable with tools tracking their activities on a computer (KnowSelf) or sharing their moods (MoodMap

App). As additional factor a lack of management or lead user support was identified as possible barrier.

As shown in Figure 3b, participants general **satisfaction** with the apps, which was measured by one item (SAT01), mounts up to *M* = 3.53 (*SD* = 0.96, *N* = 176), their perception of how useful the apps could be with regard to **professional training** reaches a value of *M* = 4.07 (*SD* = .80, *N* =123).

One-way ANOVAs with organisational sector as independent factor revealed significant effects on both satisfaction ($F_{(2,166)}$ = 12.86, p <.001) with the app and usefulness for professional training ($F_{(2,122)}$ = 6.64, p =.002). Post-hoc tests yield significant differences between the emergency sector and the two other sectors (satisfaction with app emergency vs. business: p < .001, emergency vs. health: p < .01; prof. training emergency vs. business p < .01, emergency vs. health p < .05), whereas business and health do not differ. With regard to job experience the perceived usefulness of the app for professional training was rated significantly higher by participants with more experience (for up to/at least 5 years of experience *M* =4/4.44, *SD* =.66/.51, *N* =15/18, $t_{(31)}$ = -2.19, p = .036).

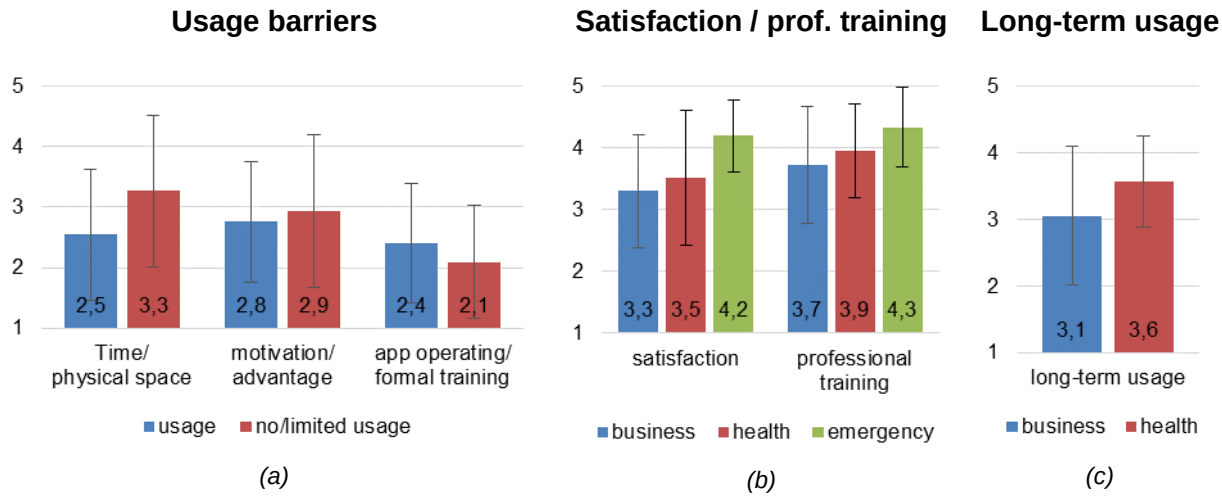

Figure 3. Differences between active/inactive users and organisational sectors regarding indicators for reaction

Finally, participants were asked about their attitudes towards using the apps as long-term applications (LT01-03). Whereas users from the health sector saw the **long-term usage** positive (M=3.57, SD=0.69, N=150), ratings from the business sector are neutral, but significantly lower (M=3.06, SD=1.04, N=102; *t(136.18)=3.61*, *p*<.001), see Figure 3c. There is no data available from the emergency sector (as they were all short-term evaluations) and we found no differences w.r.t. job experience.

Ratings from active vs. inactive users do not differ regarding satisfaction, usefulness for professional training, or long-term usage.

Summarized, the overall reaction to the MIRROR apps is that participants are satisfied with the apps, see the long-term usage positively, and agree that the apps could be useful for professional training. The latter is especially true for participants with longer job experience. Barriers for usage are mainly a lack of time (esp. for less experienced employees) or no adequate physical space, in the interviews some also reported to see no benefit in the apps. With regard to the organisational background, participants from the emergency sector reacted especially positively, whereas we found no difference between business and health.

2.2.2 Level 2: Learning

The second level of analysis is split in two aspects, namely the learning process and learning outcomes. For the former, the central question is whether and how participants engage in the reflection process in order to learn, the latter aspect concerns eventual changes in knowledge and behavioral intentions.

2.2.2.1 Learning Process

To measure the general tendency of participants to reflect and eventual changes in this tendency after app usage, the **Short-Reflection-Scale** (SRS; CR1-CR10, see appendix 5.1.1) was included in the long-term studies' pre- and post-questionnaires. For short-term evaluations, the SRS was presented only once, since a general tendency to reflect is not expected to change over a short period of time. Half of the 10 SRS items refer to individual reflection, the other half to team reflection and can therefore be interpreted as two subscales of the SRS. Furthermore, the MIRROR toolbox provides a set of 43 different **app-specific reflection questions** (CA1-CA43, see appendix 5.1.2), which were developed in order to capture reflection support by the apps on a more differentiated level. More exactly, the items are grouped along the CSRL model and thus represent the different stages of the reflection process. For each evaluation study, a set of appropriate questions was selected and presented to the participants. For the analysis at hand, we calculated for each user the mean score derived from all selected CA items of the respective evaluation. This allows for a comparison of app-specific reflection support across studies, whereas single items are considered in relation to the CSRL model in Section 2.3. In addition a control questions was included in most studies, which had the purpose to check the reliability of participants' answers. The selected control question was always part of the overall item set, but referred to a function which the app under question did obviously not support (e.g. asking whether the app supported sharing of experiences when the respective app did not provide such a sharing function).

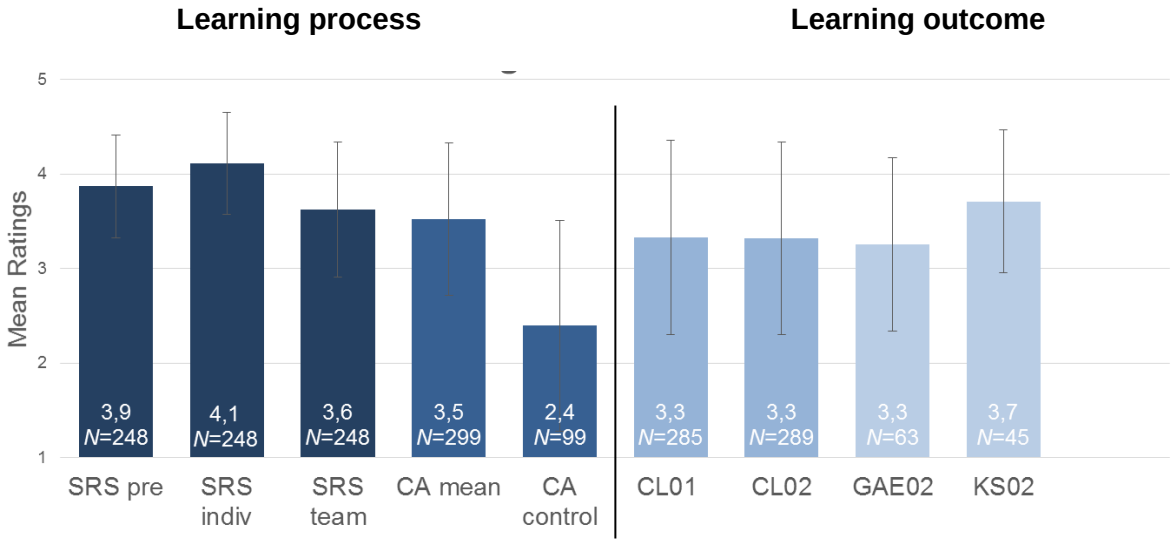

Figure 4. Mean ratings (on 5-pt. scales) for learning process and learning outcome measures.

Note. SRS: Short-Reflection-Scale; CA: app-specific reflection; CL: learning outcome; GAE: general App effects; KS:knowledge/skills;

Figure 4 (left hand) shows the mean ratings obtained from measures assessing the process of reflection and perceived support from the apps. The scores from the SRS, which are for the most part based on pre-questionnaires (except for short-term evaluations with post-

questionnaires only) show that reflection, especially on an individual level, is an important part of participants work or training already (3.6 ≤ *M* ≤ 4.1). The functions of the apps are also perceived positively regarding their perceived ability to provide reflection support. The mean rating is based on a set of 32 different items. The obtained average rating of *M* = 3.5 clearly differs from the control question with M = 2.4, which we view as evidence that participants' responses are valid indicators of their opinions.

With regard to the general tendency to reflect, the SRS was applied pre- and post app usage in order to investigate **changes in the reflection process**. Figure 5 shows the mean ratings for the subscales individual and collaborative reflection obtained in the pre- and post-questionnaires. A two-way ANOVA with the factors testing-time (pre- vs. post app usage) and reflection type (individual vs. collaborative), revealed a significant effect (with a very large effect size) for reflection type ($F_{(1,150)}$ =142.5, p < .001, η^2 = .49) but not for the factor time ($F_{(1,150)}$ = 2.27, p = .134, η^2 = .015). Thus, participants had more experience with individual reflection than with collaborative reflection as a team – however the tendency to reflect did not change during the time MIRROR apps have been used. As there is no significant interaction between the two factors, the latter is true for both types of reflection.

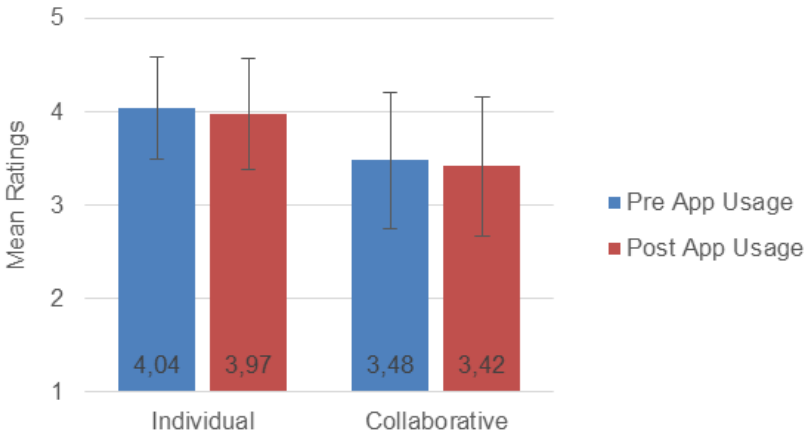

Figure 5. Tendency to reflect as individual or as a team before and after app usage (N=151)

We also investigated whether the general tendency to reflect and eventual changes in this tendency differed regarding **(a) the organisational sectors, (b) the evaluation contexts, and (c) job experience.** Two-way ANOVAs with testing-time as first factor and sector, context, and experience as second factors were performed. Results revealed main effects of sector ($F_{(1,149)}$ = 4.33, p = .039, η^2 = .028) and context ($F_{(1,149)}$ = 7.51, p = .007, η^2 = .048) as well as significant interactions of both variables with testing-time (both p ≤ .002, η^2 ≥ .061). Job experience did not affect the SRS score.

Figure 6 shows that before using the apps all participants rated their tendency to reflect about equally high. After using the app ratings from the work context and the business sector stayed about the same, whereas ratings from participants in a training context as well as those working in the health sector (which are actually in part the same) decreased significantly. We assume that this unexpected result can be attributed to the fact, that during the evaluation period - with all the workshops, explanations about reflection, and prompts to actually reflect on one's behavior – participants changed their understanding of what reflection actually means and thus revised their self-estimation of how much they regularly reflect during their work (and which parts are simply conversations about work).

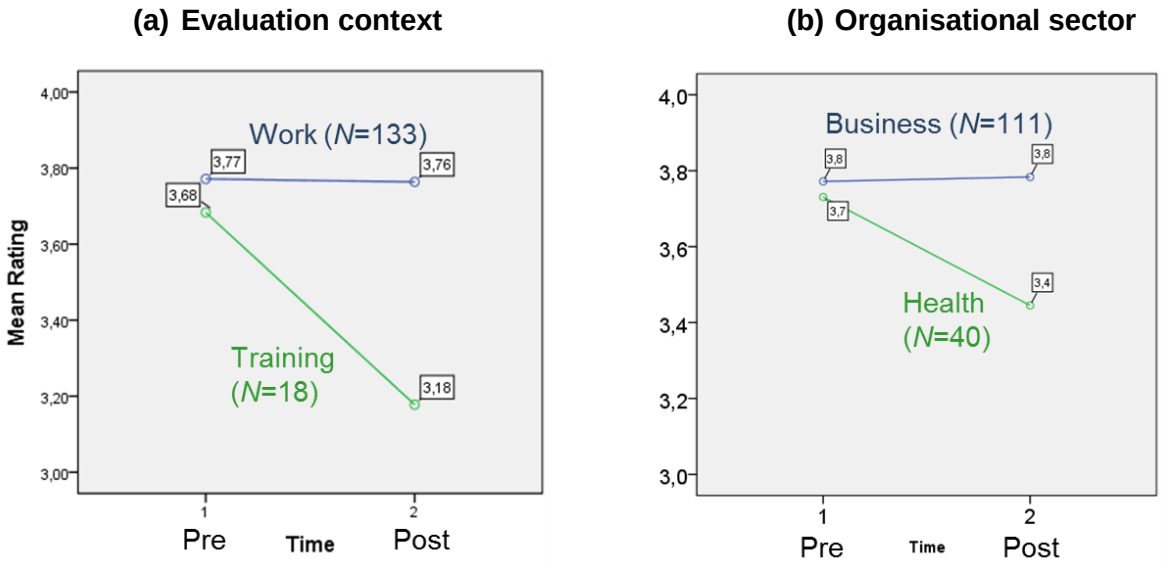

Figure 6. Tendency to reflect before and after app-usage as function of (a) evaluation context and (b) organisational sector

In order to also include short-term evaluations in the analyses business sectors, we also performed a one-way ANOVA and Kruskal-Wallis tests (in the case of non-homogenous variances or when the data was not normally distributed) with organisational sector as main factor and overall, individual and collaborative reflection as well as mean app-specific reflection ratings as dependent measures. Mean ratings are depicted in Figure 7. The results indicate that organisational sector has an effect on all variables. The effect on the overall SRS score (similar results were found for the two subscales) amounts to $X^2 = 23.03$, $df = 2$, $p < .001$ (($N_{business}$=113, N_{health}=81, $N_{emergency}$=54), post-hoc U-tests reveal differences between emergency and the other two sectors (both $p < .001$), but not between health and business. For the app-specific reflection support $X^2 = 55.12$, $df = 2$, $p < .001$ ($N_{business}$=131, N_{health}=114, $N_{emergency}$=54). In this case all three sectors differ from each other, with increasing rating from business over health to emergency. Post-hoc U-tests yield significant differences among all three sectors (all $p < .001$). For the app-specific reflection questions, the same analysis could be done for the evaluation context, but since all emergency evaluations have been conducted in a training context, data are highly confounded and thus results show the same pattern (higher ratings in the training than the work sector). Job-experience does not have an influence on these ratings.

