Which User Interactions Predict Levels of Expertise in Work-integrated Learning?

Tobias Ley[1], Barbara Kump[2]

[1] Institute of Informatics
Tallinn University, Estonia
{tley@tlu.ee}

[2] Department of Human Resources and Organization,
Vienna University of Applied Sciences of WKW, Austria
{barbara.Kump@fh-wien.ac.at}

Abstract. Predicting knowledge levels from user's implicit interactions with an adaptive system is a difficult task, particularly in learning systems that are used in the context of daily work tasks. We have collected interactions of six persons working with the adaptive work-integrated learning system APOSDLE over a period of two months to find out whether naturally occurring interactions with the system can be used to predict their level of expertise. One set of interactions is based on the tasks they performed, the other on a number of additional Knowledge Indicating Events (KIE). We find that the addition of KIE significantly improves the prediction as compared to using tasks only. Both approaches are superior to a model that uses only the frequencies of events.

Keywords: learning analytics, user modeling, knowledge indicating events

1 Introduction

Learner models [1] are at the core of adaptive learning systems, as they enable a system to adapt to individual learning needs. A variety of learner characteristics (knowledge, interest, learning style, etc.) may be represented in a learner model. Within this paper we concentrate on the learner's knowledge state. To make sure that the learning system can adapt to the knowledge of its users, continuous maintenance of the learner model is necessary.

In our previous work, we have followed two alternative paths for updating the user model which both relied on implicit information derived from natural user behavior when interacting with the system: First, an approach which is based on Competence-based Knowledge Space Theory (CbKST), and second, an approach which is based on Knowledge Indicating Events (KIE).

We have suggested Competence-based Knowledge Space Theory (CbKST) as a means to derive a user's knowledge state from tracking the user's working tasks he or she has been performing as part of his or her normal job duties [2]. The user model in this case simply adds all knowledge and skills assumed to be necessary for perform-

adfa, p. 1, 2011.

ing well in a particular task. While CbKST has been successfully applied in educational settings, certain limitations exist that make it difficult to transfer this approach to implicit user modeling in a workplace setting. There, it is difficult to observe successful task performance (implicitly) from which available knowledge and skills could be derived. At the same time, our experience has shown that users use these systems much more flexibly than simply obtaining scaffolding for performing in task. Rather, they browse or search topics, obtain learning hints, contact others or are contacted by others for help etc. [3]. As a consequence, a user model which is only based on tasks performed does not make use of all available information.

As a second approach, we have therefore suggested KIE as a means for non-invasively diagnosing user knowledge in an adaptive work-integrated learning system [4][5]. The assumption is that all behavior of a user with the system potentially can give evidence of whether a user possesses knowledge and skills. For example, an assumption could be that a person who clicks on the 'help'-button for a concept has little knowledge of this concept. While this may not be the case for each behavior, collecting enough evidence may in the end lead to a more accurate user model than if only task performance is considered.

In this paper, we compare an approach which is based on CbKST (i.e., only on tasks performed) with an approach based on KIE where a more extensive set of actions is considered. As collecting realistic user interactions in the workplace is extremely difficult and has to our knowledge not been realized, we perform an explorative study in which the work-integrated learning system APOSDLE [3] was used by six users for a period of two months. We use these two approaches to find out whether user interactions can be used to predict self-appraisal of expertise. We also compare the model to peer-assessment and to a simple frequency-based approach as a baseline.

2 Traditional Ways of Knowledge Diagnosis

In the context of adaptive learning systems, such as Adaptive Educational Hypermedia (AEH [6]), and Intelligent Tutoring Systems (ITS [7]), assessment and user model maintenance with regard to user knowledge has been typically based on *testing* for knowledge and skill. In ITS, the emphasis was placed on procedural problem solving skills. From successful or unsuccessful solutions of a problem, the skill level of a learner was deduced, and hence curriculum sequencing or learning prompts were adapted (for a recent review see [8]). The underlying models that were used in these systems include a cognitive model of the learning domain [9] [10], Constraint-based Models [11], Competence-based Knowledge Space Theory [12] or Bayesian Nets [13]. The general idea in all of these approaches has been that by tracing observable solution behavior, inferences on the underlying knowledge and skills can be made. The process of inferring knowledge in one concept from diagnosing knowledge in another concept has been called k*nowledge propagation* [1] or *knowledge update* [14].

