Iterative Augmentation of a Medical Training Simulator: Effects of Affective Metacognitive Scaffolding

Gudrun Wesiak\textsuperscript{a,2}, Christina M. Steiner\textsuperscript{a}, Adam Moore\textsuperscript{b}, Declan Dagger\textsuperscript{c}, Gordon Power\textsuperscript{c}, Marcel Berthold\textsuperscript{a}, Dietrich Albert\textsuperscript{a}, Owen Conlan\textsuperscript{b}

\textsuperscript{a}Knowledge Technologies Institute, Graz University of Technology, 8010 Graz, Austria
\textsuperscript{b}KDEG, School of Computer Science and Statistics, Trinity College, Dublin 2, Ireland
\textsuperscript{c}EmpowerTheUser Ltd., Trinity Technology & Enterprise Campus, The Tower, Dublin 2, Ireland

Abstract
Experiential training simulators are gaining increasing popularity for job-related training due to their potential to engage and motivate adult learners. They are designed to provide learning experiences that are directly connected to users’ work environments and support self-regulated learning. Nevertheless, learners often fail to transfer the knowledge gained in the simulated environment to real-world contexts. The EU-funded ImREAL project aimed to bridge that gap by developing a suite of intelligent services designed to enhance existing training simulators. This paper presents work that was a subset of this research project, reporting the iterative development and evaluation of a scaffolding service, which was integrated into a simulator for training medical students to perform diagnostic interviews. The study comprises three evaluation phases, comparing the pure simulator to a first version with metacognitive scaffolding and then to a final version with affective metacognitive scaffolding and enriched user modelling. The scaffolding service provides the learner with metacognitive prompts; affective elements are realized by an integrated affect reporting tool and affective prompts. Using a mixed-method approach by analysing questionnaires (\(N = 106\)) and log-data (\(N = 426\)), the effects of the services were investigated with respect to real-world relevance, self-regulated learning support, learning experience, and integration. Despite some limitations, the outcomes of this study demonstrate the usefulness of affective metacognitive scaffolding in the context of experiential training simulators; significant post-simulation increases in perceived relevance of the simulator, reflective note-taking, overall motivation, and feeling of success could be identified. Perceived usability and flow of the simulation increased, whereas overall workload and frustration decreased. However, low response rates to specific functions of the simulation point to a need to further investigate how to raise users’ awareness and understanding of the provided tools, to encourage interaction with the services, and to better convey the benefits of using them. Thus, future challenges concern not so much technological developments for personalizing learning experiences, but rather new ways to change user attitudes towards an open approach to learning systems that enables them to benefit from all offered features.

Keywords: adult learning; architectures for educational technology system; evaluation of CAL systems; simulations; teaching/learning strategies; metacognition;

\textsuperscript{1} Abbreviations: (A)MSS – (Affective) Metacognitive Scaffolding Service, BL - Baseline, ETU – EmpowerTheUser, SAM - Smiley Affect Measurement Technology, SRL - Self-Regulated Learning, U-Sem – User modelling infrastructure for the Social Web, UT - User Trial

\textsuperscript{2}Corresponding author. Tel.: +43 316 873-30634.
E-mail addresses: \{gudrun.wesiak,christina.steiner,dietrich.albert\}@tugraz.at, \{adam.moore,owen.conlan\}@scss.tcd.ie, \{declan.dagger,gordon.power\}@empowertheuser.com
1. Introduction

In today’s knowledge society self-regulated learning, job-related training, and life-long learning are gaining increasing importance. The demands of 21st century education require learners to take responsibility for planning, monitoring, and regulating their learning. Since formal classroom settings and human instructors are in many cases overly cost- and time-intensive, there is also a shift from traditional to technology-enhanced learning (TEL) environments (e.g. Rogers, 2000; Mayes, Morrison, Mellar, Bullen, & Oliver, 2009). TEL requires new skills (e.g. critical thinking, flexibility, communication, or information and ICT literacy) but also offers opportunities to the learner, such as personalization of the learning environment and material, independence of time and place, or self-regulation of learning content, timing, and goal-setting (e.g. Narciss, Proske, & Koerndle, 2007). Learners’ previous knowledge, experiences, and interests can be accounted for and situations that are closely related to job practice can be designed. There is already a wide variety of adaptive learning systems available, which automatically tailor to learners’ needs and provide highly motivating learning experiences (Brusilovsky & Peylo, 2003; De Bra et al., 2013). Experiential training simulators, as one kind of TEL application, are gaining increasing popularity and importance for adult training because of their high potential to engage and motivate learners by creating realistic contexts for practicing job-relevant skills (Thalheimer, 2009; Swartout, 2010). A simulation providing an appropriate environment for learning will enhance knowledge transfer and future retrieval of skills in real-world settings. Simulations have a high motivational potential and may induce flow, i.e. a positively perceived experience and state of immersion in an activity (Csikszentmihalyi, 1990). In this sense, experiential training simulators are designed to meet the challenges of adult learning, which is often characterised as self-directed, experience-based, goal-oriented, intrinsically motivated, and relevancy-driven (Knowles, 1984). Simulations enable learners to acquire, apply and practice their knowledge and skills in safe and realistic contexts, which is particularly important in healthcare and medical education (e.g. Bloice, Simonic, & Holzinger, 2013; Chakravarthy, ter Haar, Bhat, McCoy, Denmark, & Lotfipou, 2011; Galloway, 2009; Holzinger, Kickmeier-Rust, Wasertheurer, & Hessinger, 2009).

The EU-funded ImREAL project carried out research and development in the field of experiential virtual training. The project started from the assumption that with existing training simulators there is often still a gap between the simulated environments and the real-world experiences (Dimitrova, 2013; Hetzner & Pannese, 2013). Simulations cannot perfectly replicate real-world situations and necessarily only represent a part of reality (e.g. Baeyer & Sommer, 2000; Bishop, 2003; Pontonnier, Dumont, Samani, Madeleine, & Badawi, in press). Consequently, learners may fail in transferring the newly acquired skills from the simulation to their job, thus resulting in a so-called knowing-doing gap (Galarneau, 2005; Pfeffer & Sutton, 2000). ImREAL investigated ways to bridge the gap between simulated and real-world tasks by developing a suite of intelligent services to effectively align learning experiences made in the simulated environment with the real-world job practice, where the acquired skills are to be deployed. The target domain was interpersonal communication and the key pedagogical concepts for the ImREAL framework were adult learning (Knowles, 1984) and self-regulated learning (SRL) (Zimmerman, 2002), referring to learning processes under the responsibility and control of adult learners. Taking into account andragogic principles and adapting the cyclic SRL model of Zimmerman (2002), a framework for adult SRL was defined by linking simulated and real-world experiences. The model extends the three phases of forethought, performance, and self-reflection to integrate experiences made in the virtual and in the real world, and to incorporate peer experiences (Hetzner, Steiner, Dimitrova, Brna, & Conlan, 2011). Within the project, a pragmatic approach was pursued by augmenting existing experiential training simulators with the developed services. Research and

3 http://www.imreal-project.eu/
development on simulator augmentation covered three strands, namely real-world activity modelling, enriched user modelling, and affective metacognitive scaffolding.

This paper focuses on the affective metacognitive scaffolding service aiming at enhancing SRL skills and metacognition. This service provides tailored support during the learning process (i.e. scaffolding) by prompting reflection on learning and thus, stimulating SRL activities in the performance phase and fostering metacognition in terms of regulation of cognition. We report on the service integration into a medical interview training simulator, show how the simulator and augmentation have grown and become more mature over several development phases, and present a comprehensive study covering three empirical evaluations at different stages of this maturation process. Whereas in previous papers we focused on individual evaluations conducted during and right after the training in the simulator (Berthold et al., 2012; Wesiak et al., 2013), the present work gives a long-term perspective on the successive maturation of the service and simulator extension, and also reports some tentative results obtained after real-world experiences. The next two sections give a short overview on related work from the fields of SRL (Section 1.1) and scaffolding in TEL (Section 1.2). This is followed by a description of the experiential training simulator, the EmpowerTheUser (ETU) simulator, which was augmented by the ImREAL services (Section 2). Section 3 outlines the research questions and Section 4 gives a detailed explanation of the method of our study. The results derived from the three evaluations of our study are reported in Section 5, an overall discussion and outlook to future research (Section 6) concludes this paper.