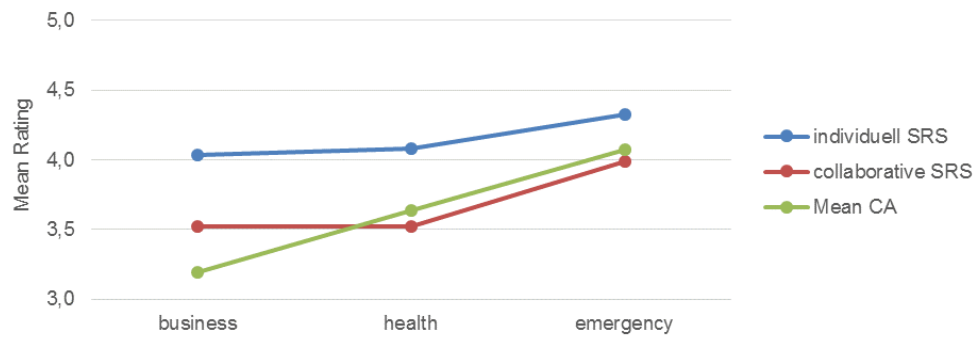

Figure 7. Effect of organisational sector on SRS-scores and ratings for mean app-specific reflection support

2.2.2.2 Learning Outcome

For measuring the perceived learning outcome, the MIRROR toolbox provides **two core questions**, which have been used in the evaluations of 10 different apps, and a large set of additional questions concerning general app effects, knowledge and skills, behavioral intentions, and the re-evaluation of experiences. Out of these **additional questions**, only two have been used in at least 3 studies with different apps and will therefore be analyzed within this context. However, an additional source for gathering data about learning outcomes are the notes participants left in the apps, e.g. when capturing their experiences at work (with colleagues, customers, or patients) or the outcomes of reflection sessions. These notes have been analyzed with respect to their reflective content by means of a coding scheme, the outcome of which is summarized below.

Table 5 lists the items that have been used for the evaluation of learning outcome in connection with at least three different apps and Figure 8 shows the mean ratings obtained across the evaluations. **Job experience** did not have an influence on the ratings of any of these four items.

Table 5. Items used to evaluate the apps' on Level 2 – learning outcome

Topic	**Item Code**	**Item text**
Core Questions	CL01	I made a conscious decision about how to behave in the future.
	CL02	I gained a deeper understanding of my work life.
General App Effects	GAE02	[The app] helped me to find situations on which we should reflect.
Knowledge and Skills	KS02	I improved my understanding in the area that I wanted to improve in.

With regard to the evaluation context, both core questions (CL01, CL02) as well as the knowledge/skill question (KS02) have been rated significantly higher by participants of the training settings than those using the apps at work. U-test for CL01/CL02/KS02 yield values of U = 2.36/2.43/3.01 with all $p \leq .018$. Context did not affect ratings for item GAE02 (p = .501). Kruskall-Wallis tests indicated effects of sector for both core questions with X^2 = 26.98/31.6, df = 2, p <.001 for CL01/CL02. Post-hoc U-test show significant differences among all three sectors for CL02 and between emergency and the other 2 sectors for CL1 (all $p \leq .003$). Ratings from the business and health sector do not differ significantly for CL01, GAE02, or KS02.

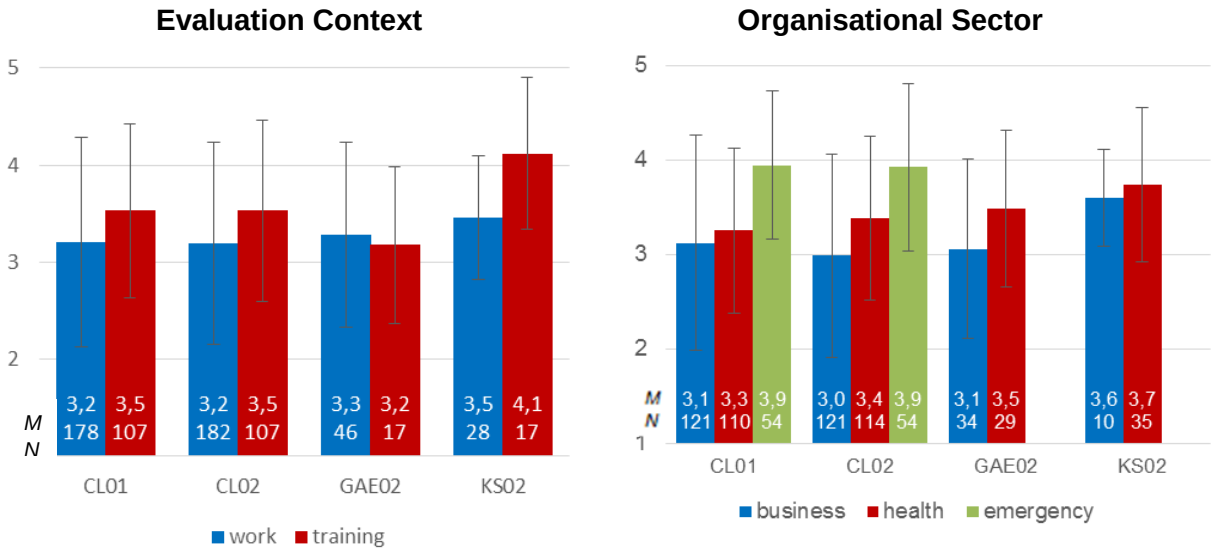

Figure 8. Effects of organisational sector and context on learning outcome variables

In order to get a clearer picture of how the different measures form the learning level relate to each other, we looked at possible **correlations** between the ratings for these items. In addition we checked whether participants with longer usage times would also have the feeling to benefit more from the apps in terms of learning. Table 6 shows the results from Pearson correlations. The data indicate that participants who used the apps more (usage) also perceived the app-specific functions as more supportive for reflection (CA mean) and gained a deeper understanding of their work life (CL02). Furthermore, there are high positive inter-correlations between the general tendency to reflect (SRS), app-specific reflection support (CA), learning outcome (CL01,02). Furthermore, the more participants perceived the app to be helpful in finding situations on which to reflect is related, the higher their subjective learning outcome and mean CA score. On the other hand, the knowledge/skills item KS02 is not related to any of the other included variables.

Qualitative data on reflective learning outcomes was gathered in interviews with participants and managers as well as in focus groups about reflection outcomes and insights gained while using the apps. Of course these insights are quite specific to the respective reflection topic, the app and the application partner. Nevertheless some examples should give the reader an impression of concrete reflection outcomes, which cannot be captured by means of rating scales. Reflection on time management supported by the KnowSelf App revealed e.g., specific time wasters and helped employees to derive lessons learned such as to daily set priorities, a more efficient way to handle interruptions and distractions. Working with the MoodMap App on the other hand showed users how emotions can influence their work and are influenced by their work. Some users reported strategies how they now handle a bad experience differently in order to improve their mood again.

Table 6. Correlations between objective usage (in min) and learning variables

	SRS	CA mean	CL01	CL02	GAE02	KS02
usage	-,016	,217*	,123	,200*	,251	,087
	,867	,017	,182	,028	,076	,685
	111	120	120	120	51	24
SRS	1	,334**	,173**	,242**	-,076	-,124
		,000	,008	,000	,576	,530
		234	233	233	57	28
CA_overall		1	,543**	,646**	,584**	,096
			,000	,000	,000	,556
			284	288	60	40
CL01			1	,662**	,542**	,073
				,000	,000	,659
				285	60	39
CL02				1	,646**	,156
					,000	,337
					60	40
GAE02					1	,263
						,291
						18

Reflection Notes

Table 7 gives an overview of the Apps and testbeds for which reflection notes were collected and coded according to the coding scheme for reflective elements developed by WP1 and WP6 (see D6.4). The scheme 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 elements are a priori excluded from the analysis. A detailed description of the categories is given in section 5.2. As can be seen in

Table 7, the number of notes entered in KnowSelf and the MoodMap App varied between 103 and 938 (for Talk Reflect only the numbers of notes with reflective content are available). Of these, 57 to 293 notes actually contained some form of reflective element and have thus been analysed. In addition 57 notes captured with Talk Reflect.

The assignment of each note to one or more of the nine categories is based on the judgements of at least 2 independent coders; in the case of disagreement the coders discussed this element and either came to an agreement or discarded the note. The inter-rater agreement ranged between 62% and 97% per category. Assignments to multiple categories are possible due to the different lengths of notes. Some contain long or several sentences, some are conversations between participants.

Table 7. Number of notes and proportion of reflective content for 6 evaluations of 3 Apps

App	**MMA** (individual notes)		**KnowSelf** (individual notes)	**Talk Reflect** (conversations)		
Test beds	BT	RG	Infoman	NBN	RNHA	RBKC/ interns
Number of notes	938	225	103			
Reflective content	293 (31%)	207 (92%)	57 (55%)	21	12	24/17

Note: At RBKC two groups used the Talk Reflect App, two departments together (RBKC) and a group of interns across departments (interns). Results about reflection notes is the only data considered in this deliverable of the second group as they did not provide post-questionnaires.

The following list summarizes the descriptions of the nine coding categories (long versions are given in Appendix 5.2) into three main stages of reflection:

Stage 1

1) **Experience**/mentioning of an **issue**
2) **Emotions**, a) own, b) other
3) **Interpretation/justification** of actions

Stage 2

4) **Linking** an experience to **other experience**
5) **Linking** experience to **different piece** of knowledge, rules etc.,giving advice
6) **Alternative perspectives**
7) **Working** on a **solution**

Stage 3

8) **Insights/learning** from reflection
9) **Conclusions/implications** from reflection

Table 8 lists the proportion of reflective elements (relative to the number of analysed notes) per evaluation for each category and stage. Due to the different types of notes (individual notes vs. conversations), they are not directly comparable and thus we refrained from aggregating the data. Overall, the data shows, that the vast majority of notes provides some description of an experience (Stage 1). Stage 2, i.e. reflecting on experiences including analysis and potential solutions, is covered by most conversations captured with the Talk Reflect, some of the notes entered in KnowSelf, and a very small proportion of MMA notes. The highest stage of reflection – learning or change resulting from reflection – was found in up to 33% of the individual notes from KnowSelf and conversations from Talk Reflect. Notes entered in connection with moods in the MMA did not contain elements of stage 3.

Table 8. Proportion of reflective elements per coding-scheme-category

App	**MMA** (individual notes)		**KnowSelf** (individual notes)	**Talk Reflect** (conversations)		
Test beds	**BT**	**RG**	**Infoman**	**NBN**	**RNHA**	**RBKC/ interns**
Stage 1: Provision and description of experience, but no (explicitly) traces of reflection **(Code 1-3)**	1: 59% 2a: 77% 2b: 21% 3: 7%	1: 23% 2a: 34% 2ap: 8%* 3: 7%	1: 95% 3: 58%	St 1: 100 %	St 1: 100 %	St 1: 95.8% / 100%
Stage 2: Reflection on experiences, including analysis and potential solutions, but no (explicit) mentioning of learning or change **(Code 4-7)**	4: 1% 7b: 0.42%		4: 2% 6a: 2% 7a: 12% 7b: 7%	St 2: 95.2 %	St 2: 83.3 %	St 2: 95.8% / 94.1%
Stage 3: Learning or change resulting from reflection explicitly mentioned **(Code 8, 9)**			8a: 9% 8b: 11% 9: 14%	St 3: 0 %	St 3: 16.7 %	St 3: 33.3 %/ 23.5%

*Note: *2ap: own physical condition; St = stage*

Summarized the results obtained on the **learning level** indicate that participants started out with a rather high general tendency to reflect, especially on an individual level, which did not change over the course of the evaluations. Exceptions are found for the training context and health sector, in which the scores decreased – most probably due to a changed understanding of the meaning of reflection. Mean ratings for app-specific reflection questions indicate that the greatest support is given in the emergency sector, followed by the health, and finally the business sector. With regard to the learning outcome questionnaire items have been rated positively as well. Again, the emergency sector reports the highest perceived learning outcome followed by the health sector. Similar, ratings from participants trying the apps in a training context are higher than those from a work context. However, since the emergency sector only tested apps in training context, analysis are not independent. Across all evaluations, it could be shown that usage time is positively related to perceived reflection support and learning outcome. Furthermore, the latter two variables are inter-related with the general tendency to reflect, whereas the knowledge/skills item is not connected to any of the other analyzed variables.

Finally, a systematic coding of participants' reflection notes reveals a high proportion of pure descriptions of experiences or emotions, as well as making explicit links between those experiences and other pieces of knowledge or working on alternative perspective or solutions. Notes or conversations regarding the highest stage of the reflection coding scheme, which covers gained insights or drawn conclusions, are clearly visible but much smaller in number. However it has to be noted that we can only report on the reflection processes captured in the tools. That of course does not cover all reflection done by the participants during the evaluation time. We know at least for groups which are co-located that they also exchanged experiences and insights face-to-face. Additionally, apps were developed targeting at different goals which is also reflected in the notes found in the apps.

E.g., the MoodMap App focused on capturing moods which explains the high proportion of reported moods while the Talk Reflect with its focus on collaborative reflection of course leads to different content.

2.2.3 Level 3: Behavior

To answer the question, whether participants changed something in their work behavior due to app usage and reflection the toolbox provides one **core question** (CB1) and a **set of work-related questions** (WK01-14). The core item and some examples for the work-related items are given in Table 9, the whole set of WK items is listed in Appendix 5.2. It has to be noted that the core question is only applicable to long-term evaluations because it asks about an actual change in behavior, which can of course not be observed in short-term settings, where the questionnaire was mostly filled out directly after using the app. The additional questions on work behavior ask in part for intentions or plans to change one's working behavior and could therefore be used in short-term evaluation as well. Out of the 14 WK items in the toolbox, 12 have been used in evaluations of 4 different apps. For the further analysis the average rating across the presented WK-items per person was used.

Overall, participants' ratings regarding improved work performance (CB01) were neutral (M = 3.11, SD = 1.08, N = 153), those for the work-related items (WK) slightly positive (M = 3.57, SD = .84, N = 120). Thus participants could benefit from the apps in rather specific ways related to single items of the WK scale than with regard to their overall working behavior.