While ITS have traditionally been more concerned with testing and practicing of skills, AEH have been more concerned with teaching conceptual knowledge. As a result, the user interaction with AEH is quite different from that in ITS: In AEH, learn-

ers consume learning material by reading texts, receiving pictures or listening to audio or video materials. Learning is supposed to be triggered by these processes and is seen to be an internal process that cannot be observed directly. For this reason, AEH attempt to guide learners through the large quantities of contents in an optimal way by recommending suitable contents or sequences of materials.

The most straightforward way of diagnosing user knowledge in an AEH system is simply asking learners questions, from which their knowledge can be deduced. This happens in the form of quizzes or tests. Usually, this form is seen as the 'most effective way' of diagnosis [15]. There has been a realization, however, that it is not always feasible to ask questions, and thus other methods have been used. Visiting pages seems to be the most common one, which assumes that a visit indicates a certain pages has been read and understood at least to a certain degree. This measure of a user's knowledge based on a page visit has been implemented, for example, in AHA! [16].

More recently, research into ITS and AEH has grown together, giving rise to Adaptive and Intelligent Web-based Educational Systems (AIWBES, Brusilovsky & Peylo, 2003). In the context of AIWBES, some steps have been made in the direction of non-invasive diagnosis of user knowledge. Brusilovsky & Millan [1] use the term *evidence-bearing event* to refer to a user activity that reveals that the user has knowledge about a certain domain element. These events can be "an answer to a test item, solution of a problem, teacher's opinion, the number of Web pages relevant to the element K that have been visited, etc." (p.26). This description of an evidence-bearing event includes the possibility of diagnosing knowledge implicitly, from a user visiting a page, as used in ElmArt [17] and LS-Plan [15], for instance. Similar approaches to assessing user knowledge from different sources of evidence have been proposed with the Cumulate server of KnowledgeTree [18] and Personis [19].

As stated above, the clear separation between ITS and AHS is gradually dissolving. Nowadays, integrated learning environments (termed, e.g. virtual learning environments or virtual classrooms) certainly contain elements of both ITS and AHS. While a learning environment may contain exercises, it could also contain learning materials that would allow a student to learn certain contents, or interactive exercises or simulations that would allow practicing certain skills. Most importantly, all modern learning environments also contain communication and collaboration functionalities which allow learners to get into contact with fellow learners, tutors or teachers. A large array of these functionalities allows for collaborative learning. Despite these developments, utilizing implicit measures for diagnosing knowledge plays a subordinate role in AIWBES. Instead, in most of these systems, testing still plays a predominant role, and other implicit forms of assessment are usually seen as less effective, less reliable and therefore as generally inferior.

Increasingly, learning environments seek to support learners not in an artificially created learning environment, but embed learning in other natural activities. Examples are game-based environments [20], or work-based learning environments that should support people in performing real tasks they are working on [23]. These systems blur the boundaries between a learning system and other types of natural activities, like natural work tasks.

3 Implicit User Modelling from Naturally Occuring Events during Work Tasks

In the context of workplace learning, where learning happens mostly in non-formal arrangements, testing for knowledge and skills (e.g. with knowledge quizzes) is not a realistic option for multiple reasons [2]. Rather, it is assumed that in workplace practices, knowledge and skills are being applied in the context of normal work tasks. Therefore, more implicit methods of knowledge diagnosis are needed.