1.1. Self-Regulated Learning (SRL)

SRL refers to cognitive, metacognitive, and affective/motivational processes that are regulated by the learner throughout a learning experience. The learner’s behaviour is goal-oriented and the applied SRL strategies describe the ways in which individuals control and direct their learning (Sitzmann & Ely, 2011, Zimmermann, 2002; Pintrich, 1999). Core processes of SRL comprise planning, knowledge acquisition, monitoring and regulation, strategy development, reflection, as well as motivational processes and motivational beliefs (Azevedo, 2005; Zimmermann, 2002). SRL itself is usually described as a cyclic process (e.g. Puustinen & Pulkkinen, 2001). Zimmermann (2002), for example, models self-regulation as a cycle made up of three phases, namely forethought (with task-analysis and self-motivational beliefs), performance (with self-control and self-observation), and self-reflection (with self-judgement and self-reaction).

A high quality and quantity of self-regulatory and metacognitive processes go hand in hand with better learning performance (at least in the long run) and achievements (Zimmermann, 2002; Schraw & Dennison, 1994). Based on a meta-analysis of 369 research reports (with 430 independent samples) of SRL in work-related training and adult education, Sitzmann and Ely (2011) suggest a framework that comprises nine fundamental self-regulation constructs as predictors for learning. Included are goal level and self-efficacy as strongest predictors, but also metacognitive strategies (with planning and monitoring), attention, time management, environmental structuring, motivation, effort, and attributions. The acquisition and use of SRL strategies can influence performance, if the strategies are applied more frequently also in subsequent learning sessions and independent practice (Bannert & Reimann, 2012; Zumbrunn, Tadlock, & Roberts, 2011). On a short-term basis, SRL-strategies are related to aspects of learning experience but not necessarily to measurable increases in performance.

It has been shown that many learners – children as well as adults – have difficulties in executing self-regulated processes, especially when learning complex topics, and that there is a need to support learners in developing and employing SRL strategies (Bannert & Reimann, 2012; Kravcik & Klamma, 2012).
Why is it so important to teach and support SRL? As Sitzmann and Ely (2011) outline, self-regulation is not only an important aspect of learning, which enables people to succeed in higher education, but it is also relevant for effectively functioning in our work and personal lives. Especially for work-related learning, SRL can enhance the motivation to take part in learning and knowledge building activities as well as in reflection about learning (Siadaty et al., 2012). With respect to computer-based learning environments, Greene, Moos and Azevedo (2011) point out that it is often due to a deficit in self-efficacy, knowledge, or motivation that students cannot benefit from TELs. Also Narciss et al. (2007) state that TEL environments place additional demands on the learner, which calls for direct investigations of SRL and metacognition in TEL environments. In this way, TEL environments should be designed to provide tools and opportunities for assisting and facilitating metacognitive skills (Azevedo, 2005), but most learners need additional help and guidance (Bannert 2007) to perform well in such environments. The support provided should not lead to a cognitive overload, in other words the design should aim at an optimal level of complexity and challenge for the student. In TEL environments this concretely means that the learner should not be overwhelmed by the number of interactive information elements to be processed (VanGog, Pass, & Sweller, 2010; Sweller, 1988).

Different types of instructional interventions are possible to support self-regulated learning – direct instruction of SRL strategies and indirect support guiding users, by providing scaffolding prompts to engage in specific SRL activities, such as setting goals, planning their tasks, monitoring and reflecting about their learning and achievements (Bannert & Reimann, 2012; Narciss et al., 2007). When implementing such SRL and metacognitive support, for example by augmenting existing learning technologies with a metacognitive scaffolding service, it is crucial to examine its effects on learning experience and SRL, to inform further design and development.

1.2 Scaffolding in Computer-based Learning Environments

Scaffolding is an important part of the educational process, supporting learners in their acquisition of knowledge and in developing learning skills. It refers to assistance given to learners as needed, and is characterised by fading this assistance with learners’ increasing knowledge and competence. Scaffolding has been a major topic of research since the pioneering work of Vygotsky (e.g. 1978) and the key work of Bruner, Wood, and colleagues (e.g. Bruner 1966, Wood, Bruner, & Ross, 1976) in traditional learning settings. Meanwhile, scaffolding has also become a key research topic in the context of modern learning arrangements. Work on the use of scaffolding in computer-based learning environments has been extensive (e.g. Azevedo, 2005; Proske, Narciss, & McNamara, 2012). Originally, the emphasis was on cognitive scaffolding, which has many forms (Clark, 2002). In the last ten years there has been a move towards research in metacognitive scaffolding (e.g. Van de Pol, Volman, & Beishuizen, 2010; Dinsmore, Alexander, & Loughlin, 2008; Greene & Azevedo, 2010). In this context, the use of metacognitive scaffolding in adaptive learning environments (e.g. Roll, Aleven, McLaren, & Koedinger, 2011), as well as scaffolding for designing and composing personal learning environments (e.g. Nussbaumer, Dahrendorf, Schmitz, Kravcik, Berthold, & Albert, in press) are also being researched.

Different forms of metacognitive scaffolding have been investigated in recent research. Molenaar, van Boxtel, and Sleegers (2011) investigated different types of problematizing scaffolds and structuring scaffolds and their effects on metacognitive knowledge and learning outcomes at group and individual level. Bannert and Mengelkamp (2013) provide an overview of research results on three types of metacognition support (reflection prompts, metacognitive prompts, training & metacognitive prompts), and could show that all of them have positive effects. Steiner, Kickmeier-Rust, Mattheiss, Göbel, &
Albert (2012) proposed a menu of adaptive metacognitive interventions, which were classified according to the targeted phase of the SRL cycle and the type of metacognitive action indicated.

Affective and motivational aspects need to also be taken into account to achieve an integrated view of scaffolding the SRL process (e.g. Efklides, 2011). Motivational aspects refer to attribution of success or failure, self-efficacy, and intrinsic interest (e.g. Zimmerman, 2002). Relevant aspects of affect consist of activity-related emotions (e.g. interest and boredom), outcome-related emotions (anger, pride etc.), or also metacognitive feelings (focused on cognition – feeling of knowing). In TEL Porayska-Pomsta and Pain (2004) explored affective and cognitive scaffolding through a form of face theory (the affective scaffolding also included an element of motivational scaffolding). Steiner et al. (2012) provided motivational interventions, based on non-invasive assessment of domain-skills and motivation, to enhance and retain learners’ motivation and engagement in game-based learning. Boyer, Philips, Wallis, Vouk, and Lester (2008) examined the balance between motivational and cognitive scaffolding through tutorial dialogue and found evidence that cognitive scaffolding supported learning gains while motivational scaffolding supported an increase in self-efficacy.

A critical aspect when providing scaffolding prompts is that a disruption of the natural flow of learning should be avoided. This is particularly relevant with immersive technologies like educational games or training simulators. If feedback or scaffolding interventions in such learning environments force the learner to stop the game or simulation experience to do something else, flow is interrupted and thus, engagement, immersion, and motivation could be compromised (Dunwell, de Freitas, & Jarvis, 2011; Van Eck, 2006). As a consequence, supporting prompts need to be strongly embedded in the simulation and learning experience, such that disruption of flow is minimised.

2. **EmpowerTheUser (ETU) Training Simulator**

The experiential training simulator used as a test bed in the presented study, which was augmented with ImREAL services, has been developed by EmpowerTheUser (ETU), a Dublin-based SME. It is an innovative RolePlay simulation platform that empowers training experts to develop easy to use tools for immersive simulations. The platform, which consists of three core tools (a Simulation Development Tool, an adaptive RolePlay Simulator, and an Analytics Dashboard), is actively used in training areas such as sales, customer services training, leadership, management, and clinical interviewing. ImREAL services have been integrated and evaluated with the clinical interviewing simulation. It provides two scenarios for training psychiatric diagnosis interview skills, mania and depression (Fig. 1), in which students watch patient videos and select appropriate statements (doctor’s reactions) out of different sets of alternatives. The statements are mapped to different steps in the interview process (e.g. introduction and negotiation of the agenda, eliciting information, or outlining a management plan), as well as to different skills necessary to perform a good interview (e.g. communication skills, demonstrating empathy, elicit information, summarise, or transition).
For each scenario, the simulator supports two running modes: assess and practice mode. The assess mode scores all learner decisions and summarizes them in a detailed performance report. This simulator mode starts with the prediction of one’s own interviewing scores for the interview steps (process scores) and required skills by means of a slider (0-100%) indicating the predicted success in the simulation. At the end of a training session the self-predicted scores are compared to the objective dialogue scores calculated by the simulator. The scoring constructs are derived from the skills defined in the Calgary-Cambridge model for medical interviewing (Kurtz & Silverman, 1996). The constructs are then loaded in the simulation model and each decision path loads a particular weighting onto the respective skill constructs that are represented by a particular decision. All weightings across all paths taken by the learner during an attempt are then computed to produce the learners’ performance score. During the simulation users have the possibility to use a built-in note-taking tool to record their reflections or take general notes (see Fig. 2(a)). Thus, with its assess mode the platform serves not only as a training tool, but also as a psychometric profiling, behavioural measurement, and skill assessment tool. In the practice mode learners have the possibility to explore the scenarios and get learning interventions like feedback on the decision path given by a virtual coach. Additionally, decision scores are overtly presented together with the selected statements and learners have the possibility to go back, replay their interactions, or view their previous notes in the learning log.