Table 9. Items used to evaluate the apps' on Level 3 – behavior

Topic	Item Code	Item text
Core Question	CB01	[The app] helped me improve my [work performance]*.
Work behaviour Examples	WK01 WK03 WK12	I used my learning on the job [The app] increased my work satisfaction. Using the app made me more confident that I can succeed in my work tasks.

*Note. *replaced in each evaluation by a specific relevant working behaviour*

Figure 9 shows the **effects of organisational sector and job experience** on mean ratings for work behaviour (WK) and the core question (CB). For evaluation context, the picture is similar with means of 3.03 and 3.69 for the work and training context respectively (means are the same for both variables, SDs vary between 0.8 and 1.1).

Because both variables are not normally distributed, effects of sector have been investigated by means of a Kruskal-Wallis test. Whereas we found no significant effect on the core question, i.e. whether work performance was improved, organisational sector does influence the mean WK ratings ($\chi_{(2)}$ = 7.5, p =.024, N = 120). Post-tests reveal a significant difference only between the emergency and health sector (p = .019). Looking at the impact of evaluation context the higher mean ratings derived in the training context both differ significantly from the work context (for WK/CB U-test results in Z = 3.06/2.47, p = .002/.013, N = 120/153).

With regard to job experience, U-tests yield no effect on WK, but on the core question. More explicitly, employees with less than five years of experience give significantly higher ratings than their more experienced colleagues, when asked whether the app helped to improve their work performance or some specified working behaviour (Z = -2.24, p= .025, N=124).

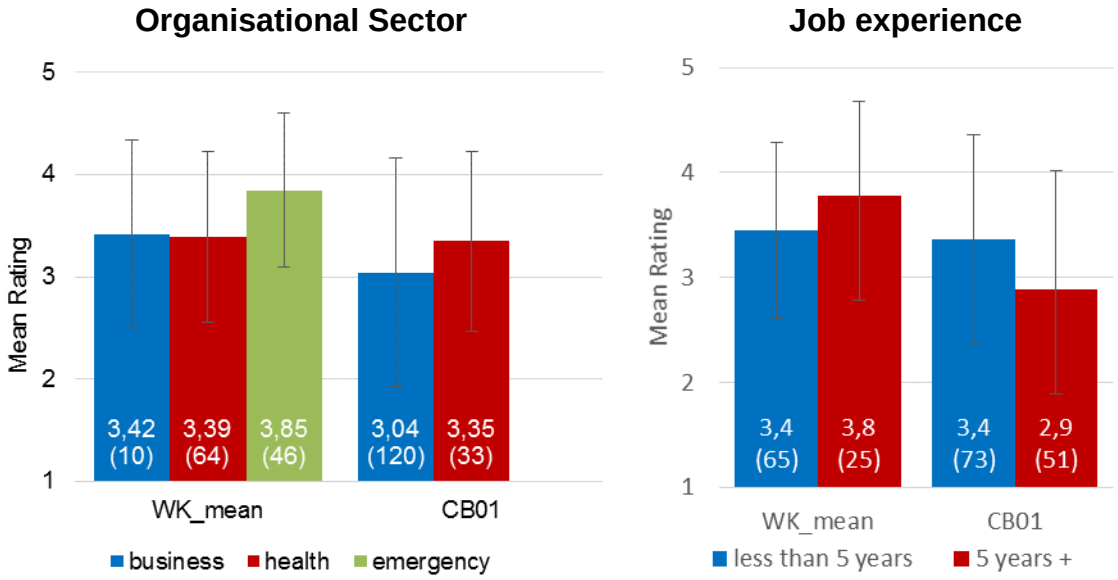

Figure 9. Effects of organisational sector and job experience on mean ratings (N) for behaviour variables

Looking at the **qualitative data from interviews,** similar to reflection outcomes behavioural change is very much specific to work and reflection focus. Nevertheless we will again highlight some concrete changes reflection supported by the MIRROR apps could trigger. Reflection on time management for example led for many participants to a more structured and efficient work behaviour. They invested more time on making plans, dealt more consciously with interruptions and tried to prevent too much work fragmentation. Work with the MoodMap App led in some cases to more attention to the mood of advisors. They tried to reflect more on negative experiences in order to be less affected by them. But, at least in the case of BT, also managers monitored the moods of their advisors and reported situations where they spontaneously intervened when they realized an advisor had a bad mood. Also collaborative reflecting had effects on work behaviour, e.g. the manager of a group of interns using the Talk Reflect App reported that they had adapted work practices after using the Talk Reflect App.

Summarised, the overall investigation of the effects MIRROR apps have on participants' **behaviour**, indicates that users could to some degree improve their working behaviour, which is especially true for the health sector and employees with less than five years of experience. For other work-related aspects, such as increased satisfaction or confidence with one's working task, highest ratings were derived from the emergency context and the group of more experienced employees. Interview data revealed insights about concrete behavior changes and therefore supported the quantitative results.

2.2.4 Level 4: Results (Organisation)

Starting out with the report of results concerning the last and highest level of Kirkpatrick's evaluation model, it has to be noted that we actually do not have any results on an organisational level as apps were only tested in certain departments or teams. Additionally, as already mentioned for the behavior level, KPIs can only be observed for long-term

evaluations. For some Serious Games short-term evaluations we tried to capture possible long-term effects with follow-up questionnaires but unfortunately participants did not answer them.

In this section we will thus present data regarding the results of the fourth level of Kirkpatricks' model that is based on a smaller scale level than organisation, but still fulfills some criteria of KPIs. The data is organized as follows:

- **Objective KPI data** from the testbeds on team level: MMA@BT, IAA/IMA@BT
- **Pre-post data** from individual participant questionnaires: MMA@BT; MedicalQuiz@NBN; KnowSelf/ARA@IMC; KnowSelf@Infoman; DoWeKnow@Infoman; IAA@NBN, BT
- **Retrospective data** from individual participant questionnaires: TalkReflect@RNHA, NBN, LondonBoroughs; IAA@NBN, BT; MedicalQuiz@NBN; MMA@Regola, DoWeKnow@Infoman
- **Loyalty metric (Net promoter score):** recommendation question for all evaluations

As there are no core questions for this level due to the great variety of relevant KPIs we deviate in this section from our overall strategy to report variables only if at least evaluations of three apps used them.

KPIs – objective data

Objective data from in the organisation established KPIs was only available for the testbed BT. Two KPIs were used for the MMA evaluation as well as for the IAA/IMA evaluation. For MMA there exist additionally data on three other KPIs, for IAA/IMA for one more. These will only be summarized shortly here, but are described in detail in D10.3. The two KPIs measured in both evaluations on a team level for three periods of measurement (before, during and after the respective evaluation) can be described as follows:

- *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)*: 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 %.

As we have data on *NPI* and *Advisor Sat* only on team level we did no statistical analysis but only report the descriptives and describe tendencies. The mean of the NPI over all four teams (MMA-1,2, IAA/IMA-1,2) which used the apps increased very slightly over the three points of measurement (pre: M=42.25, SD=17.17, during: M=43.00, SD=10.10, post: M=47.75, SD=19.82), while there is a notably decrease for the control group which did not use the app.

Figure *10* shows the trend for teams.

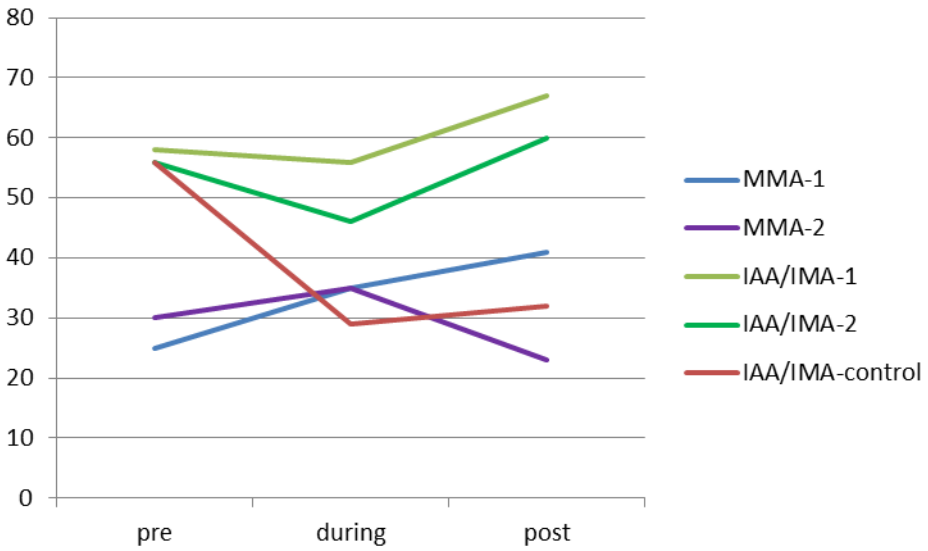

Figure 10. NPI of 5 teams for MMA and IAA evaluation at BT pre, during and post the evaluations

A similar picture shows the trend of the *Advisor Sat*. There is a really slight increase in the experimental teams (pre: *M*=88.00, *SD*=4.69, during: *M*=89.75, *SD*=3.40, post: *M*=89.75, *SD*=5.74) while the control group decreases over time. The scores for individual teams are presented in Figure 11.

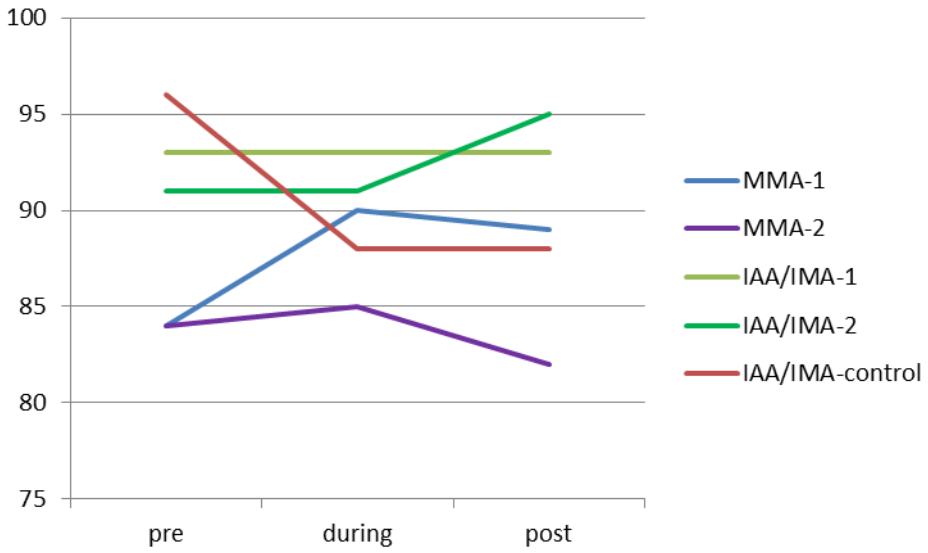

Figure 11. Advisor Sat of 5 teams for MMA and IAA evaluation at BT pre, during and post the evaluations

MoodMap evaluation at BT

For the MMA evaluation there was one more KPI reported on team level and two on participant level for two of the four participating teams:

- *Recap (team level)*: 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?'

- *Volume (individual level):* number of customers that delivered a rating to the advisor after the call
- *Average Rating (individual level):* indicates the customer satisfaction rating (0-100).

Recap developed in the same way as *Advisor Sat* in the MMA evaluation: Starting from an average score of 82, it was slightly improved during the MoodMap App period (4%), but decreased minimally after the usage cessation (2%).

For the individual KPIs there was a significant improvement for one team from pre to the during measurement ($t(18)=-3.39$, $p=.003$) in the Average Rating and significant decreases for both teams for Volume from during to post ($t_1(18)=6.30$, $t_2(14)=4.06$, both $p<.001$). All other comparison were not significant.

Issue Articulation/Management App evaluation at BT

Additionally to the KPIs mentioned above, one more KPI was monitored for the IAA/IMA evaluation.

- *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 the Average Advisor Rating, a low percentage is preferred here.

No statistical analyses were conducted as the data was only available on team level. Therefore we present here just the descriptives of the data (see Figure 12). There was no tendency notable. The testing groups started at a lower (better) 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.

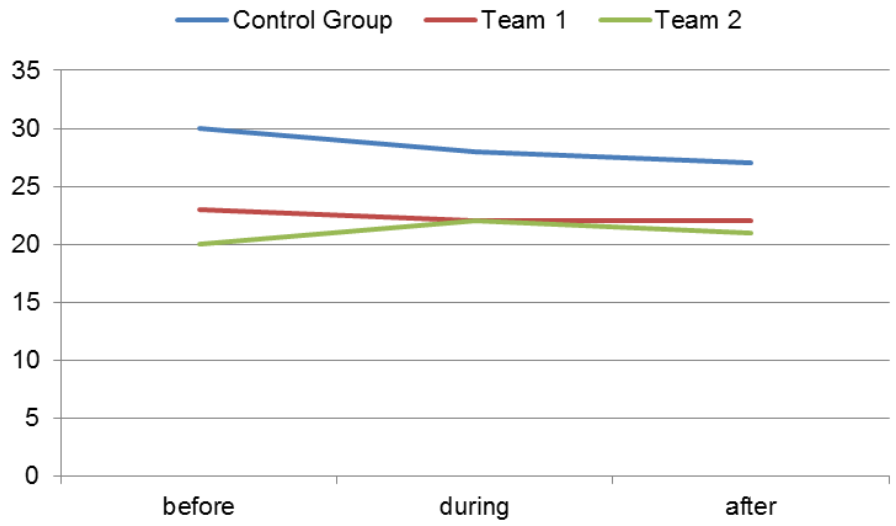

Figure 12: Repeat Calls over time

KPIs pre-post questionnaire data

For the other evaluations it was not possible to get data about KPIs monitored by the application partners. Nevertheless we collected subjective data from individual participants in the questionnaires about their work performance and other KPIs. This section reports on data which was collected pre and post app usage und could therefore be analyzed with respect to an eventual change.

For job/coaching satisfaction we conducted a one-way ANOVA with repeated measures. We found no significant improvement over time over all three evaluations for which we had data while the descriptive data shows a tendency towards improvement ($F_{(1,64)}$=3.72, p=.058).

There was no significant change in how satisfied employees were with the support for bottom-up initiated change.

For skills in time management also a one-way ANOVA with repeated measures was done. We found again no significant improvement over time, but an interaction of time and evaluation which was almost significant. As Figure 13 shows, while participants in the Infoman evaluation did not change the assessment of their time management skills, at IMC participants indicated higher skills after the evaluation (pre-post: $F_{(1,18)}$=3.69, p=.071, pre-post x evaluation: $F_{(1,18)}$=4.36, p=.051).