Several implicit approaches have been proposed. The ADAPTS system [25], for example, is used for adaptive help in complex technical maintenance tasks. It combines a domain model (different components of the maintained system), a task model (a hierarchical structure of the maintenance tasks) and a user model which tracks knowledge and experience of each technician with these components and the tasks. For each task, there is a mapping to all domain elements involved in the task. The user model uses information on which tasks were successfully performed and thereby diagnoses the knowledge of a person in the tasks and components. With this system, however, no longer field study has been conducted to collect naturally occurring interactions.

One system that implicitly infers the users' IT skills, more specifically their Microsoft Word literacy, from logs is the system OWL, an acronym for 'Organisation-Wide Learning' [21]. OWL is a recommender system for learning that logs Word commands of each user such as 'print', 'copy', 'paste', etc. The commands of different users are pooled and for each single user it is decided whether he or she over- or under-uses a certain command. Under-used commands are then recommended to the users in order to improve their Microsoft Word literacy. The appeal of this approach work lies in the operationalisation of skills: The authors diagnose the user's MS Word literacy by observing his or her interaction with MS Word – practically no inference is needed for this diagnosis. This is in contrast with approaches that try to diagnose skills which are by far more difficult to observe, such as a user's knowledge in the area of 'inference statistical analyses', or 'programming'.

Other systems that give advice in the task context are recommendation systems for software development [22]. These build a user model on a per item basis and therefore do not diagnose any knowledge or generalizable skill. They observe which methods the user has already used and refrain from recommending these items again when the methods have been used for a number of times.

In the context of our own work in the area of work-integrated learning, we have proposed an approach for inferring employee competencies from past task performance in knowledge work [2]. The user model, in this case, assumes a mapping between the working tasks of a domain on the one hand, and the knowledge and skills needed to perform these tasks on the other. CbKST is then used to infer availability of knowledge and skills when performance of certain tasks is observed [23]. Because dependencies exist in the learning domain between certain knowledge and skills in the sense of prerequisites, inferences can exploit these dependencies to make the update more efficient.

Recently, and as an attempt to extend this approach, we have suggested the idea of Knowledge Indicating Events (KIE) as a means for non-invasively diagnosing user knowledge [4]. Conceptually, the idea of KIE is in line with evidence-centered assessment design as suggested by [26]. In a nutshell, KIE are naturally occurring user actions (e.g., selecting a link, accessing a learning hint) that are interpreted as evidences for a user's knowledge state. When each of these actions is connected to a concept of the domain model, then inferences can be made about the user's knowledge level about the particular domain concept.

Clearly, in order to assess the usefulness of KIE for diagnosing user knowledge in a naturalistic setting, empirical data is needed. Such empirical data requires a field study where users interact with the learning system in a naturalistic way during their daily work tasks over a longer period of time. Within the present paper we present a first field study where we statistically compare the two approaches, CbKST and KIE with regard to their success in diagnosing user knowledge. In the next section, we will introduce the field study in which the data was collected.

4 Field Study

For the field study we used APOSDLE[1], an adaptive work-integrated learning system which aims to improve knowledge worker productivity by supporting learning within everyday work tasks. Within APOSDLE, the learner model is used for ranking learning goals, recommending useful learning content, and for suggesting knowledgeable people [3][4].

4.1 APOSDLE and its Learner Model

APOSDLE can be instantiated to various domains by creating new semantic domain models. Typically, an APOSDLE domain model consists of 100 to 150 domain concepts. The domain model also includes mappings between the domain concepts and various other elements of the system, such as work tasks, resources (document snippets), or learning paths. The APOSDLE learner model is designed as a layered overlay of the APOSDLE domain model.

To employ the KIE approach for APOSDLE, all possible user interactions with the system that could be related with one or several domain concepts were treated as potential KIE. APOSDLE tracks all user interactions with these domain model elements and from these interactions infers the knowledge state of a user for each of the concepts of the domain model based on a very simple rule-based algorithm.[2] The users receive different recommendations for a topic (different types of resources etc.), de-

[1] www.aposdle.org

[2] In the version of APOSDLE used in this study, the user model contained for each concept one of three knowledge levels (learner, worker, and supporter) which was also visualized in the open learner model [5]. For the present study, however, these inferences were not taken into account.

pending on the detected knowledge level. While these recommendations have been shown to improve task performance [24], in a next step, the simple rules should be replaced by a statistical approach. Therefore, the predictive power of each of the KIE that was used within APOSDLE should be investigated in a field evaluation. The question that we were posing at the outset of our explorative study was: Which user interactions predict levels of expertise in work-integrated learning?