For each scenario and in both running modes the simulator automatically tracks a variety of user data including number of attempts (frequency of use), training times, predicted scores, ETU-scores (overall performance scores, process scores per interview step, competence scores given per skill, and dialogue scores given per selected statement), and the text-entries made via the note-taking tool.

2.1. ETU Simulator Augmentation: Phase 1

In Phase 1 of the ImREAL project, the ETU simulator was augmented by metacognitive scaffolding. The overarching goal of this first phase was an iterative development of a coach-like service that supports users in their metacognitive and SRL strategies and thus positively influences their learning experience, motivation and long-term performance. Thus, the metacognitive scaffolding service (MSS) was developed as an analogue of a coach or mentor who sits alongside the simulator and provides scaffolding. The MSS was included in the practice mode of the simulation and supplemented the note-taking tool, both of which are depicted in Fig. 2.
The MSS was provided using calls to a RESTful service developed as part of the ImREAL project and built on work developing the ETTHOS model (Macarthur & Conlan, 2009). The scaffolds or thinking prompts presented 35 items from the Metacognitive Awareness Inventory, Regulation of Cognition scales (MAI-ROC; Shraw & Dennison, 1994), which were assigned to the specific situations occurring during the course of the interview. Thereby, instructional designers matched situations to the 5 key factors of the MAI-ROC (planning, information management strategies, comprehension monitoring, debugging strategies, and evaluation). Each factor consisted of a number of items or thinking prompts. The text of the thinking prompt item was phrased in order to elicit a yes/no response. In addition, a reflection prompt was presented with each item to trigger reflection in the open reflection section. Items to which the learner already entered a reflection were not redisplayed. Thus the MSS aimed at directly matching the learners’ cognitive activity to appropriate metacognitive support.

2.2. ETU Simulator Augmentation: Phase 2

In Phase 2 of ImREAL, an affective element was added to the MSS, turning it into the affective metacognitive scaffolding service (AMSS). To obtain information on a user’s affective state, the explicit Smiley Affect Measurement technology (Ateles-SAM, previously referred to as Smiley-Based Affect Indicator - SBAI; Moore & Ascolese, 2011) was inserted into the simulation’s note-taking tool. The user modelling service (U-Sem; see below) was utilised to provide additional information gathered from tracking users’ digital traces. In order to prompt learners to think about their learning, triggers were displayed during the training in the practice mode. Fig. 3 shows examples for (a) triggers and (b) the thinking prompts shown after the learner selected a trigger. The thinking prompts relating to the MAI-ROC items suggested learners to reflect on how their learning was going, whereas the affective prompts (Fig. 3c) asked learners to reflect on how they were feeling.
In an evolution to Phase 1, thinking prompts were displayed according to a refined map created by instructional designers. This underlying cognitive activity model was a weighted list, which mapped the 35 individual thinking prompts (i.e. items from the MAI-ROC) to specific situations in an interview. For each prompt the instructional designer set a weight that allocates the likelihood of the intervention being displayed within the set of all interventions for that section. In other words, the weights indicated how suitable a prompt was for a certain situation. These initial weights were changed according to learner’s affective state, if available. For example, a prompt considered as encouraging might have a higher weighting for learners with negative affect indication.

Generally, the AMSS is part of the so-called Ateles framework, which has been designed from the outset as a tool external to the learning environment, working alongside it. This means that it is not specifically designed for the ETU simulator but rather as flexible, complementary technology that can be incorporated within a wide range of e-learning platforms. As a consequence of the technology’s flexibility, not all capabilities of the scaffolding service may always be implemented fully within a given experiential training environment. Reasons are partly andragogical, partly technological, for instance concerning requirements of the instructional designer on how the scaffolding interventions are integrated within the learning experience or how to interface the scaffolding service with simulation framework.

U-Sem. The enriched user modelling framework U-Sem is a strand of research conducted within the ImREAL project which aimed at creating services for enriching user profiles by means of exploiting digital traces (Abel, Celik, Hauff, Hollink, & Houben, 2011). It focuses on personalization based on previous experiences and interests in a certain domain, demographic information (e.g. location or gender) and sentiment analysis (Hauff & Houben, 2011, 2012; Tao, Abel, Gao, & Houben, 2011; Gao, Abel, Houben, & Yu, 2012). The AMSS interacts with U-Sem to update the user profile information in regard to learners’ affective approach. Thereby the Twitter feed of a learner is used for sentiment analysis techniques, by which patterns of expression are derived. The affect stereotypes are categorized as positive, neutral, or negative and are used to weight information during the training in the simulator. For instance, the expression of a negative affective state during the simulation weights more for learners with a positive affect stereotype than for those who typically express themselves in a rather negative manner. The U-Sem service was implemented during Phase 2 of this study, but from the 152 participants not one provided their Twitter ID in the simulation. Thus, an evaluation of the sentiment analysis of tweet streams to derive affect information was not possible. From a comprehensive cohort survey, privacy issues emerged as the main reason for the reluctance to share social network IDs (see Moore et al., 2013 for a detailed report of these results).
3. Research Questions

The scaffolding service was designed to support and guide the user in the process of learning by means of thinking prompts, which are triggered by a continuously updated learner profile. The simulator augmentation was performed in two phases, starting with the implementation of metacognitive scaffolding in Phase 1, which further matured to affective metacognitive scaffolding in Phase 2. For the development of the service several challenges had to be considered: (a) training simulators need to be relevant to the real-world context to provide effective training for adults (Hetzner et al., 2011), (b) any kind of SRL support needs to be embedded in the simulation and the learning episode in an appropriate way, i.e., under consideration of flow, cognitive load, and their effects on motivation (Azevedo & Hadwin, 2005; Sweller, 1998; Pintrich, 2000). More generally, motivation as part of SRL and (c) learning experience needed to be taken into account, because high levels in both are important prerequisites for a long term learning benefit of the training in the simulator. This benefit does not necessarily translate into a short-term impact on performance in the simulator, although an eventual impact on learning effectiveness is supposed. In fact, affective metacognitive scaffolding primarily aims at affecting the way users think and learn, and their emotional state. Prompting learners to reflect on how they learn is assumed to activate and broaden SRL skills (Bannert & Reimann, 2012). These are acquired and performed in a gradual, developmental process, which will translate into effects on learning performance after sufficient practice (e.g. Ertmer & Newby, 1996). The possibilities of proving significant changes in learners’ general learning approach and learning performance in the case of short-term training times are consequently limited. However, in order to prove effectiveness and to support the transfer from the virtual to the real world, the augmented simulator should enhance motivation, learning experience, and real-world relevance. Effects should grow with the maturation of the simulator augmentation that is compared to the baseline (pure simulator); the affective metacognitive scaffolding should have even stronger effects than the metacognitive scaffolding only. In line with research results on SRL (Bannert & Reimann, 2012; Dignath, Buettner, & Langfeldt, 2008; Harris, Graham, Mason, & Sadler, 2002) it can then be expected that in the long run increased motivation, learning experience, and real-world relevance would lead to more frequent usage of the simulator and, as a consequence, enhance long-term performance. From this, the following main research questions were derived:

a) Is the training in the simulator relevant for real life experiences?

b) Does the scaffolding service support SRL?

c) Does the augmented simulator lead to an enhanced learning experience?

Furthermore, we used the stepwise evaluation to investigate whether the affective component led to an additional improvement in SRL and learning experience. Since the ImREAL services were implemented as external tools, we additionally investigated whether the services are integrated without negatively affecting usability. The measures used to evaluate relevance, SRL-support, learning experience, and usability are described in the next section.