At Infoman the satisfaction with standard slides which were in the focus of the DoWeKnow Evaluation at Infoman did not improve significant. As we had only pre and post data from 7 participants we conducted a Wilcoxon test (Z=-1.78. p=.075).

Figure 14 shows the pre-post comparison for job/coaching satisfaction, support of bottom-up initiated change and time management aggregated over evaluations.

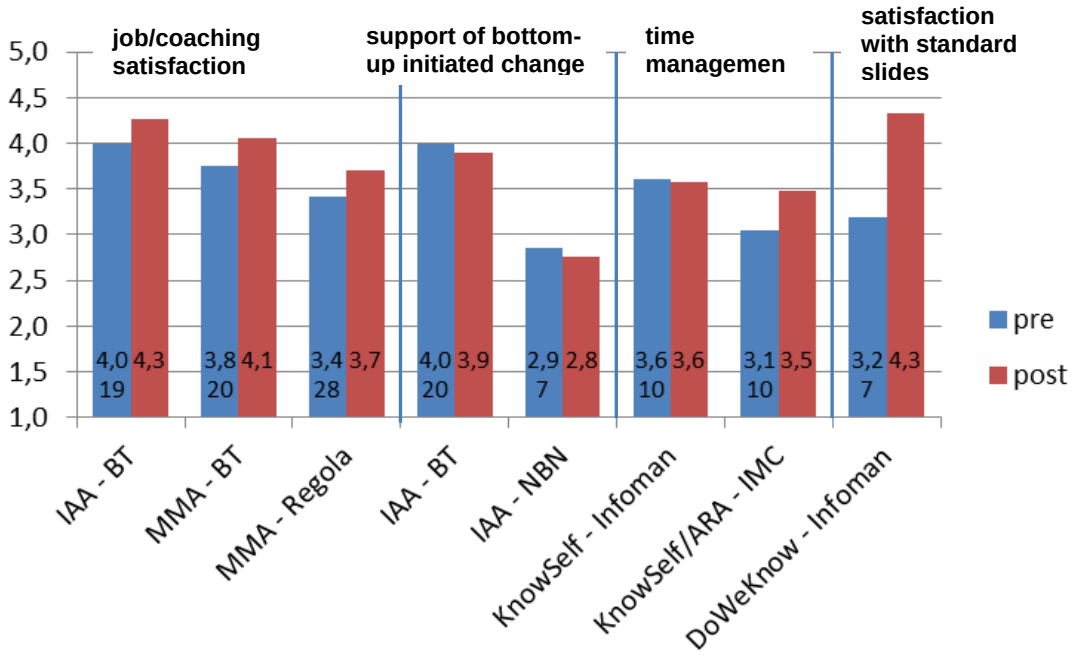

Figure 13. Pre-post Questionnaire data for individual KPIs[1] - for individual evaluations

[1] Means and N can slightly vary from the results reported in D10.3 as for the statistical pre-post comparison only data available for both times of measurement could be considered.

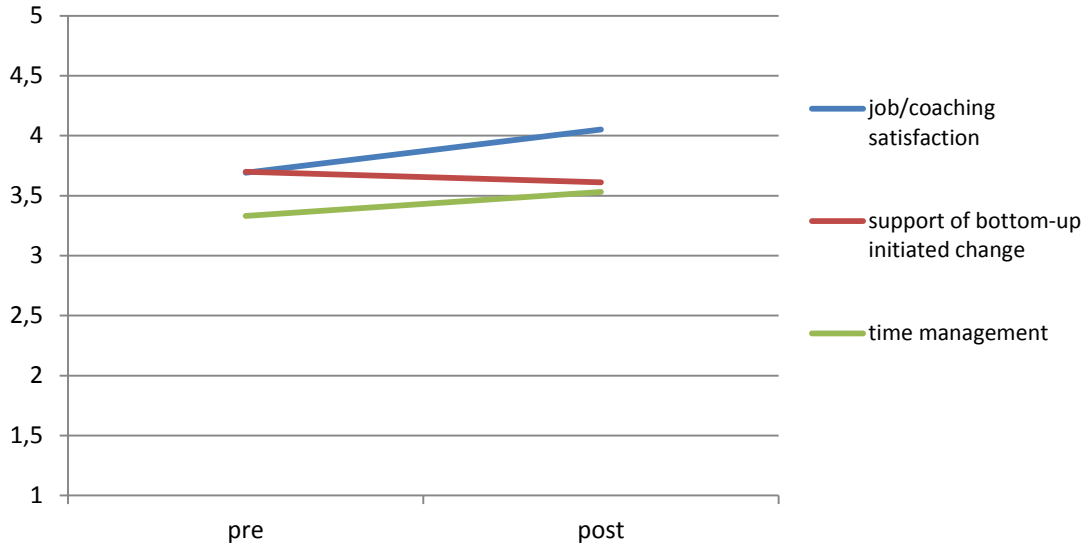

Figure 14. Pre-post Questionnaire data for individual KPIs – aggregated over evaluations

KPIs retrospective questionnaire data

This section reports on results where participants were retrospectively asked if they think KPIs on an individual level changed after using the apps.

We have data from three evaluations regarding an improvement of job satisfaction and from six evaluations for an improvement of customer (patient/resident) satisfaction. As Figure 15 shows, most evaluation participants rated this question rather neutral or slightly positive. The overall mean for job satisfaction is *M*=3.10 (*SD*=0.94) and *M*=3.01 (*SD*=1.04) for customer satisfaction. Only in two evaluations there is clear disagreement. It has to be noted that disagreement only means no improvement of the KPIs not a decrease of it.

Figure 15. Retrospective questionnaire data for improvement of job and customer satisfaction

For work performance related KPIs the picture is quite similar with neutral answers. The overall mean for work performance improvement is M=3.29 (SD=1.02). The only exceptions from this neutral evaluation are again one evaluation at NBN (disagreement) and the Infoman participants regarding the effect of the DoWeKnow App (agreement) (see Figure 16).

Figure 16. Retrospective questionnaire data for improvement of coaching, the ability to tackle difficult work and work performance

Loyalty metric (NPS)

Besides the apps' effects on employees' behavior and work performance one can also look at the success of the MIRROR apps in terms of adoption and dissemination. One possibility is to measure the chances to affect other members of an organisation and the potential uptake of the community itself by using the net promoter score which asks:

> How likely is it that you would recommend the app to a friend or a colleague? (0=not at all to 10=very likely).

The net promoter score is then calculated by the percentage of promoters (score 9-10) minus the percentage of detractors (score 0-6):

The net promoter score is a quite strict measure, because even scores in the upper half (6) are counted as detractors, whereas only the two top-scores (9-10) are counted in favor of an app. Thus, the obtained overall NPS of -28% must not be over interpreted. As for some apps we know a main point of criticism was that it contained too little content (e.g. Serious Games or Medical Quiz) we also asked for some of these the loyalty with the addition of "with more content in it". This score gives a more realistic picture how apps would be evaluated after the prototype version. This score reaches a value of 38% which is quite positive. We also analyzed the mean rating for the question to show the general attitude towards the apps. This average rating rather fits to the results reported so far with a slightly positive score (*M*=6.19, *SD*=2.79, N=239). The loyalty score for apps containing more content than now becomes notably positive (*M*=8.35, *SD*=1.79,N=71).

For the Talk Reflect App the question was already used in a slightly different form ('I would recommend the app to a colleague.' on a scale from 1= *totally disagree* to 5=*totally agree*) in formative evaluations in year 3 and then kept like this due to comparability reasons between evaluations of the same app. The mean of 3.63 (*SD*=0.98, *N*=32) is in line with that of the other evaluations with a slightly positive outcome.

In order to investigate differences between sectors a one-way ANOVA with organisational sector as independent factor revealed a significant effect on the mean of the loyalty metric ($F_{(2,239)}$=33.41, p<.001). Post-hoc tests yield significant differences between all three sectors with emergency being the most positive participants (all $p \leq .008$).

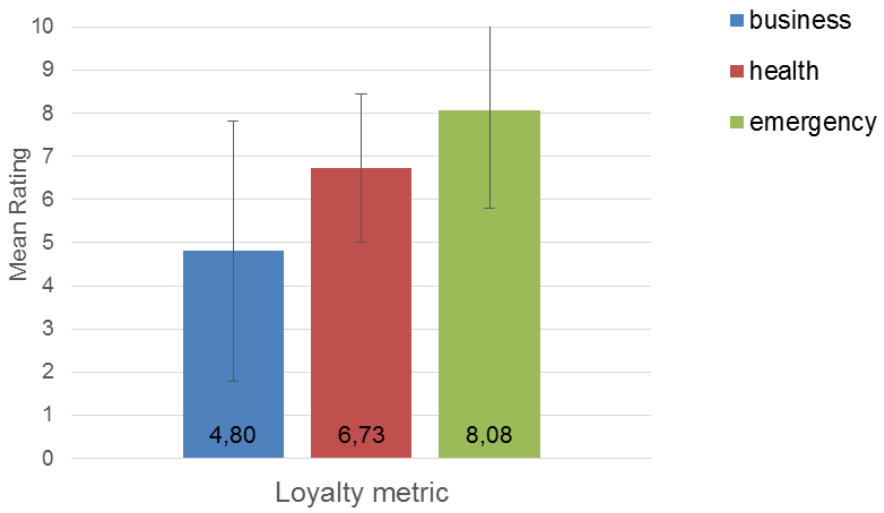

Figure 17. Loyalty metric – comparison of sectors

With regard to job experience no significant differences were found between users with more than 5 years of job experience and users with no or only little job experience (t(158)=1.9; p =.053; <5y: *M*=6.08, *SD*=2.5, N=98; 5y+: *M*=5.21, *SD*=3.1, N=62).

Qualitative data from interviews:

Data from interviews highlighted mainly two aspects. On the one hand several participants confirmed improvements in their work performance, e.g., working more efficiently after reflecting on their time management and even being more satisfied. They also mentioned that improvement achieved on an individual level would aggregate to an improvement of the whole organisation. On the other hand the limitation of purely individual reflection became obvious. In some cases people noticed that they had just a too high workload which could not be improved purely by better time management and needed their manager to find a solution. Others pointed out that time management can not only be improved on an individual level but also processes and collaboration would have to be optimized on a more global level. This became also obvious in the MMA evaluation at Regola where managers did only use the app on an individual level and missed the chance to improve things on a team level by monitoring the moods of their employees and supporting them.

Summarized we found only slight to no improvements overall in **KPIs over time**. This is not surprising as people have to use apps long and intensive enough to receive any effects. Additionally KPIs are influenced by a range of factors the MIRROR apps being only one

(potential) among them. So even if changes were found we cannot directly relate them to the introduction of reflective learning as other factors might have changed during the same period.

2.2.5 Summary and conclusion of data analysis along Kirkpatrick's model

Concluding this section, we can say that participants reacted positively to the MIRROR apps, they indicated to be satisfied with the apps and that they are in the long-run useful for professional competence development. Most positive reactions came from the emergency sector and users with more job experience rated the usefulness for professional training higher. Despite some barriers for usage, which concern mainly time and space to use the apps (also e.g., because of especially stressful periods with their primary work tasks or due to a lack of internet access or other technical problems), participants generally tended to be in favour of continuing their app usage.

Regarding the learning process, participants showed a high initial level of reflection, especially for individual reflection. This did in general not change over the course of the studies. However, in the health sector and in evaluations conducted in a training setting, the SRS scores decreased, which we attribute to a changed understanding of the meaning of reflection. The app-specific support of reflective learning provided by the MIRROR apps was perceived positive by the participants and participants also reported a positive learning outcome. The latter is also related to higher usage times, higher general tendency to reflect, and perceived reflection support. Notes entered by the participants into different applications also showed that participants did reflect about their working experience by making links to previous experiences or other pieces of knowledge or by working on alternative perspectives. However, documentations of reflection outcomes are rare.

With respect to the effect MIRROR apps had on the working behaviour of participants, the results indicate that users perceived some improvements in their behaviour at work, as well increased satisfaction or confidence with the working tasks. On the results level we found only marginal changes and these should be interpreted carefully as changes on this level are influenced by many different factors.

With the performed overall analyses across all evaluations we also investigated whether the impact of the tested apps differs with respect to some context-related variables. On the one hand, we looked at the different organisational sectors involved in the evaluation studies. It turned out, that the emergency sector perceived the apps' support most positively and also reported the highest learning outcome, increase in work satisfaction and confidence with one's working tasks. This is followed by the health sector, whereas the apps were least successful in the business sector. Similar, ratings from participants in a training context are higher than those tested in a work context and employees with less job experience seem to benefit more from MIRROR apps than their more experienced colleagues. The latter gave higher ratings for the usefulness of apps for professional training, thus it might generally be concluded that the apps are for the most part more successful in the context of professional training than integrated in the primary work process.

2.3 Data Analysis along the CSRL Model

In this section, the main focus is to investigate how supportive the apps actually are in different phases of the reflection process. We refer to D1.4, D1.4b and D1.5 for detailed descriptions of the basis of this evaluation work, but give a short summary of the main aspects. The core of the CSRL model, as it is depicted in Figure 18, is a **reflection cycle**

with the four distinct stages plan and do work, initiate reflection, conduct reflection session, and apply outcome.

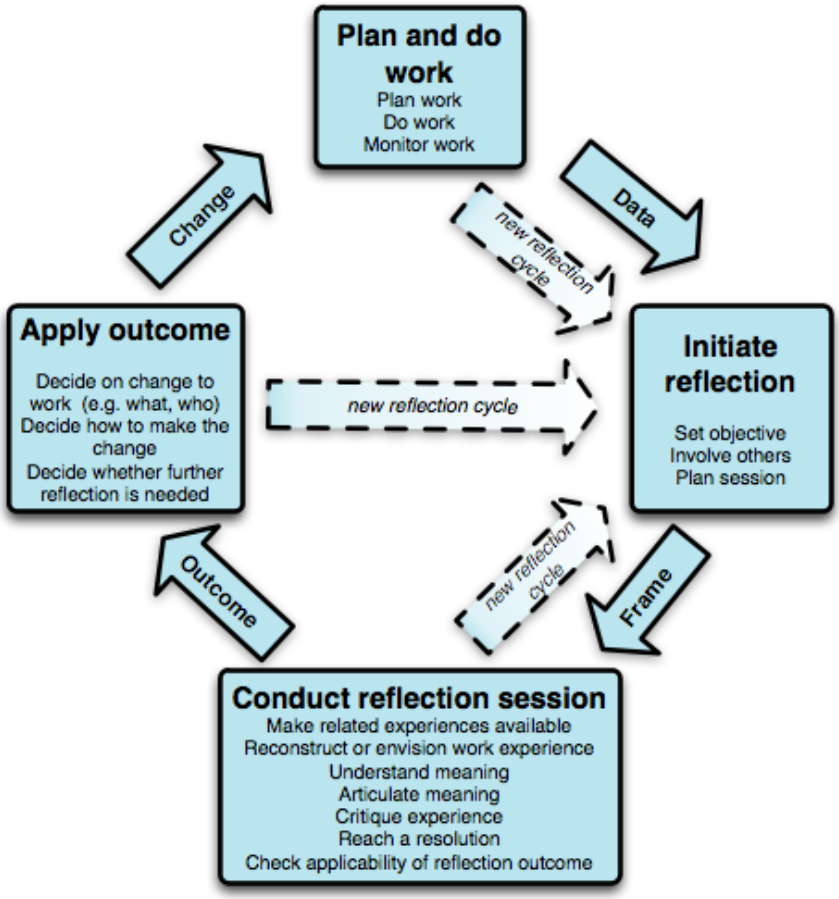

Figure 18. CSRL model

See D1.4b for a more detailed description of the current model and visit the Reflection Model Guide at http://docs.mirror-demo.eu/irg/ for an up-to-date clickable version of the CSRL model with detailed model information to each stage and transition as well as descriptions of application support for individual and collaborative reflection.