4.2 Procedure of the Field Evaluation

The field evaluation was carried out in an innovation management company in Austria. Typical tasks in the innovation management domain are for instance, *analyzing trends in a certain industry*, or *identifying strategies of industry competitors in the market*. The APOSDLE innovation management domain comprises 144 domain concepts. Examples for domain concepts from the innovation management domain are *idea management*, *market analysis*, *patent*, or *knowledge management*.

The participants in the evaluation study were 6 employees of the company who (after being trained in the use of the system) used APOSDLE for a period of approximately 2 months in their regular work. As the interaction of the users with APOSDLE should be as realistic and natural as possible, there was no specific scenario used for the evaluation. When working on their laptops and desktop computers, the innovation consultants in the innovation management company could interact with APOSDLE at any time to search for information, and to receive learning support. For example, they could look up techniques for workshop design, or view examples of customer offers from similar previous projects. Whenever APOSDLE detected a task or topic the user was working on through automatic task detection, the user was free to view the suggestions on the task or topic provided by APOSDLE, and to explore further materials. Users were in no way forced or encouraged to use the system. So we consider the data gathered as a realistic snapshot of how users naturally interact with a work-integrated learning system in a 2-month period.

In order to establish an external criterion for the knowledge levels in each topic, values of the expertise level of a user in each topic were needed. As mentioned above, objective knowledge testing is not possible in workplace learning for a number of reasons (no tests are available and there is large resistance of experts to be tested for their expertise). We therefore drew on self and peer appraisal which is a very common form of appraisal in workplace settings. For example, it is used in a number of personnel development tools, such as 360 degree feedback instruments [27]. Furthermore, while research has shown the biased nature of self-appraisal [28], there is evidence to suggest that it can be a valid form of assessing states of knowledge in technical areas and for experienced job holders [29].

Therefore, a special type of card sorting was applied. Card sorting as a technique for software evaluation has also been used, for instance by Wild et al. [30]. For self- and peer-assessment, each of the participants was provided with a set of 144 cards, one for each topic in the domain. In a first self-assessment trial, the participant was asked to sort each of the cards (topics) into one of five categories: (i) “I am rather inexperienced in this topic”, (ii) “I am neither very inexperienced nor very experienced

in this topic", (iii) "I am very experienced in this topic", (iv) "This is not my area of work", and (v) "I do not understand the topic".

In a second run, two peer-assessments were gathered for each participant. Therefore, each participant was asked to sort the same cards for two of their colleagues. In addition to the five categories in the self-assessment, a sixth category, (vi) "I cannot decide on the level of expertise of my colleague in this topic", was available for peer-assessment. That way, for each participant, we obtained one complete self-assessment and two complete peer-assessments for each topic (domain model element) in the innovation domain.

4.3 Results

4.3.1 Knowledge Indicating Events Collected in the Study

We define KIE as traceable naturally occurring interactions of a user with a learning environment that potentially allow inferences on the user's knowledge and skills. In our view, KIE can provide both positive evidence for knowledge, and negative evidence of knowledge, that is, evidence for the absence of knowledge. Examples of KIE include contacting a person about a topic, or annotating a document with a topic.

APOSDLE provides a variety of functionality, including the adaptive presentation of learning goals, the recommendation of knowledgeable colleagues, or the recommendation of documents and text passages. For the KIE approach, we identified all possible user interactions with APOSDLE that could possibly indicate user knowledge. All possible user interactions the system that could be related with one or several domain concepts were treated as potential KIE The rationale for not preselecting any type of interaction was the exploratory approach in this study: We were interested in the predictive power of the different types of KIE. The list of KIE can be seen from the first column in Table 1.