4. Methodology

The evaluation of the ETU system was carried out in three steps, starting with a baseline evaluation (BL) of the pure simulator, which was followed by two User Trials (UT1 and UT2) corresponding to the two developmental stages described in Sections 2.1 and 2.2. The study design included independent samples, thus there were no training effects from multiple use of the simulator (in all samples more than 96% of the participants had never used the simulator before). It has to be pointed out, though, that the study was

4 The following statistical abbreviations are used: $M =$ Mean, $Md =$ Median, $SD =$ Standard Deviation, $SE =$ Standard Error
carried out in a real-world context using a commercial simulator that was enhanced during the course of the project. Therefore, the setting of the study could not meet the methodological requirements of an experiment (e.g. control of all other factors that might influence students’ learning experience). Diversity between the samples was minimised, as all participants were enrolled in their third year medical curriculum at Trinity College, Dublin and over 97% already had experience with conducting clinical interviews in real-world settings. The three samples also did not differ significantly regarding sex ratio and age. Generally, the use of the simulator or completion of the questionnaires was not mandatory for the students, but participation rates for the three periods reached 92%, 87%, and 97% (of 142, 165, and 157 students respectively, which represented the total classes). We have no available data on why some students did not participate but one could speculate that for a given year, a number of students might drop out, be out sick etc. Response rates for the questionnaires were much lower, especially the questionnaires presented after real-world interviews, which have been completed by only 8% of the participants across all studies. Whereas the materials for this post-interview part of the study are described in this section together with the remaining materials used, the results based on this small and self-selected sample are reported separately at the end of the results section.

**Procedure.** Students were offered to use the ETU simulator as an optional instrument to train their clinical interviewing skills as part of their curriculum in the spring semesters (2011, 2012, and 2013) of their third year in medical school. Log-data (ETU simulator and [A]MSS data) were collected during the simulation, questionnaire data right after the simulation (post-simulation-survey) and after students had conducted real-world interviews with patients (post-interview-survey). Additionally, in UT2 participants completed a cohort-characterization survey before entering the simulation. Links to this survey were sent out via e-mail. During the training phase students could use the simulator as long and as often as they wished and deal with the two psychiatric scenarios. In BL and in UT2 both scenarios were presented together, that is participants could choose which scenario they entered first, whereas in UT1 the mania scenario was activated only two weeks after the depression scenario. In all three studies, both scenarios could be entered in assess mode as well as practice mode. The scaffolding service was available in practice mode only and the simulation needed to be accessed via the assess mode before the practice mode could be entered. At the end of the training, an integrated link directed students to the online post-simulation-survey, asking for feedback on the simulation and on students’ learning experience. Simulator usage was possible without answering the questionnaires. The times between simulation and real-world interviews varied for each student but all carried out the interviews within the same semester. Links to the post-interview-survey were sent out via e-mail after all students had conducted their patient interviews.

The following subsections report the methods for each of the three evaluations. For the latter two only modified or additional materials are described. In essence, a major part of the data collection instruments have been applied over all three evaluation phases in order to enable statistical comparisons; some evaluation instruments have been added in successive trials in order to refine the evaluation methodology and to account for additional research foci introduced in the respective period. Besides the reported surveys, participants of UT2 completed questionnaires regarding their personality, learning styles, cultural background and social network behaviour (cohort-characterization-survey). Since these are not relevant in the context of this paper, we refer the interested reader to Moore et al. (2013), Steiner et al. (2013a,b), and Wesiak et al. (2013).

### 4.1. Baseline Evaluation (BL): Pure ETU Simulator

**Participants.** 131 medical students from Trinity College, Dublin used the pure simulator during the spring semester 2011. Ratings regarding their prior experience (a 6-pt. scales from 1-never to 6-very often) indicate that they had not used the ETU simulator before ($M = 1.05, SD = .28, Md = 1$), whereas
they did have experience with clinical interviews ($M = 4.45$, $SD = 1.06$, $Md = 5$). From the overall sample, 76 completed the questionnaire containing demographic questions. Their average age was 23.04 years ($SD = 2.85$), 45 were female (59%). Overall, 27 students completed all subsections of the post-simulation-survey, and 7 completed the post-interview-survey.

Materials and Measures. General log-data collected during the training with the simulator included the number of participants entering each scenario and the time spent per scenario. In alignment with the research questions outlined in Section 3, the following instruments were used: Real-world relevance was measured by 2 survey items asking (a) how relevant the training experience with the ETU simulator was for real clinical interviews with patients (4pt.-rating scale from 1 - not relevant to 4 - very relevant) and (b) how well prepared students felt after their training in the simulator (in per cent). Relevance was included in both the post-simulation and the post-interview survey, preparedness only in the latter. In order to assess differences in self-regulated learning we employed the questionnaire on SRL (QSRL; Fill Giordano, Litzenberger, & Berthold, 2010), an instrument to assess learners’ usual learning approach. The QSRL consists of nine subscales covering cognitive and metacognitive strategies, time and resource management. Answers are given on an analogue scale from 0 to 100 to indicate agreement. Since there were no normative values available for our student population, scores were interpreted with respect to differences between samples, not regarding students’ absolute SRL-skills. The QSRL was presented after the simulation. As further indicators for SRL effort was measured in terms of time spent in the training simulator, state motivation was inquired via a specific question posed after the simulation and after the real-world interview (4pt. rating scale from 1 - not motivated at all to 4 - very motivated), and subjective learning effectiveness was evaluated via the NASA TLX subscale “performance” (Hart & Staveland, 1988). This item, which was included in the post-simulation as well as post-interview survey, asked how successful participants were in accomplishing what they were asked to do. In addition, the text entries logged by the simulator were analysed with respect to content type (Moore et al., 2012), whereby we differentiated between the following four content types: (1) positional notes that concern time, date, simulator, etc. (e.g. “Rapport and initial inquiry have been completed. Must now concentrate on moving the interview forward”); (2) technical notes addressing the system, questions or underlying model (e.g. “Great concept. Has a lot of potential, will be much better when more points to consider and points of information are added to the overview that accurately reflect performance.”); (3) patient notes, i.e. observations of the patient (e.g. “Px. unable to quantify length of period she has been like this for.”); (4) reflections, i.e. reflective text (e.g. “It is quite important to deal with the patient with empathy to make them comfortable and also to try and illicit the cause of the bout of depression, in this case. I felt I did not do too bad as far as expressing empathy is concerned, allowing the patient to open up and try and formulate a management plan. More experience in history taking will be the key.”). Entries with text concerning more than one content type were coded for all relevant types. In the baseline evaluation, learning experience was assessed in terms of perceived workload and costs that the learning experience imposes on the user via the NASA-TLX. The six subscales indicating mental, physical, temporal demand, performance, effort and frustration on a scale with 100 points range were presented after the training in the simulator. For an optimal workload, users should be challenged but not overloaded with information. For evaluating integration, the System Usability Scale (SUS, Brooke, 1996; see Holzinger et al., 2011, for an application in the medical area) was employed in the post-simulation-survey to investigate the general usability of the ETU-simulator. An overall SUS-score with a range from 0-100 was calculated from 10 items on various usability aspects (with 5-pt. rating scale from 1 - strongly disagree to 5 - strongly agree).

4.2. User Trial 1 (UT1): ETU Simulator with MSS
Participants: From the 143 Trinity College, Dublin medical students using the simulator during the spring semester 2012, 39 also completed the post-simulation survey. They were on average 22.7 years old ($SD = 2.4$) and 54.8% were male. Nobody had used the ETU simulator before, 97.4% indicated that they had conducted real-world clinical interviews before, but only 15.4% with psychiatric patients. Eight students responded to the post-interview survey, seven completed it.

Materials and Measures. Data collection followed the procedure of the BL study, but was adjusted and extended to account for the newly implemented service. The simulator with MSS offers two training modes, assess and practice, thus log-data were collected separately for each mode. Regarding real-world relevance, both items (relevance and preparedness) were included in the post-simulation as well as the post-interview survey. For SRL, measurement of the general learning approach via QSRL, and assessment of effort, and subjective learning effectiveness followed the baseline design. Text entries were collected from the simulator’s notepad and the open reflections in MSS; content analysis followed the same four categories as used in the BL. The assessment of motivation was split up into four questions referring to learning about clinical interviews, improving interviewing skills, performing a good interview, and applying what has been learned. In the post-interview survey the first question was replaced by the motivation to use the ETU simulator again. Learning experience was captured in the post-simulation survey. Besides perceived workload (NASA-TLX) also flow and the perception of thinking prompts were assessed. Flow, which refers to the fluency and smoothness of the learning process, was measured via the Flow Short Scale (Rheinberg, Vollmeyer, & Engeser, 2003). Answers were given on 7-pt. rating scales (from 1 - strongly disagree to 7 - strongly agree). To assess the perception of thinking prompts provided by the MSS, users rated their helpfulness and appropriateness (w.r.t. time and content) via 10 items with 5-point rating scales (from 1 - not at all to 5 - very much). Integration was measured by the SUS and by three items regarding the integration of MSS in the simulator, i.e. whether the service and thinking prompts were perceived as integral part of the simulation. Responses were given on a 5-pt. agreement scale (from 1 - strongly disagree to 5 - strongly agree). Furthermore, technical flow was assessed in connection with the Flow Short Scale via three questions capturing technical flow in terms of a smooth interplay of software components.