In a second step these types of **tool use were mapped to the MIRROR CSRL model** and simultaneously refined to cover the entire reflection cycle. Figure 19 (or correspondingly Figure 10 in D1.5) depicts the result of this process. The four phases of the reflection cycle are shown one below the other (light yellow areas) with the main tasks and processes inserted for each phase (dark yellow areas). On the left and right hand of the model, 23 different (sub)-categories (1, 2a, 2b, …12) of tool use are mapped to the four phases. These categories are on the one hand assigned to one of the four reflection phases and on the other hand divided into four groups of tools: tools for capturing data (light blue, e.g., 2b, 7c, or 10b), tools for providing data to the learner (red, e.g., 3b or 6b), tools for scaffolding the process (grey, e.g., 2a, 3a, or 5), and tools for simulating the work processes (purple, category 1). Each of the MIRROR apps supports certain processes during reflective learning und thus offers functionalities that correspond to a subset of these 23 categories.

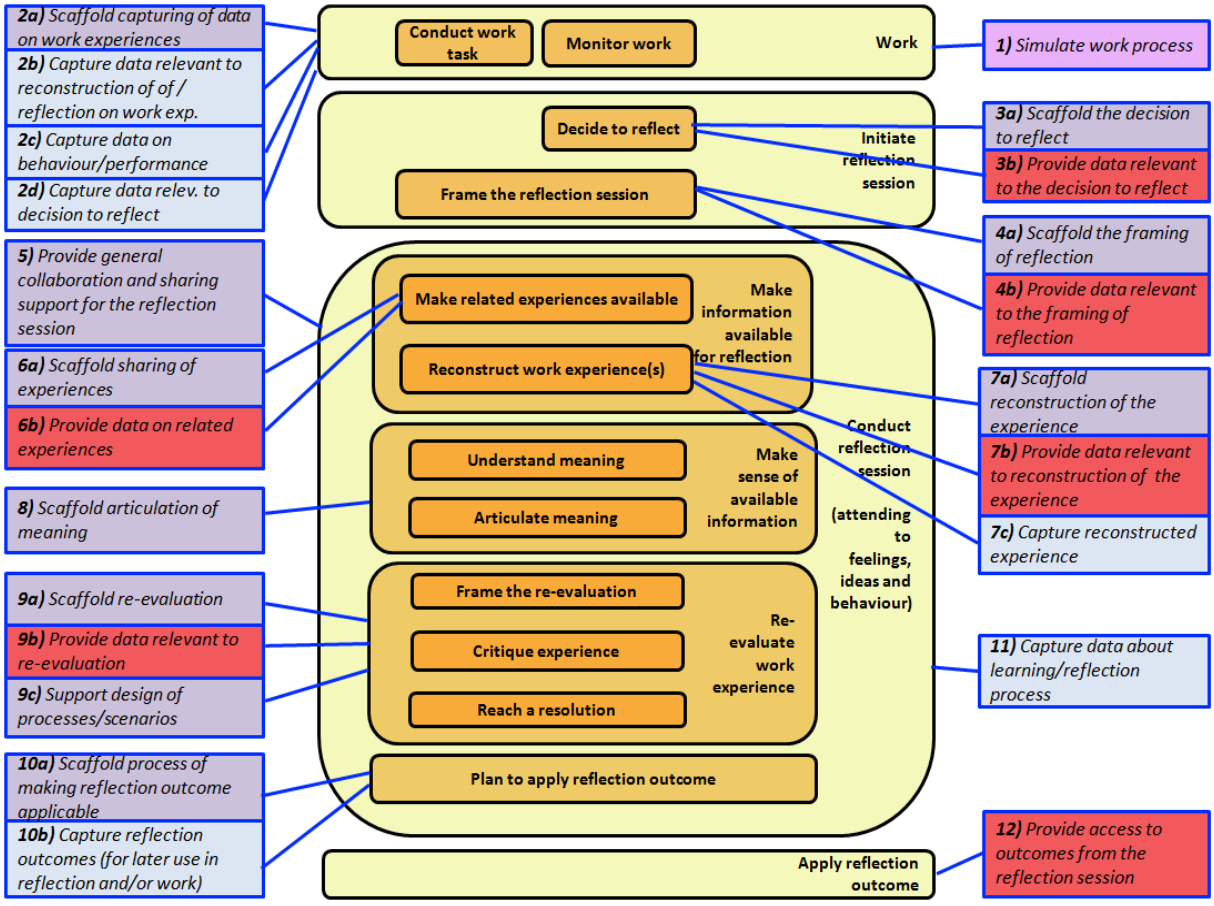

Figure 19. Categories of tool use mapped to the CSRL model

In this model version (D1.4), as it was used in the time of the evaluation methodology development, the **transitions between stages** (data, frame, and outcome) are not explicitly shown. Tool use categories for the transitions *data* and *frame* are both covered in the second stage (3b – provide data relevant to the decision to reflect and 4a – scaffold framing of reflection), for the transition *outcome* in the third stage (10a,b – capturing reflection outcome and supporting process of making it applicable). The transition *change* which originates in apply outcome and feeds back into plan and do work (changes planned and applied to work) are not covered in this model version or the questionnaire items for app-specific reflection support. This is because model and version have been derived from user studies and practical needs regarding tool functionalities. The toolbox which has been developed in year 2, has not been changed because no practical need to extend the scale emerged from later tool developments.

Finally, as a third step, 43 questionnaire items have been developed to evaluate how well the single apps support the processes they intended to support. For each of the 23 categories between one and four questions have been formulated. These questions constitute a main part of the MIRROR toolbox in which they are subsumed under app-specific reflection questions (CA). The items are specified as core questions, i.e. for each MIRROR app the relevant subset of questions was used in the summative evaluations. The amount of questions per app varies, because it depends on the breadth of functions the app in questions provides. An overview of the exact questions and their assignment to each of the 23 tool use categories is provided in D1.5 (Section 4.3), the list of items is also given in this deliverable in Appendix 5.1.2.

2.3.1 App-support for reflective learning in each CSRL model stage

In the following, the overall evaluation results from the app-specific reflection questions are presented in relation to the four stages of the CSRL model. Whereas we used the mean score of the scale (CA_mean) for the analysis along Kirkpatrick's model of evaluation (see Section 2.2.2.1), we consider each question separately for the analysis of app-support along the CSRL model.

Overall, 32 out of the 43 provided questions have been used in the evaluation of at least one app. Depending on the variety of tasks and processes within a stage of the CSRL model and on the number of different functions MIRROR apps as a whole provide for each stage, different numbers of items are applicable. Table 10 gives an overview of the tool use categories and corresponding items for each stage of the model. It also shows which items have not been used in the summative evaluations. It has to be noted, though, that the fact that some item has not been used in any of the evaluations does not mean that the respective function or process is not covered by the MIRROR apps. Each evaluation has been prepared individually by the respective app developers who selected the most appropriate app-specific reflection questions out the whole set. In order to keep participants motivated to fill out the questionnaires, they had to be as short as possible – thus often only the most important questions have been presented to the participants. In addition, especially for the apply outcome stage, questions have rather been discussed in interviews than presented as simple rating item in a questionnaire.

Table 10. Overview of CSRL stages, corresponding tool use categories, and questionnaire items

Reflection Stage	Categories covered by apps	Items (CA) used	Items not used
Plan an do work	1: simulate work process	42,43	
	2a-d: capture data and scaffold capturing	1,2,5,21	3,4
Initiate reflection	3a,b*: provide data for and scaffold decision to reflect	10-12,22	
	4a*,b: provide data for and scaffold framing of reflection	13	14,15,23,24
Conduct reflection session	5: provide collaboration and sharing support	40,41	39
	6a,b: scaffold sharing and provide data on related experiences	16,25,26	
	7a-c: provide data for and scaffold reconstruction, capture reconstructed experience	6,17,27	
	8: scaffold articulation of meaning	28-30	
	9a-c: provide data for and scaffold re-evaluation, support process/scenario design	19,31,32, 37,38	18
	10*a,b: capture reflection outcome and scaffold process of making outcome applicable	7,8,33,34	35,36
	11: capture data about learning/reflection process	9	
Apply outcome	12: provide access to reflection outcomes		20

*Note. *refer to transitions between stages (3b: data, 4a: frame, 10a,b: outcome)*

Figure 20 shows for the first three stages of the reflection process the average ratings obtained for the corresponding questionnaire items.

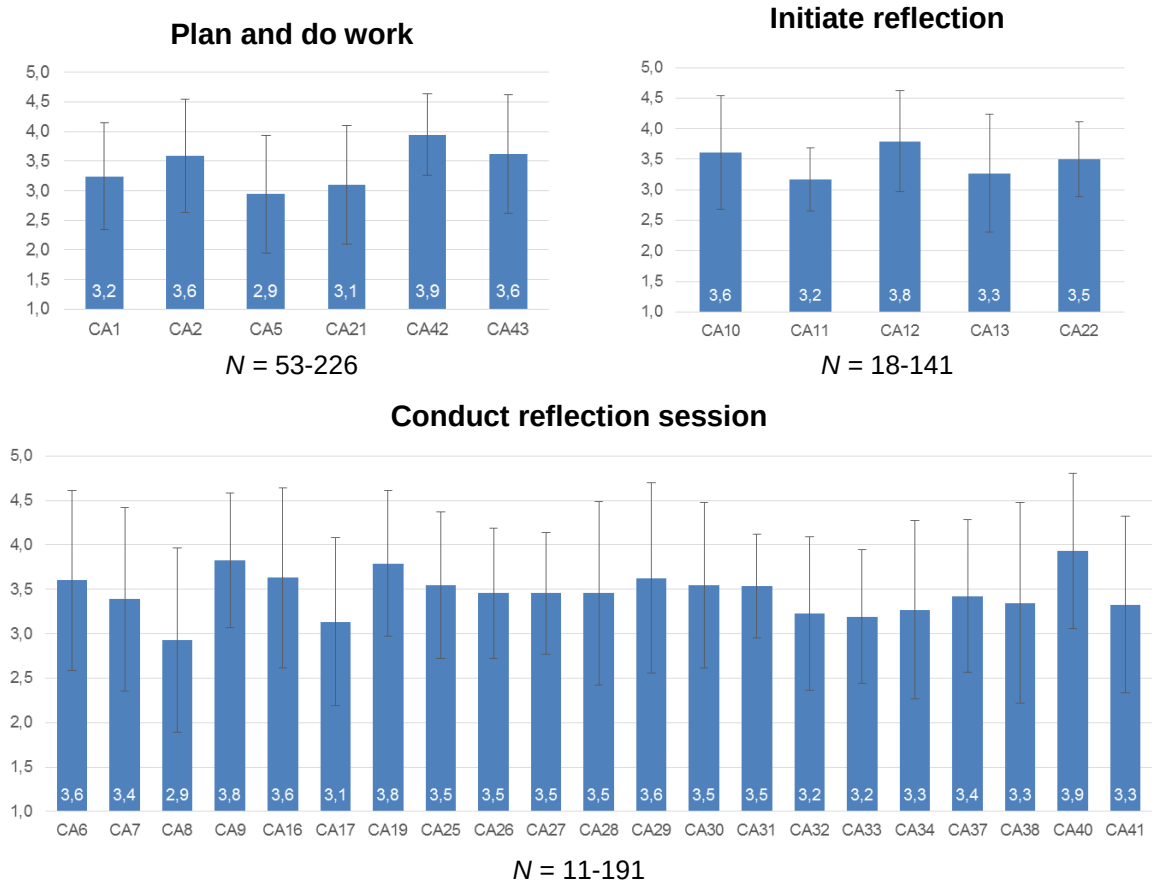

Figure 20. Mean ratings (and SDs) for app-specific reflection questions per CSRL model stage

The **plan and do work** stage (top left in Figure 20) was covered by 6 items. The mean ratings derived from 53 to 226 responses per item ranged between 2.9 for collecting information on supporting the decision when to reflect (category 2d) and 3.9 for support in simulating the work process (category 1). Except for category 2c, which refers to tools that capture data on behaviour or performance, the summative evaluations show that the plan and do work stage is very well covered by the apps and that participants view the provided functions as helpful for supporting reflection at this stage.

At the **initiate reflection** stage, reflection objectives are set, colleagues might be involved and the reflection session is planned. For this stage five different items have been presented to the participants, which covered three out of the four tool use categories specified for this stage. Missing is category 4a, which refers to the transition frame (scaffolding the framing of reflection). For the five presented items, between 18 and 141 responses have been obtained. Their mean ratings (see Figure 20 top right) indicate a rather positive view of the app support ($3.2 \leq M \leq 3.8$) for initiating reflection.

Conducting a reflection session can be seen as the main process in the reflection cycle, as it deals with the actual process of attending to one's experiences, feelings, ideas, or behaviours, re-assessing and understanding them, and drawing conclusions out of this process. The comprehensiveness of this stage is reflected in the number of categories and corresponding items assigned to it. Altogether there are 11 (sub)categories with 19 items for this stage plus two subcategories with six items provided for the transitions outcome, which

already feeds into the apply outcome stage. Out of this item set, 21 different items have been presented to the participants. Responses stem from 11 to 191 individuals, who rated the app-support in this phase of the reflection cycle from *M* = 2.9 to 3.9 (see bottom of Figure 20). The number of different questions included in the evaluations shows the variety of functions MIRROR apps provide as a whole. The perceived support of the functions for conducting reflection sessions is for the main part clearly positive (for 12 items *M* ≥ 3.5).