The KIE *perform topic* indicates that APOSDLE automatically detected a topic on which the user was working on (e.g., the user was creating a presentation where he or she used the term "innovation management"). Another KIE is called *select learning goal – topic*. This KIE implies that a person has clicked on an item in a list of learning goals which lead to the display of further information for the topic mentioned in the learning goal. The KIE *view resource* indicates that a user viewed a resource (e.g. a report, presentation, image or video) annotated with a certain topic. The KIE *edit annotation* means that a user annotated a resource with a certain topic, or modified an existing annotation. The KIE *perform task* means that a user carries out a task which requires knowledge about certain topics. This task-topic relationship is represented in the APOSDLE domain model. The KIE *contact peers* means that a user contacted another APOSDLE user via APOSDLE to communicate about some topic within the APOSDLE domain model. A KIE *select learning goal task* exists that indicates that a person has clicked on an item in a list of learning goals which were presented for a task the user was working on. The KIE *get learning hints* means that a user requested additional 'learning hints' (questions, exercises) for a topic at hand. The KIE *create learning path* indicates that a topic occurred in a intentionally created 'learning path' (a list of learning goals which are arranged in a didactically beneficial sequence). Fi-

nally, the KIE *being contacted* means that a user was contacted by another user via APOSDLE about a certain topic.

Table 1 gives an overview over the system usage of each participant in the testing phase of APOSDLE in terms of the frequency of how often each KIE was applied by each of the participants (P1 to P6). On average, a user applied 303.83 KIE (s =288.81) during the evaluation phase. P1 applied 875 KIE; hence, she used the system at most and rather frequently. P4, P5, and P6 showed moderate usage in terms of number of KIE. P2 and P3 showed a low usage of the system with less than 120 KIE.

What becomes obvious already from the table is the unbalanced distribution of KIE types: the most frequently applied KIE were performing a topic (855), selecting a learning goal for a topic (399), viewing a resource (180), editing an annotation (176), and performing a task (149). All other KIE occurred much less frequently (some even only sporadically).

Table 1: Frequency of knowledge indicating events (KIE) for each of 6 participants (P1-P6)

KIE	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*	*Total*
Perform topic	375	59	36	118	143	124	855
Select learning goal-topic	158	16	29	66	65	65	399
View resource	116	3	7	8	36	10	180
Edit annotation	143	3	2	2	18	8	176
Perform task	68	3	26	23	12	17	149
Contact peers	2	6	7	2	--	8	25
Select learning goal-task	5	5	3	1	1	--	15
Get learning hints	--	--	5	--	--	6	11
Create new learning path	5	--	--	1	1	2	9
Being contacted	3	--	--	1	--	--	4
Total KIE	875	95	115	222	276	240	1823

4.3.2 Predictive Models

In order to predict the level of expertise, a model was needed that could predict as a criterion a user's level of expertise (the user's self-assessment of being a beginner, advanced or an expert) for each of the 144 domain topics using different types of predictors (e.g. tasks only vs. all tracked events). We chose to use a multiple linear regression model as a prediction model which fits a linear regression line through the data by minimizing the squared difference between actual category values (self-assessed expertise) and prediction. The overall regression coefficient (R) gives an overall indication of the fit of the model, while the standardized regression weights (beta) give an indication of the weight each predictor carries in the prediction.

We were only concerned in this analysis with predicting the level of expertise (beginner, advanced, expert) for each topic and each user. We therefore filtered the data for all cases where we had both a valid expertise rating from the user (leaving out the ones were the user indicated they did not understand the topic or it was not in his/her working domain) and for which there was at least one event that had been tracked by APOSDLE. The reason for the latter was that we did not want the data to be too

skewed by the different levels of engagement of the different users. So the model should have a realistic chance to make predictions from the collected data. Analyses were performed on a "per person" and "per topic" level so that each data point corresponds to one person and topic combination. The dataset we used contained 340 such data points of 864 possible ones (6 users times 144 topics). All analyses reported below were performed on this dataset of 340 data points.