4.3. User Trial 2 (UT2): ETU Simulator with Affective MSS and Enriched User Modelling

Participants. 152 students used the ETU simulator with integrated AMSS and augmented user modelling during the spring semester 2013. Demographic data was collected as part of a more comprehensive cohort characterization survey before the training. The 95 respondents were, on average, 22.81 years old ($SD = 3.79$) with an equal sex ratio. The post-simulation and post-interview-surveys were answered by 40 and 19 students, respectively (with 36 and 16 complete answer patterns). From the first group only one person had used the ETU simulator before, whereas all but one had previous experience with clinical interviews (25.6% with psychiatric patients).

Materials and Measures. Principally, data collection followed the procedure of UT1. The additional cohort-characterization-survey covered demography questions and a range of different questionnaires on personality, learning, and social network use. Since this paper focuses on the comparison of the three evaluation stages, the reported results are confined to basic demographic data. The post-simulation survey was adapted to account for the affective aspect of the AMSS. The Total Affective State Scales (TASS; Nicolescu et al., 2006) were used as subjective self-report on affect after the learning episode. In addition, the Ateles-SAM provided the opportunity of capturing affect self-reports during the learning activity. Furthermore, the post-interview survey was extended to gather data on specific aspects of the training in the simulator. These included five questions concerning the perception of the two different training modes (assess vs. practice) and three questions each on the use of the note-taking/reflection tools and the in-scenario affect self-report via Ateles-SAM.
5. Results

Results for the three evaluation phases are reported together in order to get a clearer picture of changes across the three stages of development. In each trial, from the participants using the simulator in the assessment mode of at least one scenario (whole sample), different subsamples used the practice mode and answered the presented questionnaires. Thus sample sizes vary throughout this section and are reported together with the results. Despite the large initial sample sizes, for some questionnaire sections the number of responses is rather low and comparisons have to be interpreted with caution. Power-analysis (using G*Power 3.0; Faul, Erdfelder, Lang, & Buchner, 2007) show that even when assuming large effect sizes (ES) and an $\alpha$-level of .05, sample sizes of at least 21 participants per group would be required to reach conventional power estimates of 0.8 (Cohen, 1992\textsuperscript{5}). As mentioned above, sample sizes for the post-interview questionnaire do not reach this criterion and are therefore reported separately at the end of this section.

5.1. Simulator Usage

Log-data was collected separately for the two simulation scenarios. Fig. 4(a) shows the usage duration in minutes of each scenario in the three studies, plus the overall time spent on the training with the simulator. The number of participants entering the scenarios (and therefore contributing to the depicted means) varied. Whereas almost all participants entered the depression scenario ($N = 131/143/140$ for BL/UT1/UT2), only subsamples of BL and UT1 used the mania scenario ($N = 118/18/152$). On the other hand, most UT2 participants entered the mania scenario first and thus spent less time with the depression scenario (see Fig. 4a). Due to these differences in usage of the two scenarios and because this work’s focus is on the maturation of the simulator and augmentation in general, for further analyses log-data are aggregated across the two scenarios. Questionnaire data are, in any case, scenario-independent.

Fig. 4 (b) shows the different usage of the practice mode and assess mode in UT1 and UT2. Practice mode was only offered after integration of the ImREAL services, thus in BL there was only one mode available. Additionally, the practice mode could only be used after the assess mode. In UT1 41% entered the practice mode, in UT2 38%. In both trials and both scenarios students also spent more time in assess than in practice mode. In UT2, for example, students spent on average 11 min ($SD = 14.2$) in the practice mode and 20 minutes ($SD = 7.3$) in the assess mode of the mania scenario, the depression interview was trained for only 4.8 ($SD = 8.9$) and 7.3 minutes ($SD = 6.33$) in the practice and assessment modes, respectively. Especially for the practice mode, the variation within the sample is very high (see $SD$s).

\textsuperscript{5} Effect size conventions after Cohen (1992) index as small/medium/large effect size: $d=0.2/0.5/0.7$, $f=0.1/0.25/0.4$, $f^2=0.02/0.15/0.35$;
Fig. 4. (a) Mean training time in minutes (with SE) per scenario across both training modes; (b) usage and mean training time of assess and practice mode in User Trials 1 and 2

5.2. Real-World Relevance and Integration

Natural requirements for the practical deployment of the (augmented) simulator are on the one hand its relevance for real-world situations, that is *do students benefit from the training with regard to real-world situations*, and on the other hand its usability (with and without ImREAL services).

Two questions regarding the *relevance of the simulation* were asked right after the training in the simulator. Fig. 5 depicts the mean ratings and scores obtained in the three studies. Comparisons among the three studies show that ratings regarding the perceived relevance increased significantly from BL to UT2 (Kruskal-Wallis $\chi^2 = 6.302, df = 2, p = .043$; U-test $z = 2.22, p = .027$), whereas ratings from UT1 do not differ from the other trials. Overall, the obtained median of 3 in all studies indicates that participants perceived the simulation as ‘relevant’ with and without ImREAL services.

![Fig. 5](image)

**Fig. 5.** Mean (and SE) post-simulation ratings and scores for the relevance and usability of the ETU simulator

The question asking how well prepared students felt for the real-world interview based on their experience with the ETU learning environment was answered on a scale from 0-100 by UT1 and UT2 participants. The bar-group in the middle of Fig. 5 shows an increase from UT1 to UT2. However, with average scores of $M=51.3$ ($SD=17.8$; $N=33$) in UT1 and 59 ($SD=16.16$; $N=37$) in UT2 the difference did not reach statistical significance ($t_{(68)}=1.9, p=.062$).

*Usability* and integration of the system, measured via the SUS, increased relative to the initial version (see Fig. 5 for exact scores; $F_{(2,167)} = 4.4, p = .014, \eta^2_p =.05$). Post-tests after Scheffé locate the difference between BL und UT2 with $p = .021$. *Technical flow*, which was evaluated in UT1 and UT2 in conjunction with the Flow-Short-Scale, also increased ($M_{UT1} = 3.75, SD = 1.16$; $M_{UT2} = 4.69, SD = 1.15$; $t_{(75)} = -3.55, p = .001$) and the *integration of metacognitive scaffolding*, i.e. the thinking prompts was seen as neutral to positive with no difference between UT1 and UT2 (mean ratings for 3 questions with 5-pt. scales range between 3.25 and 3.46). Overall, augmenting the simulator with the scaffolding service did not negatively affect the perception of usability and integration, the service was assessed as

---

6 Conventions for effect sizes $\eta^2_p$ indexed $\eta^2_p=.01/.06/.14$ as a small/medium/large effect
well integrated in the simulator and in certain aspects the augmented simulator with AMSS was even perceived better than the baseline version or, respectively, the simulator with MSS.

5.3. Self-Regulated Learning

Questionnaire on SRL. Data (N = 76/20/25 for BL/UT1/UT2) from the nine QSRL subscales revealed means ranging between 54.07 (SD = 14.4) for ‘memorising strategies’ in BL and 71.74 (SD = 10.1) for ‘elaboration strategies’ in UT2. A one-way MANOVA for independent samples revealed no significant difference among the three user trials, neither on a multivariate level ($F_{(18,222)} = .524$, $p = .945$, $\eta^2_p = .041$) nor for a single subscale (all $p > .13$, all $\eta^2_p < .035$).