As mentioned above, the item (CA 20) covering the **apply outcome** stage has not been presented to participants of the summative evaluation studies. But information to this stage has been collected in interviews with participants and managers, focus groups and open questions of questionnaires. Participants reported from changing strategies regarding time management and implementing changes w.r.t. time planning, dealing with interruptions and focussing more on work tasks to prevent work fragmentation. Other participants reported on planned changes regarding emotion regulation. They mentioned the decision to let things at work affect them less, to take care about "appropriate moments in which to do a break" and they wanted to try to be more positive regarding distressed customers. In a collaborative reflection setting interns from different departments reflected on how to deal with difficult situations and their manager reported changes in their work processes due to that.

However, there exist also examples where 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.

2.3.2 Summary of app-specific data per CSRL model stage

Aggregating all items that have been presented to evaluate the app-specific reflection support for a single stage, allows for a comparison of the three stages covered in the summative evaluations. Figure 21 depicts the CSRL model and summarizes for each stage of the reflection cycle the number of items developed and used to evaluate this stage, the number of responses (data points) gained in the summative evaluations, and the resulting median. The medians for the three stages plan and do work, initiate reflection, and conduct reflection session give evidence that for each stage at least half of the participants agreed (rating of 4 on the 5-pt Likert scale) or strongly agreed that the tested apps support the reflection process by means of the functionalities they provide (all *Md* = 4).

To get a more differentiated view of the support per reflection stage, on the right bottom corner of Figure 21 the means and standard deviations derived from the averaged responses per stage are depicted as bar graph. With mean ratings ranging from 3.46 (*SD* = 0.88) for conduct reflection session to 3.71 (*SD* = 0.77) for initiate reflection, the data indicate that there are no extreme differences regarding the perceived support for each stage. A Friedmans test for repeated measures also results in no significant differences among the stages (*p* = .752).

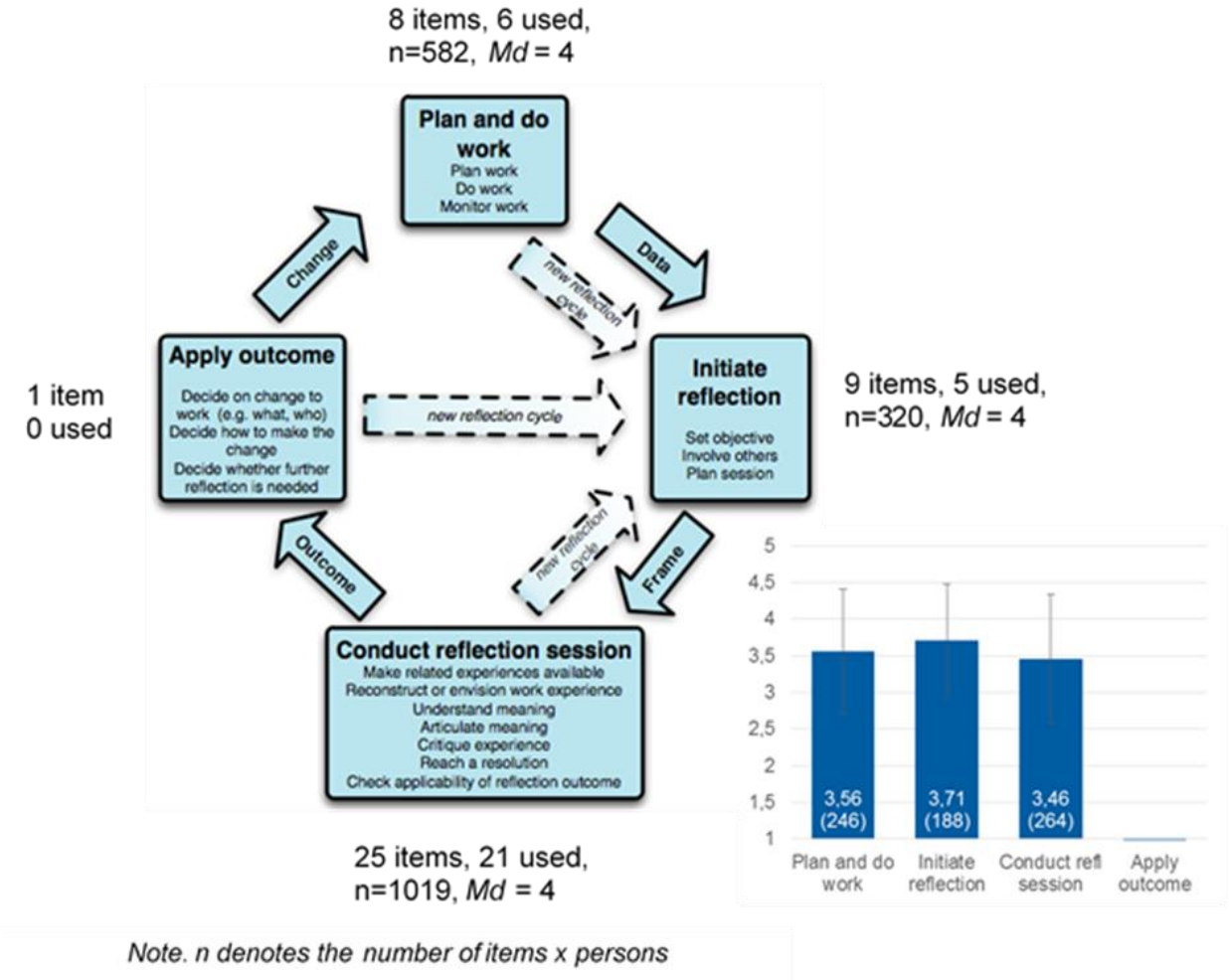

Figure 21. App-specific reflection support per CSRL model stage

2.3.3 Effects of evaluation sector, context, duration per CSRL model stage

In this section, the variables investigated in relation to the four levels of Kirkpatrick's model, are also linked to the stages of the CSRL model. This should allow for a more differentiated picture regarding the perceived support of the apps during different phases of the reflection process. Figure 22 shows the effects of organisational sector, evaluation context, evaluation duration, and job experience on the mean ratings for the respective items at the three different stages of reflection. As already mentioned above, there is no data available for the apply outcome stage and there is no main effect of reflection stage. The exact results from statistical analyses are summarized in Table 11.

For the organisational sector the general picture is the same as already found along Kirkpatrick's evaluation levels. The emergency sector perceives the apps as most supportive, followed by the health sector. This is true for all three stages with significant differences among the three sectors at all points of the reflection process. Kruskal-Wallis tests yield significant main effects of sector for all three stages. Post-hoc pairwise comparisons show that the differences are true for all sector pairs at all stages with only one exception: business and health do not differ at the plan and do work stage. There are also slight differences in the data patterns of each sector, e.g. the emergency sector rates the support for conduct reflection session highest, whereas the business sector perceives the app-support for this stage as lower than for the other two stages.

With regard to the evaluation context, ratings from participants testing the apps in a training context are in all stages significantly higher than those testing the apps during their normal work. Differences are especially strong in the conduct reflection session stage, which can be attributed to the decreasing ratings of the work context in this stage. Very similar results are found for the effect of duration, with significantly higher ratings from short-term evaluations. It needs to be pointed out again, that the two variables context and duration are confounded, because most studies that took place in a training context have been short-term.

Finally, the effect of job experience reveals a different picture. Although the ratings of participants with less than five years of job-experience are slightly higher for the stages do work and conduct reflection session, these differences are statistically not significant. But there is a significant difference in the initiate reflection stage, for which participants with longer job experiences perceive the apps as more supportive than their less experienced colleagues.

Figure 22. Effects of evaluation sector, context, duration, and job experience per CSRL stage

Table 11. Effect of sector, context, duration, and job-experience per CSRL model stage

	Plan and do work			**Initiate reflection**			**Conduct reflection session**		
	N	Test-stat*	*p*	*N*	Test-stat*	*p*	*N*	Test-stat*	*p*
Sector (df=2)	246	29.75	<.001	188	20.92	<.001	264	52.14	<.001
business–health		-2.24	.075		-2.59	.029		-3.47	.002
business–emergency		-5.42	<.001		-4.53	<.001		-7.12	<.001
health–emergency		-2-83	.014		-2.81	.015		-3.73	.001
work–training context	246	4.63	<.001	188	2.91	.004	264	6.46	<.001
long-term–short-term	246	4.63	<.001	188	4.9	<.001	264	6.88	<.001
< 5 years–5 years + experience	176		.165	151	2.10	.036	194		.383

*test-statistics: χ^2 *(df=2)* from Kruskal-Wallis tests or standardized *Z* from Mann-Whitney U-tests

2.3.4 Summary and conclusion of data analysis along the CSRL model

Summarized, the overall analysis along the CSRL model showed that the evaluated MIRROR apps successfully support reflection at the first three stages of the reflection process, i.e. there are functions providing support for planning and doing work, for initiating reflection, and for conducting reflection sessions. Also the transitions between these stages as well as the transition to the apply outcome stage are covered by the tools and could also be assessed during the summative evaluation studies. However, the apply outcome stage was only covered by a single item in the set of 43 app-specific reflection questions, which was at the end not used in the evaluations. On the other hand 32 different aspects of app-support could be assessed and the data show that they are perceived as being supportive by at least half of the participants (median ratings for all three stages amount to *Md* = 4). From a purely descriptive perspective, the highest mean rating was obtained for initiate reflection, followed by plan and do work, but the differences do not reach statistical significance. The effects of organisational sector, evaluation context, and duration found in relation to the Kirkpatrick model proved to be valid in each phase of the CSRL model. More explicitly, during the whole process of reflective learning (except for apply outcome) app-specific reflection support was perceived higher whenever the apps have been used in the emergency sector (followed by health), in the context of a training-setting, and for only a short period. With regard to individual differences, participants with longer experience at their current position rated the support for initiating reflection higher than their colleagues. Overall, it can be concluded that the support of the apps for reflective learning works well regarding the first three stages of the reflection cycle. Regarding the apply outcome stage so far not many apps did focus on that aspect. But apart from supporting the documentation of reflection outcomes and experiences with implemented changes, support for actually applying reflection outcomes and the transition of changing back into the plan and do work stage might be something that rather needs support by a coach or managers who support employees in their reflection. Thus this last stage of the reflection cycle often takes place outside of the apps.

3 Lessons Learned: Part 2 - Project Perspective on Reflective Learning

This part of the deliverable is concerned with insights on the project side, i.e. the experiences gained by the MIRROR partners during four years of researching technology-supported reflective learning. We report insights w.r.t. different forms of reflection, technical, organisational, and context-related aspects for the successful introduction of reflective learning, potential and effects of the apps in organisations, and also methodological aspects to be considered in such comprehensive and diverse evaluation studies. As we collected the contributions from MIRROR partners we considered the perspective of all scientific partners, app developers, as well as application/testbed partners. Additionally, insights from a related discussion session at the last general assembly of MIRROR (GA8), have been integrated. This way it was possible to gain comprehensive insights on reflective learning at the workplace and to examine the topic exhaustively.

The base of this part are individual contributions of all partners from the consortium answering the following questions:

- Leading Question: What are your 'lessons learned'?
- Additional questions:
  - What did you learn through the project?
  - What aspects of reflection support with apps did work well?
  - What problems were you faced with?
  - What would you advise others who would like to support reflection at the workplace with apps?

We then aggregated the aspects mentioned by the different partners and, based on the partner contributions, the following main topics emerged:

1. **Potential** for reflective learning
2. **Forms** of reflection
3. **How to successfully introduce reflective learning?**
   a) **Technical** aspects
   b) **Management support**
   c) **Testbed** characteristics
   d) **Introduction** of reflective learning & apps
   e) **Data capturing**
   f) **Long term** process of reflection

4. **Effects** of apps in different testbeds
5. **Evaluation** aspects

In the next sections we will report in more detail about these aspects.

3.1 Potential for reflective learning

Especially during the user studies (see D1.2) conducted in the first year of the project the potential for reflective learning was researched. One important insight was that reflective learning has already been part of the working processes in our testbeds, but only in a non-systematic way (and to different degrees). Thus the systematic support of reflective learning by adequate technology was a great opportunity to improve reflective learning practices at the workplace. Main issues would be to provide an objective basis for reflection (e.g. by on-going data capturing), to scaffold reflective learning process, and to share the outcomes of reflection. It also became clear during the user studies that there are different ways to help employees in learning by reflection and thus the need for several types of apps. Examples are activity tracking of ongoing work-tasks, capturing moods, devices to collect and share experiences, or apps that capture raw data by sensors or similar technologies. All these different approaches have the potential to trigger or support an ongoing-reflection process. The potential of technology and software uptake to support creative problem solving in the care domain was also identified through the project. Especially in the case of dementia care with its unique problems apps supporting creative problem solving can help to improve care. As the care sector is not very familiar with new technologies the combination of reflective learning which already is a recognized skill in social care and technology was seen as something new and interesting in social care. This sector is also an example for the potential of supporting reflective learning by means of serious games which allow users to gain experiences in their field of work and to reflect on it without having the risk of negative or even fatal consequences of wrong behavior.

During the user studies a potential for the support of collaborative reflection was also found by helping to articulate experiences and to transfer reflection outcomes to organisational levels of knowledge. Regarding organisational levels it was also found that technology support could be used to make processes more transparent, thereby uncover potentials for improvements and thus increase the awareness about the advantage of using process monitoring and controlling.

While insights in this section were already gained quite early in the project the next sections report more on insights gained through the complete four project years.

3.2 Forms of reflection

Due to the variety of organisational sectors with their specific needs regarding reflective learning, also different forms of reflection emerged. These concern the level of reflection, the context of reflection, the process of reflection (and its duration), as well as how a reflection session is framed.

During the project we gained experience regarding individual and collaborative reflection and how it influences work at the individual, collaborative and organisational level. Besides considering those levels in app developments, also conceptual work has been carried out to account for these different forms of reflection and the connections between them. In WP 1

the CSRL model was developed which describes reflection on a general level as reflective cycle and suits individual as well as collaborative reflection (see D1.4 and D1.6). Furthermore, to be able to consider the special characteristics of collaborative reflection a blueprint cycle of collaborative reflection was created in WP 6 based on user studies and refined by formative and summative evaluations (see D6.2 and D6.3). The transition model, developed by WP 4, 6 and 8 (Prilla, Pammer & Balzert, 2012, see also D4.2, D6.2 and D8.2) shows the steps from a work related experience to the application of reflection outcomes and how individual, collaborative, and organisational levels can be involved. Rather early during the project time, also clear differences between the levels emerged. Regarding organisational reflection WP 8 came to the conclusion that organisational reflection does not follow the same principles as individual and collaborative reflection as an organisation cannot reflect. It is always an individual or a group of individuals who reflect. This may take place on behalf of the organisation but it is the impact on an organisational level that characterizes organisational reflective learning. To achieve changes on organisational level it is necessary to transfer reflection outcome to higher levels in the organisation as the possibilities for change are restricted for individuals on the worker level.