We are giving below the results of four different predictive models. All models use the user's self-assessment as a criterion. In a baseline model, we use the two assessments by colleagues as predictors. Although the input data for the human baseline model differs from the other models (colleagues assessed users on a 3-point scale, while the KIE models use all gathered activities as predictors), we still use the colleague assessment model as a baseline to compare all models based on KIE to human judgment. Note that it is likely that humans form judgments about other's expertise in a similar way as APOSDLE, namely by gathering evidence from behaviors and forming a composite judgment.

We then employed three KIE models to compare their predictive potential. First, we employed a model that took only the performed tasks into account. This is in line with CbKST which considers for each task all the skills assigned to that task and assumes the user possesses these skills if the task is performed successfully. In our case, we do not know about the success of task performance, but we assume that the higher the number of detected or selected tasks, the higher the probability that the user can in fact perform the task successfully, and, hence, possess the skills assigned to the task. We call this model *Task KIE model*.

The second KIE model used as predictors the frequency of all events, not only the performed tasks. As can be seen from Table 1, the amount of information this model used is significantly larger. While the KIE task model drew only on 149 events, the *Full KIE model* drew on 1823 events.

Finally, we also computed a model which did not take into account the type of event that was performed for a concept, but rather drew only on total frequency of events for a concept (*Frequency KIE model*). So, for example, if a user had performed one task related to the concept, had looked at two resources and had created one learning path for the topic then the Full KIE model would take these different events into account by using all of them as predictors. The Frequency KIE model instead would only use one predictor for the topic, namely total number of events, in this case 4.

To summarize, the following 4 models were computed

- *Human judgment model* (baseline model): 2 Predictors: two expertise ratings from colleagues
- *Task KIE model*: 1 Predictor: Frequency of KIE relating to the performance of a task
- *Full KIE model*: 9 Predictors: Frequency of KIE for all 9 event types.
- *Frequency KIE model*: 1 Predictor: Frequency of all KIE for a topic, regardless of the type.

4.3.3 Comparison of Linear Regression Models

Table 2 gives an overview of the outcomes of the linear regression to predict self-assessed knowledge levels using the four models. The baseline model (human judgment) reveals a coefficient of R=0.405 and is significant with F(2,534)=53.198, p<.001. These results indicate that, taken together, the two peer assessments predicted the self-assessment to a moderate degree, and predictions were clearly better than chance. The beta weights indicate that both peer judgments contribute significant amounts of variance to the model: $beta_{peer1}$=0.340 (t=8.268, p<0.001), $beta_{peer2}$=0.139 (t=3.370, p=0.001). Results for these peer assessments are in line with correlation coefficients usually obtained in studies that compare self- and peer assessments, for instance [31] and [32] (see below in the discussion section).

As can be seen, two of the KIE models are significant (Task KIE and Full KIE model) while the Frequency KIE model is not. On the one hand, this gives general support to the two approaches we have pursued. As CbKST suggests, looking at the tasks performed at the workplace, it is possible to predict to some extent expertise needed to perform these tasks. Also our second approach to track activities with APOSDLE was successful to predict expertise better than chance.

Moreover, if these two models (tasks only vs. all events) are compared to each other, then the addition of explained variance (the difference in R^2 of .053) turns out to be significant ($F_{change}(8,330)=2.382$, p=0.017). This indicates that the Full KIE model is superior to the Task KIE model in that it adds to the prediction some additional variance.