Reflection Notes. Another indicator for SRL strategies is the kind of information students recorded during the learning process. Overall 1092 thinking prompts were provided during the second user trial, 2001 during UT1. The strong decrease of prompt frequency is due to an evolution of the scaffolding in which the rules for triggering prompts were refined to achieve a better timing of the prompts. Triggering rules were defined by instructional designers, i.e. subject matter experts. With each prompt a text entry field popped up for collecting users’ reflections (AMSS text entries) and students were asked about the usefulness of the last part of the simulation. From 69 yes/no responses in UT2 88.4% were positive, that is users perceived the simulation as useful. Additionally, the ETU-notepad could be used at any time to reflect or take notes. Table 1 compares the number and types of text entries from (A)MSS and ETU notepad in each of the three studies. Text entries were coded in the four categories: positional, technical, patient notes, and reflections. Entries with text concerning more than one content type were coded for all relevant types, therefore resulting in some entries having multiple codings. The number of users entering text decreased over the three studies. Considering the content of the notes, the proportion of entries that were actually reflective increased in UT2. Simultaneously the portion of all other entry types (positional, technical, and patient notes) decreased. A comparison of the distributions revealed a significant difference between the three studies ($\chi^2 = 60.14$, $df = 6$, $p < .001$ based on 817 codings), with all UT2 cells showing significant deviations from the expected values. The highest difference between observed and expected frequencies was found for the reflection notes in UT2 ($f_o = 80$, $f_e = 50$, $\chi^2 = 18.08$, $p < .001$). Besides that, only the number of observed patient notes in BL and technical entries in UT1 are higher than expected.

Table 1 Content types for entries from the AMSS and ETU note-taking tool

<table>
<thead>
<tr>
<th>Users</th>
<th>Text entries</th>
<th>Positional f (%)</th>
<th>Technical f (%)</th>
<th>Patient Notes f (%)</th>
<th>Reflections f (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>81</td>
<td>358</td>
<td>54 (15.08)</td>
<td>118 (32.96)</td>
<td>90 (25.14)</td>
</tr>
<tr>
<td>UT 1</td>
<td>58</td>
<td>135</td>
<td>23 (17.04)</td>
<td>77 (57.04)</td>
<td>22 (16.3)</td>
</tr>
<tr>
<td>UT 2</td>
<td>35</td>
<td>86</td>
<td>1 (1.16)</td>
<td>11 (12.79)</td>
<td>6 (9.68)</td>
</tr>
</tbody>
</table>

Note. f = absolute frequencies; due to the multiple coding of entries concerning more than one content type f values (and %) add up to more than the total number of text entries (more than 100%).

Effort. Time spent with the training in the simulator (as depicted in Fig. 4) differs among the three trials ($F_{(2,428)} = 28.971$, $p < .001$). Post-tests after Scheffé show that the effect is due to UT1, where training times were shortest. Considering only the practice mode of the simulator, there is also an increase from UT1 to UT2 with $t_{(93,6)} = -2.53$, $p = .013$ (in BL this mode was not offered).

Motivation. Fig. 6 shows the mean ratings obtained for four questions concerning participants’ motivation in UT1 and UT2 and an overall score (as average of the four single items) for comparison with the single question in BL. Since data are not normally distributed for the four single scales, medians and results from Mann-Whitney-U-tests are reported. Right after the training in the simulator responses
on 4-pt. scales resulted in $Md = 2$ for the BL sample ($N = 95$) and in $Md \geq 3$ for all four questions of UT1 ($N = 38$) and UT2 ($N = 39$). The overall scores ($Md = 3.5$ and $3.25$ for UT1 and UT2 respectively) reflect this rather high motivation of the participants in the two user trials with integrated ImREAL services. An ANOVA confirms this increase of overall motivation ($F_{(2,170)} = 40.7$, $p < .001$, $\eta^2_p = .324$; post-tests show differences between BL and UT1 as well as UT2, but not between UT1 and UT2). Comparisons of the two user trials regarding the four single questions reveal a higher motivation for learning more about interviewing from users of UT1 than of UT2 ($z = -2.3$, $p = .016$). Otherwise, we found no differences in motivation between the two samples.

![Mean motivation ratings (4-pt. scales) and SEs from the post-simulation phases of the evaluations](image)

**Fig. 6.** Mean motivation ratings (4-pt. scales) and SEs from the post-simulation phases of the evaluations

**Feeling of Success.** Another question with respect to SRL considered the self-evaluation of participants by asking how successful students felt in accomplishing their task (subscale ‘performance’ of NASA-TLX, see below). Mean scores (on scales from 0-100) obtained right after the training in the simulator indicate that students from UT2 ($M = 69.16$, $N = 38$) had the strongest feeling to be successful, followed by the baseline ($M = 57$, $N = 76$) and UT1 ($M = 45.17$, $N = 35$; Kruskal-Wallis $\chi^2 = 28.4$, $df = 2$, $p < .001$; for all post-hoc U-tests $p \leq .014$).

**5.4. Learning Experience**

**Workload.** Average NASA-TLX workload scores for the baseline and the two user trials are depicted in Fig. 7. Overall scores differed among the three user trials ($F_{(2,149)} = 3.81$, $p = .024$, $\eta^2_p = .05$). For UT2 ($N = 40$), it reached a level of 38.72 ($SD = 11.6$), which is significantly lower than in the baseline evaluation ($M = 44.81$, $SD = 12.01$, $N = 76$). Results from single subscales indicate that the training in the simulator leads especially to load for effort and mental demand. All other subscales have mean values below 50%. Besides the reported differences in performance (feeling of success), results from a one-way ANOVA indicate that students from UT2 felt less frustrated than participants from the two previous evaluations ($F_{(2,146)} = 8.9$, $p < .001$; $\eta^2_p = .11$; post-test $p \leq .19$).
Perception of Thinking Prompts. Ten questions asking how helpful the thinking prompts provided by the (A)MSS were perceived have been answered by 16 students of UT1 and 21 of UT2, who had used the simulator’s practice mode and received prompts. On a 5-pt. rating scale (ranging from 1 - not at all to 5 - very much) the overall score reached 3.36 (SD = .33) in UT1 and 3.55 (SD = .72) in UT2, i.e. students evaluated the prompts as being helpful with regard to content, timing, support in planning, monitoring, improving, and analysing one’s learning performance. Comparisons of the two trials revealed no significant differences between the overall scores or between scores for the individual questions (all \( p > .2 \) for unrelated samples T-tests, but the effect size for the overall score amounts to \( d = .42 \)).

Flow. Ratings (on 7-pt. scales) from the Flow Short Scale are available from 40 UT2 and 37 UT1 users. The average UT2 rating of 4.75 (SD = .79) for overall flow was significantly higher than in UT1 (\( M = 4.35, SD = .91; t(75) = -2.12, p = .038 \)). Higher ratings stand for a higher agreement concerning the fluency and smoothness of the learning process as well as for a higher involvement in the task.

Affect. The overall mean score derived from seven subscales of the TASS (\( M = 64.4, SD = 16.4 \)) shows that students were in a positive affective state regarding different dimensions like mood, motivation or thinking activity after they had finished their training in the simulator. The SAM was displayed 352 times, i.e. between 1 and 8 times per user (\( M = 2.23, SD = 1.36, Md = 2 \)). Only one learner, however, actually made use of the affect report to indicate her current emotional state, while all other learners did not (see below for a discussion of this issue).

5.5. Supplementary Results

This section reports results from the post-interview-surveys. As the number of participants answering these questionnaires was very small in all three studies (7/8/19 in BL/UT1/UT2), the interpretation of results has to be done with caution. On the one hand, the effects found in the data are based on a self-selected sample, i.e. those students who voluntarily responded to this follow-up questionnaire. On the other hand, power-analysis showed that the probabilities to discover medium effects with the given sample sizes, reach a maximum of .31 for two samples (based on unrelated t-tests) and .22 for all three samples (based on one-way ANOVAs).

In all three phases of the study, the post-interview questionnaires contained items regarding the relevance of the simulation, perceived performance and participants’ motivation. With medians of 3 in all samples, students’ ratings regarding the relevance of the learning experience with the ETU simulator
for the real-world interviews did not differ among the three evaluations. There was also no change from the post-simulation to the post-interview phase. Preparedness for the real-world interview based on the experience with the ETU learning environment and perceived performance were answered on a scale from 0-100. Whereas we found no significant difference for preparedness (M = 63.6/42.7/59.11, SD = 11.4/29.3/19.7, Md = 60/60/66 for BL/UT1/UT2), students’ feeling to be successful in accomplishing the real-world interview increased steadily over the three studies, with the highest value of M = 68.38 (SD = 15.15, N = 16) reached in UT2. A Kruskal-Wallis test revealed a significant difference between the BL (M =40, SD = 32.15, Md = 60, N = 7) and UT2 (χ² = 6.1, df = 2, p = .047; U = 22, z = -2.29, p = .022).