With Serious Games another aspects was integrated into the CSRL model as during the project it became clear that these games support the first phases of the model in a slightly different way: **Virtual experiences** are another way to gather experiences in the 'plan & do work' stage on which one can then reflect. However, the experiences are not work-integrated, but take place in a training context, just as the reflection processes are part of the training. However, this reflection process usually also considers and integrates real world experiences (by comparing the situations faced with in the game to situations experienced at work) and the outcomes of the reflection processes can and should be applied into the real-life work-setting.

One other aspect that became clear during the project was that reflection can happen as a **campaign or** integrated in the workplace as a **continuous process**. What suits best for the given situation depends among other things on the topic and the goal of reflection. For concrete goals like improving time management, app developer as well as application partner reported a campaign to be a good choice as we found in an evaluation that employees will only focus a certain time on that aspect of work, i.e. how the structure of work-processes can be improved. After a first improvement of their time-management, there is no more need to continuously reflect on it. A re-evaluation might take place after some time, but that would again be a campaign of limited duration. Other areas like improvement of dementia care or collaborative reflective learning of interns of different departments benefit more from a continuous concept. Here employees are consistently confronted with new problems as in the case of dementia care but also benefit from the experience and solutions already found by others.

During the evaluations we also found that reflection can take place in **formal sessions as well as in informal settings.** From app developer side for example it was reported that they heard about structured debriefing sessions after an emergency or an emergency training but that volunteers also shared "war stories" as a way to release stress, build identity and reputation, and as a way for collaborative sense making and reflection.

3.3 How to successfully introduce reflective learning?

In these sections we describe several potential risks for a successful introduction of reflective learning in an organisation and derive possible solutions and advices to prevent them.

3.3.1 Technical aspects

From several partners it was reported that technical aspects play an important role when it comes to the introduction of computer supported reflective learning. Besides attitudes and prejudices against technology which might still be the case in some sectors the major difficulty was an often rather limited IT infrastructure of organisations. Barriers might be no or really low access to hardware, firewalls or other internet restrictions in organisations. But well-functioning technology is necessary because otherwise users get frustrated and will not use the apps. Also a technical implementation into existing systems is often not easy or impossible to achieve but is an important fostering factor as it makes it easier for employees to integrate reflective practices into their everyday work processes. Several app developers highlighted that another aspect which fosters acceptance and adoption of new reflection supporting technology is the integration of end-users right from the beginning of the development process. It has to be made clear to end-users what it means to test technology under development instead of commercial or market-ready products. On the one hand the benefit is, that during this process developers get to know and can react to the exact needs of users while on the other hand users have to test prototypes which are never free of bugs and can run instable at some points. This helps users to form realistic expectations and prevent disappointments. The importance of such pre-information became obvious as some application partners reported that apps still being prototype versions had a negative effect on their evaluation.

Advice:

- Get a clear picture of requirements and preconditions quite early in the process and aim for the largest transparency possible between (end-)users and developers.
- Provide a stable IT infrastructure and technical support during the whole reflection process as this seems to be a crucial aspect for successful introduction and adoption of new technology.

3.3.2 Management support

All kind of partners (application, scientific) reported management support to be another crucial factor for a successful introduction of reflective learning supporting technologies into an organisation. Through the project we had several positive as well as negative experiences. Examples for the former were managers who planned reflection as part of the work process and encouraged people to reflect, examples of negative experiences were managers who gave no support because stakeholders changed positions and the new managers not seeing the added value of using the apps. Therefore it is really important that the management is convinced of the benefits reflection and reflective learning supporting technology bring to their organisation so that they can transfer their motivation to the employees and also give them opportunities (time, space) to reflect. Another crucial point is that management not only accepts, allows, or motivates employees to reflect but also that they are ready to support organisational change which might be the consequence of reflection.

Advice:

- Clearly present the benefits of reflective learning to the management level.
- Do not start to introduce your approach without strong management support. Make sure that employees will get the opportunities to reflect and that there is willingness on the management level to consider reflection outcomes and to engage in organisational change if necessary.

3.3.3 Testbed characteristics

Several testbed and user characteristics can influence the success or failure of the introduction of reflection apps. Main characteristics reported by partners were individual characteristics, job experience, work load, and privacy.

During the project it became obvious that reflection is in a lot of jobs only a secondary work process and will not be conducted if the **work load** of primary tasks is too high. And even if there is time to reflect about experiences from work it is still really helpful to integrate the reflection processes as much as possible into the primary work process of employees to keep the threshold for reflection and data capturing quite low. Basically, it needs an optimal level of workload for participants to start a reflection process. One partner reported about being successful with reflection about time management for people with a medium workload. If the workload is too high people feel they do not have the time to reflect but have to manage to somehow finish their primary work tasks first. On the other side if work load is too low there is no need e.g., to improve time management.

Other characteristics are not per se barriers for reflection but the successful introduction of apps really depends on the match between testbed characteristics and app. This is shown, for example, by the aspect of **job-experience**: some partners reported about more experienced employees to benefit more from using the apps (CaReflect, WATCHiT) or that managers (who are generally more experienced) are more prone to reflection than lower hierarchy levels (as it was experienced in the KnowSelf/ARA evaluation at IMC). On the other hand, some apps are clearly more useful for newcomers in their jobs (Virtual Tutor Serious Games). Also for time management (KnowSelf, ARA) it seems that less experienced people had a greater benefit from reflecting about it than employees who had handled time management issues before.

Regarding **individual characteristics**, an application partner reported about people being more comfortable with self-reflection than others which seemed to be quite resistant to change throughout several approaches and settings for reflection.

Another testbed specific aspect is **privacy**. While several partners report that from the participant perspective there were no privacy issues and no restraints from sharing data the organisational privacy policy has to be respected so apps should be adaptable to the needs of a given context.

Advice:

- Integrate the apps as much as possible into the existing infrastructure and the whole approach in the established work processes.
- Consider specific characteristics of the target group and try to fit the best matching approach for their needs and requirements.

3.3.4 Introduction of reflective learning & apps

According to the experiences we had in the project one of the most critical moments for the successful introduction and adoption of reflective learning is the introduction itself. Potential risks for people not using the apps and thus a failed introduction process is a lack of understanding what reflection really means and/or a lack of motivation on the side of the users. Therefore a good introduction has to address these two aspects. First, many partners reported from a lack of **understanding what reflection actually is** or how it could be conducted. So a good introduction should explain the concept of reflective learning, define goals for users and also help to interpret captured data if this will be part of the reflective learning approach. It is not enough to give people an app and ask them to use it, for most settings more guidance is needed. One evaluation could show that coaching might be a good approach to be combined with reflection. Second, people have to be **motivated to use the apps** by conveying to them what the potential advantages and benefits of technology-supported reflection are, for them personally but also for the organisation. As mentioned above reflection is a secondary process and seldom mentioned directly in the job descriptions of employees. Therefore people have to understand why they would profit from performing these processes. Benefits should be identified for all levels involved from the management level to the employee. An especially critical situation occurs if apps are more addressed to organisational than individual aspects. In this case, the individual employees have to capture experiences and give input in order to help managers to identify problems and derive solutions. Thus there is often no direct benefit employees can see for themselves. This has to be made very transparent and the secondary benefits have to be explicitly communicated to the employees, who would of course profit from e.g. improved organisational processes. An even better solution would be to create apps in such a way that they combine both aspects: to support employees in their individual reflection and to use the input on a higher hierarchical level to help managers on a team level or initiate organisational change.

Advice:

- Introduce the apps and reflective learning approach sufficiently. People need to understand the concept of reflection and how to use the apps.
- Explain the benefits of reflective learning for all involved levels. Help users to define reflection goals which will motivate them to reflect.

3.3.5 Data capturing

As reflection is learning from experiences, experiences must be available in order to reflect on them. Data capturing apps can be very useful as they help to remember experiences and also allow sharing them. Capturing data can be done in an automated way (which has the benefit that it does not require additional work load or distracts from work) or manually by capturing certain variables or writing down experiences (which might have the benefit of already triggering reflection). For a manual capturing of data one scientific partner pointed out that this must be really easy, effortless, and adaptable to the work process as e.g., emergency volunteers do not have the time and cognitive resources to focus on complex data capturing devices. For automatically captured data, the above mentioned privacy issue has to be considered as some users might not feel comfortable with automatic tracking software. This is mostly due to an (implicit) fear of being monitored or even controlled by their superiors. Thus, it has to be made transparent, who has access to the data (often only the tracked individuals themselves) and for what purposes it will be used. In cases where the data is shared, a healthy error-culture would be a pre-condition for successful introduction of

the app. It was also reported by scientific partners that users have to be supported in understanding and interpreting the captured data especially if it is about large amounts of data as for the KnowSelf or the WATCHiT App. So developers should aim at providing easy to understand visualizations and also some guidance on how to interpret the data.

Advice:

- Provide easy to understand visualizations of captured data.
- Support and guide users to interpret their data and derive insights.
- Make sure that the sharing of automatically captured data is transparent to and accepted by the users.

3.3.6 Long term process of reflection and adoption of apps

For settings in which reflection is more than a campaign (see above) the question is how to help and motivate people going long-term. While a good introduction of reflective learning and other factors like a stable technology are the base for long-term adoption there are other factors which influence if people continue using apps and reflecting or stop it after some time.

As mentioned above reflection is mostly seen as a secondary process which has to wait in the line when primary work tasks are more urgent. So people really need to have the **opportunities to reflect**. That might be reflection sessions in the form of team meetings, coaching sessions, or just a certain – pre-assigned – amount of time for themselves dedicated to reflection.

Some kind of **external motivation and social control** is also really helpful. Partners had the experience that a supporting manager or a lead user who motivates his colleagues to reflect is really beneficial for an active reflection process. This is true for collaborative reflection processes but also for reflection which is more targeted at individuals. For example, reflecting on one's individual time management benefits from some kind of social control as it was the case in the coaching scenario with the KnowSelf/ARA evaluation. If employees know that they will discuss their observations and reflection insights with a coach they probably will be more motivated and feel more committed to actually reflect than when no other person is involved.

Another issue reported by some scientific as well as application partners is that people need to see an effect of reflecting and of sharing their reflection outcomes. Especially for topics which they cannot solve by themselves it is important that employees see that there will be actions taken towards the goal from the side of the management. If people feel their input is being ignored, effort and participation in reflection will decrease. Documentation of outcomes was also reported to be helpful. A final important aspect is the definition of clear goals that users want to reach by means of reflection. This helps users to focus on certain aspects and also to see progress and therefore stay motivated.

Advice:

- Provide a good introduction of reflection and apps.
- Provide a stable technology and technology support.
- Support users in setting goals.
- Make sure users have the opportunity to reflect (time, space).

- Encourage managers or other contact persons to motivate employees and to engage as a lead user. Consider other motivating/monitoring persons such as coaches to help people to reflect and to persist in doing so.
- Follow up on reflection outcomes and give employees the feeling that their insights and inputs are appreciated and followed up on.

3.4 Effects of apps in different testbeds

Regarding effects the apps had in the different testbeds we first of all have to say that it is not possible to isolate the effect of a single app on work performance in a field study due to other external uncontrollable factors that might influence the specific users at the same time. This will in detail be discussed in section 3.5 ‚Evaluation aspects'. But also if it might not be due to external factors like structural changes or a new manager or the like it has still not necessarily been the app itself that led to increased reflective learning. As it became clear during the project, supporting reflective learning needs more than an app. It is a socio-technical approach that helps employees to learn from reflection. As we highlighted before people need to be instructed and guided. But if an approach exists of more than just the pure app of course it is not clear anymore to which extent different aspects of the approach had an effect on the reflective learning process. It might be that just the introduction of something new had already a positive effect. It could also be that a good instruction and presentation of benefits of reflection might foster reflective learning. Nevertheless during the summative evaluation in the project we could show that an approach in which the apps were an essential part could help people to engage in reflective learning and to gain insights from it.

Besides this more general aspect, there are insights some partners reported from specific effects in the different testbeds:

At NBN the MIRROR project and the insights gained through the project led to some crucial organisational changes as the workflow and work processes became more transparent for employees. Due to the processes and discussions that emerged from the project, potentials for improvement and some weak areas were identified and solutions derived. E.g., it became obvious that physicians need to document their training achievements, but that this was almost never done. Now documentation is introduced and senior physicians check on these documentations on a regular basis. Other improvements are newly scheduled department meetings and adaptions of documentations proofs.

At BT self-reflection became a useful method to understand how much someone's energy and confidence can impact the opinions of customers and peers. In the call centres self-reflection clearly enabled a bottom up approach to coaching as call takers could indicate what their feelings and issues were and managers could respond. So reflection provided support for effective coaching based on team members' needs.

At RNHA the acceptance for a technology-based approach was especially pleasing as the care sector has been notably for its lack of technology uptake, either for service delivery or for resident use. The apps provided carers with new and novel ways to tackle difficult issues in their everyday work. Another positive effect was that collaboratively used apps enabled carers to have their work and expertise explicitly recognized by their peers which again motivated them to invest more effort in care note recording (a basis on which can be reflected).

3.5 Evaluation aspects

During the project we did not only gain insights about computer supported reflective learning but also about how to evaluate such approaches. These insights are in the area of planning and conducting (summative) evaluations as well as data analysis and interpretation.

As for the introduction of the reflective learning approach itself, the key to a successful evaluation is a **contact person at the testbed** which is in contact with the participants, is preferably some kind of superior for them, and monitors the progress of the evaluation during the complete process. While it might be difficult sometimes to convince employees of the benefit of reflection it can be even harder to motivate them to take part in an evaluation of an app. For a successful evaluation it is therefore quite helpful to have a clear and realistic picture in the beginning about which people will be involved, about participation rates, and about motivation of potential test users: are they interested in using the app, will they receive a reward for taking part, will it just be a (extra) task assigned to them by their manager, etc.

To create good evaluation methods there is a need for **expertise in three areas**: technical knowledge about the apps to be evaluated, knowledge about the target group (to phrase questions in a way they understand and feel natural), as well as methodological expertise in creating evaluation methods. If an app developer is not in the lucky position to combine all three it is important to have a good coordination between experts of all three areas. With respect to log-data it should be clarified beforehand, which data can be used as indicators for a successful introduction of the app and how these data can be easily connected to other data sources. Questionnaires, interviews, and the like really need to be adapted to the possibilities and requirements of the target users to receive valid results, e.g. people with low education level might not be used to questionnaires and might be really challenged by them. And even for more experienced people evaluation might be seen as an extra effort besides using the apps and thus as rather disturbing task. Clear communication and transparency right from the beginning as well as management support are here crucial factors as well. Some scientific partner also experienced that being on-site during the beginning and the end of evaluations where introduction and final questionnaires, interviews and the like take place played a key role in getting good evaluation data.

