

# Evaluating the Adaptation of a Learning System before the Prototype is Ready: A Paper-based Lab Study

Tobias Ley<sup>1,2</sup>, Barbara Kump<sup>3</sup>, Antonia Maas<sup>2</sup>, Neil Maiden<sup>4</sup>, Dietrich Albert<sup>2</sup>

<sup>1</sup> Know-Center, Inffeldgasse 21a,  
8010 Graz, Austria  
tley@know-center.at

<sup>2</sup> Cognitive Science Section, University of Graz, Universitätsplatz 2,  
8010 Graz, Austria  
{tobias.ley, antonia.maas, dietrich.albert}@uni-graz.at

<sup>3</sup> Knowledge Management Institute, Graz University of Technology, Inffeldgasse 21a,  
8010 Graz, Austria  
bkump@tugraz.at

<sup>4</sup> Centre for HCI Design, City University London, Northampton Square, College Building,  
London, EC1V 0HB, United Kingdom  
N.A.M.Maiden@city.ac.uk

**Abstract.** We report on results of a paper-based lab study that used information on task performance, self appraisal and personal learning need assessment to validate the adaptation mechanisms for a work-integrated learning system. We discuss the results in the wider context of the evaluation of adaptive systems where the validation methods we used can be transferred to a work-based setting to iteratively refine adaptation mechanisms and improve model validity.

**Keywords:** Adaptive Learning Systems, Evaluation, Task-based Competency Assessment, Learning Need Analysis, Knowledge Space Theory

## 1 Evaluating Adaptive Systems in Due Time

Learning systems that adapt to the characteristics of their users have had a long history. Due to the complexity of most adaptive systems, it has been acknowledged that rigorous evaluation is indispensable in order to deliver worthwhile adaptive functionality and to justify the considerable effort of implementation. This is also reflected in the substantial amount of evaluations that have been published so far. Van Velsen et al. [1] present an overview and have noted several limitations in current evaluation practices. A variety of evaluation frameworks have been presented [2], [3], [4], all of which propose to break down the adaptive system into assessable, self-contained functional units.

The core research question when evaluating an adaptive system concerns the appropriateness of the adaptation. Typically, two aspects are distinguished, (a) the

*inference mechanisms* and (b) the *adaptation decision*. While endeavors related to (a) seek to answer the question if user characteristics are successfully detected by the adaptive system, evaluations of (b) ask if the adaptation decisions are valid and meaningful, given selected assessment results.

It is recommended that these two research questions are investigated in an experimental setting using a running system (or prototype) where the algorithms are already implemented [1], [3]. The problem is that in many situations the development cycle of the software product is short and the evaluation might become obsolete as soon as a new version has been developed [4].

For this reason, we are pursuing a multifaceted evaluation approach for adaptive systems. By gathering both field and experimental evidence, we are checking validity of models and appropriateness of the adaptation mechanisms over the course of design, implementation and use of the system in an iterative manner. With this article, we describe an experimental evaluation that seeks to answer the above mentioned research questions in a controlled lab situation but *without* a running prototype, that is, in due time *before* the system is actually developed. After a brief presentation of the results, we will discuss the wider implications of our approach for evaluation research for adaptive systems.

## 2 Evaluation of an Adaptive Work-Integrated Learning System

Our paper-based evaluation has been conducted in the course of the APOSDLE<sup>1</sup> project. APOSDLE is a system for supporting adaptive work-integrated learning (WIL). With WIL, we refer to learning that happens directly in a user's work context, which is deemed beneficial for maximising learning transfer [5]. APOSDLE offers learning content and recommends experts based on both the demands of the current tasks, as well as the user's state of knowledge with regard to this task. APOSDLE is currently available for five different application domains. The experiment in this article has been conducted for the *requirements engineering* domain.

### 2.1 Adaptation in APOSDLE

Corresponding to the basic ideas of *competence-based knowledge space theory* [6], the users' knowledge states in APOSDLE are modelled in terms of sets of competencies (single elements of domain related cognitive skill or knowledge). In order to make inferences on a user's competencies, APOSDLE observes the tasks a user has worked on in the past. Each of the tasks is linked to a set of competencies (*task demand*). Taking into account the task demands of all previously performed tasks, their frequency and success, APOSDLE builds the user's instance of the user model by making inferences on the likely state of knowledge. In the following, this procedure shall be termed *task-based competency assessment*.

---

<sup>1</sup> APOSDLE ([www.aposdle.org](http://www.aposdle.org)) has been partially funded under grant 027023 in the IST work programme of the European Community.

In order to adapt to the needs of a user in a given situation, APOSDLE performs a *learning need analysis* (also termed competency gap analysis elsewhere): The task demand of a task is compared to the set of competencies of the user. If there is a discrepancy (*learning need*), APOSDLE suggests learning content which should help the user acquire exactly these missing competencies in a pedagogically reasonable sequence. In order to perform these adaptations, the domain model of APOSDLE contains *tasks* and *competencies* as well as a mapping that assigns required competencies to tasks. A prerequisite relation exists both for competencies and for tasks.