Motivation was assessed via four questions in UT1 and UT2, which were aggregated to an overall motivation score. This mean score could then be compared to the single score collected in the BL phase. In the post-interview phase all but one rating reached a median of 3 (on 4pt. rating scales). Average ratings are depicted in Fig. 8. As only exception, UT2 participants indicated to be very motivated (Md = 4) to perform a good interview. Mann-Whitney-U tests for the four single items and a Kruskal Wallis test for overall motivation showed no significant differences between the samples.

![Fig. 8. Mean motivation (4-pt. scales) from the post-interview phases of the evaluations](image)

In UT2, we used the post-interview survey to gather further user feedback on open issues identified from the cohort and post-simulation surveys. Five questions tried to find out learners’ attitudes towards the two different simulation modes and reasons for waiving the opportunity of exploring the simulation in the practice mode. Moreover, users’ opinion on the possibility to take reflection notes and on the smiley-based tool (SAM) for affect reporting was queried by three questions in each case.

From the 19 participants responding to the post-interview survey, 65% indicated to have used both assess and practice mode. Reasons for using only the assess mode referred for the most part to time constraints. From the students using both modes, 55% said they have learned more in the assess mode and that they would prefer the assess mode for future training. On the other hand, 60% indicated that they found the practice mode more helpful in preparing for the real-world clinical interview. Thus, there is no clear preference for one mode over the other. The questions whether students used the possibility to take (reflection) notes was affirmed by only 12.5%; they indicated that the notes helped in structuring the agenda for the interview and to follow the options. Reasons for not making use of the note-taking were for the main part, that students didn’t perceive it as necessary or useful, but also that they have not been aware of this option. Response to the SAM-items collected in the post-interview survey of UT2, revealed
that participants simply did not notice the smiley-based tool for indicating their emotions (16 out of 17 answered “No”). Its usefulness could therefore not be evaluated.

6. Discussion

In this paper we presented the iterative evaluation of an augmented training simulator for medical interviews, which involved three different student samples using the simulator in three consecutive years. The initial evaluation (BL) involved the pure simulator without additional services and served as baseline for the two augmented simulator versions. In Phase 1 the metacognitive scaffolding service (MSS) has been designed, implemented, and evaluated (UT1). The findings showed a positive effect on state motivation, positive perception of the thinking prompts, and consistently good usability. In Phase 2, the service was enhanced by refining the presentation of thinking prompts in collaboration with instructional designers and by adding an affective part and thereby turning the service into the AMSS, which was again evaluated (UT2). Due to this stepwise procedure, differences can be attributed to the two stages of development. However, in doing so, some weaknesses of the study as discussed below need to be taken into account. Data collection involved a mixed-method approach including log-data as well as questionnaire data, which together served as a basis for answering our three main research questions: does the simulator augmentation by means of the scaffolding service (a) improve the real-world relevance of the training, (b) support SRL, and (c) enhance users’ learning experience? Additionally, the added value of AMSS compared to MSS only was investigated for each of the three mentioned questions. We start the discussion by elaborating on the limitations of the evaluations, then addressing each of the research questions separately in more detail, followed by a more general conclusion and outlook to further research.

Limitations of the Study. The main weakness of the study presented in this paper is related to the number of respondents. While extensive log-data was available from all participants in the three evaluation phases, response rates to the post-simulation survey were varying and rather low, representing only a small portion of the overall sample and thus being quite self-selective. Differences in the number of respondents exist between the three evaluation phases (BL, UT1, and UT2), as well as between the questionnaires presented within one phase, since additional dropouts occurred between different survey parts. Only very low numbers of participants could be involved in a follow-up data collection after real-world experience (i.e. post-interview-survey), where feedback on relevance aspects and longer-term motivational effects were intended to be investigated. The samples in the post-interview phases were highly self-selective, nevertheless the available data has been analysed and reported as supplementary results. Although the related outcomes are only indicative they give an idea of possible tendencies on results as a basis and benchmark for future user trials that could investigate in more detail the benefits and perceptions after real-world experiences. The reasons for the small sample sizes in the post-simulation- and post-interview-surveys and potential systematic differences between the groups of responding and non-responding students remain unclear; and the self-selected data collection and its potential bias on obtained results need to be taken into account when interpreting the results. Nonetheless, it should be noted that in addition to the impact determined on questionnaire results, also positive effects of simulator augmentation and maturation could be identified on the level of log-data – in terms of longer interaction times and increased reflective note-taking. In Phase 2 of simulator augmentation the metacognitive scaffolding had been refined and complemented by an affective element, which was investigated in UT2. Students’ reluctance to make use of the services integrated in the training simulator complicated an in-depth analysis of effects on users’ affective states. Future studies need to address this question in more detail. Another drawback of the presented study is that in the different evaluations there was a difference in the availability and sequence of simulation modes and scenarios, which might have mediated or moderated the effects identified. While from a methodological
point of view keeping the simulation modes and scenarios constant and consistent over all evaluations would have been desirable, these alterations of the experimental protocol were due to pragmatic and technological reasons. These were given by the possibilities and limitations of a user trial setting integrated in a real-world curriculum and, respectively, by technical necessities and decisions made around the functionalities of the simulation framework. Even though the study presented features a number of limitations, it could demonstrate a post-simulation increase in perceived relevance, reflective actions, preparedness, and motivation with iterative simulator augmentation, without any negative effects on the usability and flow of the simulation.

**Real-world Relevance and Integration.** The enhancement of the simulation platform with ImREAL services (A)MSS (and U-Sem) was evaluated with respect to integration and usability as well as perceived relevance of the learning experience. Relevance ratings increased from baseline to UT2 and with a median of 3 on a 4pt. scale users see the training in the simulator definitely as relevant preparation for interviews with real patients. Overall, we found no changes from the post-simulation to the post-interview phase. Thus the positive results obtained right after the training remained until after the real-world experience. Therefore, we can conclude that the simulation embeds an authentic scenario relevant to the real-world application context, as called for by adult learning theory (Knowles, 1984; Swartout, 2010). Despite the design of ImREAL services as external tools to the simulator, usability ratings and technical flow increased from the first user trials, thus indicating a good integration of the ImREAL services and a smooth interplay of software components, which allows a fluent interaction with the simulation.

**SRL.** Short-term changes on learner’s self-regulation and metacognitive strategies, as measured with the QSRL, could not be observed. Considering that training times in the simulator lasted about half an hour, this result is not surprising, since changes of the general learning approach are not immediate but rather long-term processes (Pressley, 1995). Looking at special aspects of SRL, there are clear indications that the addition of thinking prompts provided by the scaffolding service was beneficial to the learners. Effort in terms of time spent with the simulation increased in the practice mode of the simulator from UT1 to UT2, which implies a positive effect of the refinements of thinking prompts and/or affective element added to the scaffolding service. Subjective learning effectiveness was increasingly high, i.e. learners of UT2 felt significantly more successful in performing the simulation than UT1 and BL participants. Supplementary results showed that strong feelings of success continued until after the real-world interview, again especially UT2 learners had a strong feeling of success in conducting the interviews. This improved perception of students’ own learning achievement would be expected to be related to their actual performance and their motivation as well. Objective performance was only measured in assess mode of the simulator and due to constraints coming from the simulation developers, students had to enter the assess mode before entering the practice mode. With respect to self-evaluation, the only point in time students were asked to predict their scores was before entering the learning scenario of the assess mode. Since only a few students went back to assess mode after practicing and thus updated their score profile, most scores – as well as score predictions - do actually not reflect the additional learning effects from the (A)MSS. Concerning motivation, data also shows a positive impact of the augmented simulator right after the training, as overall motivation increased from the BL to UT1 and UT2. The small set of data obtained after the real-world interview show the highest ratings for UT2 on all questions, however differences among the three samples are not significant. Motivation to perform well and to further improve their skills was especially high. Differences were not significant, but power analyses suggest that the small sample sizes in BL and UT1 could be responsible for this result. With the improved service, results from the note-taking tools show that the notes students took were increasingly of reflective nature. Thus the type of notes taken by the students supports the assumption that AMSS fosters metacognition and reflection, which points to an additional
benefit that users get out of the enhanced simulator. However, the number of unique users who actually worked with the thinking prompts is rather low: in UT2 30 students left text entries in the AMSS open reflection section and/or the ETU notepad. One important reason is that the ImREAL services were only available in practice mode, but almost 60% of the students did the training only in the assess mode (which is prerequisite for the practice mode). Thus, ways of engaging more learners to use the simulator’s practice mode, to actively use the thinking prompts and to provide even more reflections should be explored. This might simply consist in opening up the possibility to precede the assess mode with a training in practice mode. Using the practice mode first would also allow an evaluation of the AMSS with respect to objective as well as predicted performance. For the latter, self-assessments undertaken after the training but before receiving feedback from the simulator, would also give an impression of students’ self-evaluation skills.