Besides the challenges coming from data collection, **measuring reflection** and interpreting the results of such a measure has to be done very carefully, as well. On the one side participants do not always document the entire reflection process, thus there is often more reflection taking place in an organization than is found in the notes collected with the apps. Methods like reflection diaries can capture more data about reflective behavior but they might also influence it in one or the other direction. On the one hand, the documentation of reflective behavior might decrease reflection by introducing an extra effort to reflection, namely not only to reflect but also to document or evaluate it. On the other hand reflection diaries or journal can also increase reflective behavior by reminding or triggering reflection..

Furthermore, **identifying the outcomes of reflection** is not so easily done either. In some cases the realization of outcomes might need some time and cannot be observed during the evaluation period. On the organisational level we had the experience that the measurement of relevant KPIs is often not established in the organisations. And even if this is the case, an organisation has first to be convinced to hand over these internal data. Secondly, even if researchers manage to get this valuable data it is still not possible to definitely isolate the

effects of the reflective learning approach. Pre-post-comparisons and comparisons with control groups might give a hint but when conducting evaluation studies in the field with pre-existing groups as participants one cannot rule out that other processes also influenced the observed measures.

In order to meet these challenges, we found the concept of triangulation with using a combination of different methods, to be a reasonable approach to evaluate the introduction of reflective learning in organisations. First qualitative data from reflection notes, diaries, etc. constitute a direct image of the (documented) reflection thoughts and outcomes triggered by using the apps. Second, quantitative log-data recording the actual usage of apps and self-report data from questionnaires collecting participants' attitudes towards the apps enable us to capture data of many people and also to compare different groups/apps/organisations etc. Third, qualitative interview or focus group data make the picture more complete and help us to interpret the data and to understand even different results across methods (e.g. questionnaire vs. interview data from the same evaluation).

Finally, a more general point regarding the introduction and evaluation of apps at the test-beds, is the difficulty to introduce several apps in a combined approach. There was one combined study with KnowSelf and ARA which actually proved to be very successful, a second one combining MMA with IAA/IMA could not be finished, and a third study combining KnowSelf and MMA could only be carried out on a formative level (see D4.4, Section 3.4.2 for a description of this evaluation). More combined studies would have been desirable, however this requires the apps to be in a really mature status of development in order for the testbeds to be willing to learn about and integrate two or more new technologies into their daily work. Additionally individual evaluation studies of the single apps should be available in order to separate effects per app from effects resulting from the combined usage. Thus such app combinations would actually be the natural next step in the evaluation process, if the project was continued.

3.6 Summary and conclusions of lessons learned

Throughout the MIRROR project the partners of the consortium (scientific and application) gained a lot of valuable insights about reflection and important aspects for the introduction of a technology-based approach for reflective learning. In this summary we highlight again the aspects which were reported by several partners. Other aspects mentioned above were reported by only one or few partners. As turned out in discussion groups, not mentioning a certain aspect in their reports, does of course not mean that other partners would disagree or have not experienced this aspect, they did just not report it in their contribution because it did not have such high priority. Thus the following list includes those aspects which were perceived as important issues within larger parts of the consortium.

- Involving app users in the development process is an important strategy to foster technology adoption.
- Integration of software into the infrastructure of organisations is beneficial, because it facilitates usage.
- The adaptation to suboptimal technical infrastructures at the application side can turn out to be a true challenge.
- Reflective learning follows different approaches: reflection campaign vs. reflection as continuous process.

- Reflection is often perceived as a secondary process so high work load might prevent people from reflection.
    - Management support is especially important in cases of high work load (maybe make reflection a primary task).
    - Opportunities to reflect should be integrated in the work process so that employees do not perceive reflection as extra effort.
- The importance of a good introduction of the reflective learning approach cannot be highlighted too often: People need to understand what is meant with reflective learning and how they can benefit from it. This is especially important as the term reflection is used in everyday language and people's understanding of reflection often differs from the scientific meaning.
- Several characteristics can influence to which degree a single employee benefits from (self-)reflection. Depending on how experienced employees are in their jobs they might benefit from different apps and approaches and might have different topics to reflect on and areas to improve.
- Especially for reflection as a continuous process it is important that users experience the beneficial effects of reflection. People need to have the feeling that the process of reflection and the gained insights are relevant. Regarding individual reflection, this might include signs of improvement, but also a coach or superior to whom insights of self-reflection are reported. For topics which apply to the whole team/department or organisation employees need to see that insights are used and reflection outcomes are considered in organizational development.
- Challenge of transferring reflection (outcomes) on an organisational level: This is highly related to the aspect mentioned above. On the one side people need so see that their outcomes have an effect and can actual change something in an organisation, on the other side this change can involve many people and especially large organisations might need some time to change processes. A related question is: Is the organisational culture actually 'ready for reflection'? Often reflection is a bottom-up approach which might not suit strongly hierarchic structured organisations.

4 Overall Conclusion

In this section we integrate results gained from the overall analyses of user data in Section 2 with the projects side's insights of Section 3. This will give a more complete picture and also help to interpret the results found in Section 2.

A prerequisite for computer supported reflective learning is the actual app usage which might be prevented by certain **barriers**. In the overall data analysis we found 'not having enough time' to be the main barrier for not using the apps. This reflects the insight gained during the project that reflection is viewed as a secondary process besides the primary work tasks with a clearly higher priority. So when employees have an overly high workload they will not engage in reflection as this needs additional effort for a task that is not directly part of most employees' job description. Input from interviews added 'lack of motivation' as a barrier, which was rated as rather neutral in the questionnaires. Some participants did not see an advantage in using the app and therefore lacked the motivation to engage in reflection with the MIRROR apps. This stresses the importance of a dedicated introduction of app supported reflective learning demonstrating the benefits from engaging in reflection. For some apps this might be even harder as employees would maybe not benefit so much individually but their input is necessary for reflection on a management level or for coaches. If this is the case it has to be made transparent and employees need to see effects of their input on an organisational level.

All these barriers are strongly related with management support which was identified as crucial factor for the success of reflective learning during the project. Management has to see the **benefits** of reflection and to support their employees in engaging in reflection. This means on the one hand motivating people but on the other hand giving employees the time and the space needed for reflection.

The initially high **tendency to reflect** (measured via the Short Reflection Scale) did not further increase. The found decrease in the SRS score of some evaluations was attributed to a changed understanding of what reflection actually means and thus a revised self-estimation on how much reflection occurs regularly during their work. This interpretation relates quite closely to the reported experience that (a) a sound introduction of reflective learning is necessary and (b) that in most cases it is not enough to give users an app and let them go. Regarding (a) this aspect stresses the importance that the process of reflection although – or maybe because – also used in our common everyday language needs to be explained to people.

In fact a **socio-technical approach** is needed which imbeds the apps in a framework of reflection. On the one hand people need to be guided in reflection. Support is beneficial in understanding what reflection is, conducting reflecting sessions, and defining goals for reflection. This framework should also consider that apart from purely individual reflection where people can derive solutions which they can influence by themselves, in many cases reflection influences the complete organisation. Processes might have to be changed in order to allow reflection. Furthermore, reflection might facilitate the identification of problems which have to be tackled on an organisational level. Bottom-up reflection also does not fit very well to very hierarchical, top-down organized structures. So there has to be a certain organisational culture which allows reflection to happen and to change something. The need for such a socio-technical framework to actually achieve change in working behavior is also reflected in the respective questionnaire item, which shows also only slight improvements of work behavior.

On the level of **organisational results** it was rather difficult to show significant effects that can be contributed to using the apps. This corresponds to the insights partners reported. Many factors influence variables such as KPIs and also with methods such as pre-post comparisons and control groups one cannot control every possible influencing factor. Teams in organisations are not randomized as in a laboratory study and therefore may differ also in other factors apart from app usage which is the typical problem of field studies. Another challenging factor was that KPIs are not defined and monitored on a regular basis in all organisations.

An interesting result we found in the data analysis as well as in overall experiences of projects partners were differences between participants with more and less **job experience.** While less experienced participants seemed to change their work behavior more with the support of the apps, more experienced participants rated the usefulness of the apps for professional competence development higher.

Additionally we found that participants with less experience perceived lack of time more as a barrier than more experienced participants. This may be caused by the case that more unexperienced employees may need even more effort to master their primary work tasks while more experienced employees may often be also in higher positions and be more independent in arranging their work day.

While overall results showed less experienced participants to improve their work performance more after using the app some partners reported from more experienced workers to benefit more from reflection. The reason might be while overall the participants who are in their job for not such a long time of course have less experience and have therefore a greater potential to improve their work performance by reflecting on it. But still in certain areas also experienced employees can profit by reflection and may even be in the situation that they can value reflection more from the perspective of years of experiences they gained during their work life. Overall reflection seems to be especially relevant and useful where new processes happen. This might be employees in the beginning of their work life or organisational changes in processes or tasks. This might also explain why in some cases reflection as a campaign might be enough, as the need for reflection might drop when solutions for relevant problems are found. While e.g. the KnowSelf/ARA evaluation was conducted as campaign, the duration of six weeks was right for most participants while for some the motivation dropped already after four weeks – maybe because participants already felt their problems with time management to be fixed.

Another result we found was the relatively higher benefits of participants in the training contexts and the emergency sector. This might be the result of the same processes as all in the emergency sector conducted evaluations were trainings and the two factors are to some part confounded. Additionally, participants rated the usefulness of apps for professional training quite positive. One reason for the more positive evaluations of trainings might be that in training sessions (such as workshops) the work load might not be a barrier as the time is especially dedicated to the training. Also in these situations reflection might not be seen as a secondary process. This is related to partner insights reporting that managers could sometimes profit more from the apps as they used them to monitor their employees and as a consequence used the inputs to support the employees by directly reacting to the observed patterns. Thus, also in this case the apps were actually used to advise or train their users at work.

Regarding the CSRL model we could show that the MIRROR apps all together covered almost all stages and transitions of the model really well apart from the apply outcome stage which is not covered explicitly by most of the apps. For a more detailed look on how the

individual apps support the different stages we refer to D1.6 which discusses these processes in more detail.

Regarding evaluation methods our analyses show again the benefits of triangulation, the combined usage of different evaluation methods such as quantitative data from questionnaires, qualitative data from interviews and content in the apps as well as log files. Different strengths and weaknesses of the types of data can level out and taken together all these data produced a much deeper picture than data from only one source.

To conclude this report, the overall analysis of data gathered in very different evaluation settings combined with the individual insights scientific and application partners from the MIRROR consortium gained throughout the project, show that the introduction of technology support for reflective learning at work is able (a) to trigger new reflection processes on individual, team, and organisational level, (b) to improve employees working behavior by connecting different experiences, increasing awareness of problematic situations or processes, and providing an environment for simulating difficult situations, and (c) to foster the entire reflection process about these issues. Future work should concentrate on improved processes for introducing technology support and fostering its adoption by providing a socio-technical framework with a holistic approach to reflective learning at work. This includes apps on the technical side as well as human support, e.g. in terms of introduction sessions etc. Also, the combined usage of MIRROR apps is a very promising approach for further developments of reflective learning at work in order to foster reflection processes in a broader and more integrative way.

5 Appendix

5.1 Appendix A.5.1: Reflection related scale and items

5.1.1 Short Reflection Scale (SRS)

Core Question Short **R**eflection Scale (CR)

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 (shaded): CR1, 3, 4, 5, 10

Subscale team reflection: CR 2, 6, 7, 8, 9

5.1.2 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 reflect about my work.	☐	☐	☐	☐	☐
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.	☐	☐	☐	☐	☐

5.2 Appendix A.5.2: Work behaviour and other work-related criteria

Note. WK02 and WK14 have not been used in any of the analysed evaluation studies.

ID	Question	strongly disagree	disagree	slightly disagree	neutral	slightly agree	agree	strongly agree
WK01	I used my learning on the job	☐	☐	☐	☐	☐	☐	☐
WK02	The app helped me to improve my work experience.	☐	☐	☐	☐	☐	☐	☐
WK03	The app increased my work satisfaction.	☐	☐	☐	☐	☐	☐	☐
WK04	The app helped me to improve my [work performance]	☐	☐	☐	☐	☐	☐	☐
WK05	I kept up my change of behaviour	☐	☐	☐	☐	☐	☐	☐
WK06	The app helped to improve my performance	☐	☐	☐	☐	☐	☐	☐
WK07	The app helped to improve team performance	☐	☐	☐	☐	☐	☐	☐
WK08	The app helped me save time	☐	☐	☐	☐	☐	☐	☐
WK09	The app helped me to focus on my work tasks	☐	☐	☐	☐	☐	☐	☐
WK10	The app helped me to satisfy my customers faster	☐	☐	☐	☐	☐	☐	☐
WK11	The app helped me to tackle difficult work situations	☐	☐	☐	☐	☐	☐	☐

Other work-related critera

ID	Question	strongly disagree	disagree	slightly disagree	neutral	slightly agree	agree	strongly agree
WK12	Using the app made me more confident that I can succeed in my work-tasks.	☐	☐	☐	☐	☐	☐	☐
WK13	Using the app supported me to master my work-tasks.	☐	☐	☐	☐	☐	☐	☐
WK14	The app improved my work satisfaction.	☐	☐	☐	☐	☐	☐	☐

5.3 Appendix A.5.2: Coding scheme for reflective elements

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 had 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 …”)

References

1. *D1.1 Specification of Research Methodology and Research Tooling*
2. *D1.2 Report on User Studies*
3. *D1.4 Model of Computer Supported Reflective Learning - version 1*
4. *D1.5 Specification of Evaluation Methodology and Research Tooling*
5. *D1.6 Model of Computer Supported Reflective Learning - version 2*
6. *D4.2 Individual reflection Apps - version 1*
7. *D4.4 Individual reflection Apps - version 3*
8. *D6.2 Prototypes of Annotation and Scaffolding - version 1*
9. *D6.3 Enhanced prototypes of Annotation and Scaffolding (version 2); prototype for synergizing - version 1*
10. *D 6.4 Prototypes of annotation and scaffolding in version 3, prototypes for synergizing in version 2*
11. *D8.2 Prototype for organisational learning intelligence - version 1*
12. *D10.1 Initialization and preparation of test beds for use and evaluation of MIRROR tools and methods*
13. *D10.2 Formative evaluation of MIRROR Appsphere usage and effectiveness at test beds*
14. *D10.3 Summative Evaluation of MIRROR Appsphere usage and effectiveness at test beds*
15. *Prilla, M., Pammer, V., & Balzert, S.: The Push and Pull of Reflection in Workplace Learning: Designing to Support Transitions between Individual, Collaborative and Organisational Learning. EC-TEL 2012: 278-291*