For the present study, the domain model was modelled in terms of the tasks in the requirements engineering domain (e.g. *Complete the normal course specification for a use case*, or *Carry out a stakeholder analysis*), as well as the competencies needed to perform these tasks (e.g. *Understanding of strategic dependency models*, or *Knowledge of different types of system stakeholders*). The model has been constructed, initially validated and refined in a previous study [7].

## 2.2 Design, Procedure and Hypotheses of the Study

The aim of our study was to test different algorithms for task-based competency assessment and learning need analysis. The participants were a sample of nineteen requirements engineering (RE) students. We had selected eight tasks from two sub-domains of the RESCUE process (Requirements Engineering with Scenarios in User-Centred Environments, [8]). According to the domain model, 22 competencies were required in total to perform well in these eight tasks.

Each student had to work on four exercises which had been constructed to directly map to the tasks from the task model. For example, they were asked to write a use case specification for an iPod, or to carry out a stakeholder analysis for a realtime travel alert system of an underground. The exercises were constructed to be ecologically valid, i.e. that they corresponded well to tasks that would have to be conducted by requirements engineers in a work-based setting. The sequence of exercises was randomized across participants.

Before conducting the exercises, students gave both competency and task self appraisals. Performance in the exercises was measured by marks assigned by a professor of RE. After each exercise, students were asked for an appraisal of their performance for the exercise just conducted. They were also asked to indicate which additional knowledge they would have required to perform better, both in a free answer and a multiple choice format. Answers from the free answer format were later subjected to a deductive content analysis that mapped each free answer to a competency from the domain, or a new one. The multiple choice items contained all competencies assigned to the particular task in the domain model as well as a number of distractors, i.e. other competencies not assigned to that task. Competencies had been reformulated to describe personal learning needs (e.g. *I would need to learn what is a domain lexicon and how to apply it*).

Self appraisal was included in this study as it is a common and economical way to assess competencies or performance in the workplace [9]. In accordance with prior research [10], we expected that self appraisals would correspond to actual task

performance (hypothesis 1). The second hypothesis looked at the personal learning needs indicated by the students. We assumed that competencies selected by the students for each task would, in a substantial proportion of cases, correspond to competencies assigned to the task in the domain model. If this were not the case, learning need analysis based on the task-competency assignment in the domain model would not be possible. Lastly, we employed different algorithms for task-based competency assessment and investigated whether they would correspond to competency self appraisal by the students (hypothesis 3).

## 2.3 Results of the Study

### 2.3.1 Hypothesis 1: Self appraisal and task performance

A one-way Analysis of Variance which compared the marks received for the exercises between those students that had indicated they were able to perform the task without assistance and those that had indicated otherwise showed that contrary to our expectations there was no relationship between self appraisal *before* task performance and task performance as assessed by the marks received ( $F_{(1,69)} = .007$ , ns.). There was, however, a moderate relationship between self appraisal *after* task performance and task performance itself as measured by a Spearman Rank Correlation Coefficient ( $\rho = -.38$ ,  $p < .01$ ). It appears that students were not able to realistically predict their performance in the tasks before they conducted the exercise. Their appraisals after task performance, then, were slightly more accurate.

### 2.3.2 Hypothesis 2: Personal Learning Needs

Asked for their personal learning needs after the exercises, the students were significantly more likely to chose learning needs assigned to the particular tasks ( $M = 2.63$ ) than distractors ( $M = 1.06$ ) ( $t = 5.23$ ;  $p < .001$ ). This confirms the hypothesis and is an indication of the overall validity of the modeled structures. Similarly, learning needs extracted from student free answers in the content analysis were to a large degree those originally assigned to the particular task (60 vs. 39). Two of the tasks account for more than two thirds of the contradicting answers, namely task 3 (17 contradicting learning needs) and task 4 (11 contradicting learning needs). This gives strong reason to believe that there had been missing competency assignments for these tasks. Particularly, six new competencies that had not been part of the original list were suggested from analyzing the free answers. These include items like *Knowledge about different types of requirements*. These missing competencies may have also led to violations of the prerequisite relation on tasks which were found when comparing the relation to the obtained answer patterns.

### 2.3.3 Hypothesis 3: Task-based Competency Assessment

Assessing competencies from the observation of task performance is one of the key benefits of using competence-based knowledge space theory. The usual way to do this is to take the union of all assigned competencies for all successfully mastered tasks. As [11] has shown previously, this method may lead to contradictions, especially in

the case where the numbers of competencies assigned to tasks are large, and therefore suggests using both positive as well as negative task performance information. In the present study, we have compared two algorithms to predict the knowledge state of the students from task based information. Three predictors for task information were used (task self appraisal prior to task, task self appraisal after task, and task performance assessed by the expert) and each was correlated with competency self appraisal.

Although in all three cases, correlation coefficients were higher for the algorithm that took negative task performance information into account, the coefficients were of only small magnitude, ranging between  $\rho=-.017$  and  $\rho=.129$  (Spearman Rank Correlation), and with only one becoming significant. We partly attribute these low correlations to the fact that competency self appraisal is probably not a very accurate criterion for the actual knowledge state of our subjects.