Comments from the post-interview-questionnaire revealed that reasons for not using the note-taking tool were either that students didn’t think it was necessary (6 comments) or that they weren’t aware of it or how it is used (2 comments). Thus, for future applications, it is necessary to find ways to attract more students into the practice mode and to actually use the prompts. Since the usage by those who saw the prompts was good, it is important that in the future the instructions point out more clearly that this service is available and how it is used. Supplementary results from the post-interview phase suggest that there is no clear preference for one mode over the other which might be due to the fact that each mode offers specific features that are helpful to students (scores in the assess mode, thinking prompts in the practice mode). Beside the fact that not all students engaged or could engage with AMSS and the reflection tool, the decreasing number of text entries made requires reflection. Since the trigger rules of the scaffolding service had been refined in Phase 2, however, less thinking and reflection prompts were presented, which naturally led to less frequent possibilities to take reflective notes in the open reflection section of AMSS.

**Learning Experience.** Overall, the obtained data indicate that the improved service with affective elements enhanced the learning experience for students. Flow experience increased from UT1 to UT2, and overall workload decreased from BL to UT2. Additionally, the level of frustration was lowest in UT2. Thus the simulator augmentation conveyed a stronger feeling of task involvement and fluency and the AMSS did not put additional cognitive load on learners. On the contrary, with the highest workload scores for mental demand and effort, medium scores for temporal demand, decreasing frustration, and increasing feeling of success, the enhanced simulation seems to yield an optimal distribution of workload. Thinking prompts provided by the (A)MSS were well received and perceived as being helpful with regard to content, timing, and support in SRL activities such as planning, monitoring, and analysing one’s learning performance. This can be viewed in connection with motivational beliefs, which may be influenced by positive sounding scaffolds and hints to optimize learning. If students see the prompts as support they can develop a positive attitude towards the whole learning process, which again influences the initial motivation of the learner (Vollmeyer & Rheinberg, 2006). Finally, participants were in a positive affective state after the training in the simulator (only measured in UT2), however we could not capture their emotional state during the training. The smiley affect measurement technology (Ateles-SAM) had been validated as a suitable approach for an interactive affect indicator in simulator-independent studies (Moore, Steiner, & Conlan, 2013). The affect technology was integrated in the simulator in Phase 2 and provided students with the opportunity to indicate their emotional states via this non-verbal scale when entering the simulation as well as during the training via the note-taking tool. However, this possibility of affect reporting during interaction with the simulator was not used by trial participants – mainly because they were unaware of this feature.
From a methodological point of view, additional future studies targeting the services in a more isolated manner and under more controlled conditions (i.e. involving guidance and instruction to use and how to use the services) may complement and refine the evaluation results obtained so far. As mentioned above, the scaffolding service was developed as an external tool to be used for augmenting different kinds of experiential simulators. The compromise between flexibility and specificity lead to several solutions regarding the integration of the service, which turned out to be not optimal with respect to a comprehensive and most efficient support of students’ learning experience. Furthermore, the application of a longitudinal evaluation design instead of a cross-section evaluation would give further insights in the underlying effects and better meet the requirements of a longer-term process.

**Conclusions and Outlook.** Overall, the evaluation outcomes clearly point to the usefulness of affective metacognitive scaffolding and give valuable information for further research. Results obtained in a series of three evaluation studies demonstrate that the simulator augmentation via AMSS enhanced user support as well as flow and fidelity of the simulation, leaving learners more motivated and engaged, without overly burdening them or interrupting the flow of their learning experience. Main avenues for future research include (a) stimulating interaction with the service, (b) investigating new models for triggering thinking prompts, and (c) meliorating the introduction to the system. Usage of the simulator’s practice mode and consequently, interaction with thinking prompts and SAM have been rather low, social network IDs have not been provided at all by our samples and thus, compromised use of the U-Sem service in Phase 2. This is partly due to the aforementioned trade-off between working with existing simulators that are actually deployed in practice instead of prototypes on the one hand, and having control over the entire process of technical integration on the other hand. In the presented case, the integration was carried out very cautiously in cooperation with developers and instructional designers in order to avoid interference with flow and additional workload. Evaluation results showed that this led to a lack of awareness regarding the availability of the tools, as well as a lack of understanding of their usefulness. As the SAM technology was most likely not noticed by a large part of participants, as a first step a more prominent presentation of the affect reporting tool has already been achieved by updating the user interface for a better visibility of SAM.

The fact that workload, especially frustration, decreased and perceived flow increased with the integrated service indicates that the current integration of prompts was not perceived as disruptive to the learning experience. This suggests that there is room left for further exploring more explicit ways to prompt users into indicating their emotions and reflecting upon their learning, e.g. by raising the visibility of triggers to draw learners’ attention to the opportunity to pause and reflect. The aim is to find an optimal balance between keeping the natural flow of the simulation and fostering interaction with the AMSS. Thereby, integration plans should also account for the constraints of a specific simulation, e.g. simulation length and knowledge domain. In addition, the target audience needs to be considered, since different levels of support in terms of intervention points or triggering rules and types of prompts might be particularly suitable for certain groups of learners (see e.g. Verpoorten, Westera, & Specht, 2011 for a classification of reflection prompts). First approaches in this direction concern, for instance, the elaboration of different scaffolding strategies in line with cultural variations in learning and in the perception of thinking prompts (see Steiner et al., 2013a). With respect to appropriate timing of reflection triggers, thinking prompts are currently linked to specific points in the simulation, as authored by the instructional designer, or are displayed on request by a learner. Krogstie, Prilla, & Pammer (2013) have developed a model for computer-supported reflective learning at work. Research is also underway to automatically identify moments of reflection, which could lead to frameworks for the automatic generation of reflection prompts at appropriate times. Thus, future triggers could be event-driven, cued as a response to the learners’ behaviour and characteristics.
Finally, future work should concentrate on ways to support awareness and perceived usefulness. Results demonstrated that although learners may think of themselves as self-aware and possessing emotional intelligence, they do not necessarily see a role for this with their professional training. Thus, a careful introduction in the system and its features is needed as well as a clear explanation of expected learning benefits and use of given information. Explanation should be given to users why the use of features like affect reporting or note taking and reflection are meaningful and should be used. Benefits concern the tools’ potential to improve learning competence, to develop higher self-awareness, as well as their effect on personalising the learning experience. With respect to the use of information from users’ social digital traces, it is key to explicitly explain to users what kind of information is used, how the information is retrieved, how this information can be used for personalised learning. Giving users knowledge and control over the used information and the services is assumed to help to foster their willingness to disclose personal information and to make use of the different features. Apart from that, opening up information about the research group behind the technologies may help building trust, acceptance, and usage of the services.

The diffident use of the augmented simulator features may, to some extent, be correlated with students’ background. The AMSS was developed as a tool external to a specific simulator, working alongside it and prepared for use with different training environments. The service has been integrated also with simulations on intercultural competence training in business and in academic contexts provided via a different simulation infrastructure (see e.g. Dimitrova et al., 2013; Hetzner & Pannese, 2011), which provides a starting point for further studies in another domain and with cohorts from different disciplines.

Our study was grounded on the overall hypothesis that affective metacognitive scaffolding positively affects self-regulated learning, learning experience, and learning effectiveness of simulations. Effects on affective states need deeper investigation in future studies, ensuring adoption and usage of the affective elements of the service. While effects on metacognition and learners’ subjective experience could be identified, the effectiveness of learning can only be demonstrated by increases in performance, which was deliberately not measured in this study. Actual changes in learning performance take place only after longer time intervals following the acquisition and practicing of SRL skills, which would necessitate comprehensive longitudinal studies and analyses.

In our study, students using the offered services perceived them as positive and helpful for their learning. The challenge is not so much to develop technologies for further supporting and personalizing learning experiences, but to change user attitudes and workflows towards a more open approach to learning systems in order to enable learners to perceive the benefits of these systems and take advantage of all offered features.

References