### 3 Discussion and Outlook

The results caution towards the use of self appraisal information as a criterion variable for evaluating the adaptation of a learning system, but also as an input variable for the user model. Self appraisal by our subjects showed to be unrelated to their actual performance. A possible reason for this may be that the students were rather inexperienced in the domain. We assumed this also holds for the case of work-integrated learning, which is in line with [12] who found high validity of self appraisal only for experienced job holders. Also social desirability may have resulted in answer tendencies, as all performance appraisals before task execution were much higher than after.

The results for task-based competency assessment were largely unsatisfying due to low validity of the criterion variable. Future research will show whether our algorithms prove to be more successful than traditional measures. In any case, the question of a valid criterion variable for a knowledge state (which at the same time has ecological validity), will continue to be a challenge in work-based learning.

Checking for personal learning needs has proven to be a promising way to identify parts of the models with low validity (missing competencies in our case). In combination with indicators that estimate violations of the prerequisite relation from answer patterns, these methods can be used to iteratively refine models once they are in use.

We are currently planning an extensive summative evaluation of the APOSDLE system and the components contained therein. A purpose of the study reported here was to gain an understanding of how paper-based methods could be applied for evaluating the adaptation of a learning system specifically in the context of adaptive work-integrated learning so that they may be incorporated in a more comprehensive evaluation approach in a field setting. For that reason, all the validation methods employed here can be easily transferred to a setting where the learning system is in operation and provides suggestions for learning needs and learning content during actual task performance. The role of the RE professor in our study could then be taken by supervisors of those working in the tasks. Short and unobtrusive system dialogues after task execution could be used for collecting self appraisal as well as indications

of actual personal learning needs from the learners. This information could then be fed back to adaptation designers to iteratively refine the adaptation decision or the underlying domain model, such as suggesting additional competency assignments for particular tasks or missing competencies altogether.

### Acknowledgments

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG. Contributions of four anonymous reviewers to an earlier draft of this paper are kindly acknowledged.

### References

1. Van Velsen, L., Van Der Geest, T., Klaassen, R., Steehouder, M.: User-centered evaluation of adaptive and adaptable systems: a literature review. *The Knowledge Engineering Review*, 23 (3), 261-281 (2008)
2. Brusilovsky, P., Karagiannidis, C., Sampson, D.: The Benefits of Layered Evaluation of Adaptive Applications and Services. In: S. Weibelzahl; D. Chin; G. Weber (eds.): *Empirical evaluation of adaptive systems. Workshop at the UM 2001*, pp. 1-8, (2001)
3. Paramythias, A., Totter, A., Stephanidis, C.: A modular approach to the evaluation of adaptive user interfaces. In: S. Weibelzahl, D. C. a. G. (eds.): *Empirical evaluation of adaptive systems: Workshop at the UM 2001*, pp. 9-24 (2001)
4. Weibelzahl, S., Lauer, C. U.: Framework for the evaluation of adaptive CBR-systems. In: I. Vollrath; S. Schmitt; U. Reimer (eds.): *Experience Management as Reuse of Knowledge. GWCBR 2001*, pp. 254-263, Baden-Baden, Germany (2001)
5. Lindstaedt, S. N., Ley, T., Scheir, P., Ulbrich, A.: Applying Scruffy Methods to Enable Work-integrated Learning. *Upgrade: The European Journal of the Informatics Professional*, 9 (3) 44-50 (2008)
6. Korossy, K.: Extending the theory of knowledge spaces: A competence-performance approach. *Zeitschrift für Psychologie*, 205, 53-82 (1997)
7. Ley, T., Ulbrich, A., Scheir, P., Lindstaedt, S. N., Kump, B., Albert, D.: Modelling Competencies for supporting Work-integrated Learning in Knowledge Work. *Journal of Knowledge Management*, 12 (6), 31-47 (2008)
8. Maiden, N. A., Jones, S. V.: *The RESCUE Requirements Engineering Process - An Integrated User-centered Requirements Engineering Process, Version 4.1*. Centre for HCI Design, The City University, London/UK (2004)
9. Hoffman, C., Nathan, B. & Holden, L.: A Comparison of Validation Criteria: Objective versus Subjective Performance Measures and Self- versus Supervisor Ratings, *Personnel Psychology*, 44, 601-619 (1991)
10. Mabe, P. & West, S.: Validity of Self-Evaluation of Ability: A Review and Meta-Analysis, *Journal of Applied Psychology*, 67, 280-296 (1982)
11. Ley, T.: *Organizational Competency Management - A Competence Performance Approach*. Shaker, Aachen (2006)
12. Muellerbuechhof, R. & Zehrt, P.: Vergleich subjektiver und objektiver Messverfahren für die Bestimmung von Methodenkompetenz am Beispiel der Kompetenzmessung bei technischem Fachpersonal. *Zeitschrift für Arbeits- und Organisationspsychologie*, 48, 132-138 (2004)