Table 2: Results of four linear regression models to predict self-assessment of expertise for 144 domain topics

	No. of Predictors	*R*	*F*	*p*
Human judgment model	2	0.405	53.198	<0.001
Task KIE model	1	0.147	7.514	0.006
Full KIE model	9	0.274	2.979	0.002
Frequency KIE model	1	0.071	1.714	0.191

4.3.4 Split Half Validation of Beta Weights

Next, we look at the beta weights of the predictors in the Full KIE model. Because beta weights as such should not be interpreted as signifying their general importance in predicting the criterion, a split half validation was performed. The data set was split into half by random assignment of all cases. Then the linear regression model was calculated for each half and beta weights were derived. This procedure should ensure that the most important beta weights could be identified. Table 3 shows the most important three predictors in each of six iterations and the sign of the beta weight. The signs of the weights in the table show the direction of the prediction where a positive sign indicates that the predictor indicated increased expertise. All signs are in an anticipated direction, except for “Get a Learning Hint” which indicated increased expertise, rather than less.

The table confirms that performing a task and creating a learning path were the two dominant predictors as they appeared among the top three in each of the cross validation datasets. The third predictor varies across the datasets indicating that additional variance cannot be explained in a stable manner.

Table 3: Three strongest predictors and the sign of their beta weight in six split half cross validation samples

Perform Task (+)	Create Learning path (-)	Create Learning path (-)	Create Learning path (-)	Create Learning path (-)	Perform Task (+)
Create Learning path (-)	Perform Topic (+)	Perform Task (+)	Perform Task (+)	Perform Task (+)	Create Learning path (-)
View Resource (+)	Perform Task (+)	Get Learning Hint (+)	View Resource (+)	Select Learning Goal (-)	Get Learning Hint (+)

5 Discussion and Conclusion

We conclude from the results that it is possible to predict to a certain extent self-assessed expertise in a domain by considering events tracked by a work-integrated learning system in a natural use of the system at the workplace. We note further that an approach, based on CbKST which predicts expertise solely from tasks selected is already successful. If we add further events which are assumed to indicate knowledge (KIE), then the prediction can be significantly improved.

The correlation coefficients derived from judgments of peers (one peer, R=0.340; two peers R=0.405) are in line with studies that compare self- and peer assessment in workplace settings. For example, two meta-analyses find correlation coefficients between self- and peer ratings of ρ=0.36 [31] and r=0.37 [32]. Considering this as an upper baseline, we consider the performance of the Full KIE model (R=0.274) as a rather promising result, considering the fact that the KIE model did only have the opportunity to track events for 2 months, and there was a great variation in how frequently APOSDLE was used by the participants.

Our study also clearly shows that it makes sense to look at qualitative differences in the events rather than just count frequencies. There is a good amount of evidence in the behavioral and educational sciences that supports this finding: When compared to novices, experts in a domain don't just do more of the same, but instead their behavior and thinking is qualitatively different than that of novices. Comparing the Task and Full KIE models to a model based on tracking frequencies of events only, clearly shows the superiority of taking into account different types of events. This is also supported by different signs of the predictors when the beta weights were more closely analyzed.

It should be mentioned that we do not intend to suggest that a linear regression model is the most suitable statistical model to predict expertise. Instead, the purpose of the present paper was merely to compare the predictive power of different KIE models in a situation of real workplace behavior. Linear regression models have a wide applicability and they have been shown to be fairly robust, and this is why they

have been our choice in this case. We did employ some alternative regression models (like logistic regression) and found the general conclusions to be the same. We are fairly certain that the predictive accuracy could be improved by employing more sophisticated machine learning models.

Finally, it should be mentioned that the generalizability of these findings needs further research in other domains. Especially, the beta weights clearly need further validation. It is quite likely that the importance of each of the types of events for predicting expertise will vary significantly between domains, or even from person to person. If enough data was available, it might be worthwhile to build a predictive model for each person which from a user modeling point of view may even improve the performance of the algorithm.

Acknowledgements

APOSDLE (www.aposdle.org) has been partially funded under grant 027023 in the IST work programme of the European Community.

6 References

[1] Brusilovsky, P., Millàn, E.: User models for Adaptive Hypermedia and Adaptive Educational Systems. In: Brusilovsky, P., Kobsa, A., and and W. Nejdl (eds.) The Adaptive Web. pp. 3-53. Springer-Verlag, Berlin, Heidelberg (2007).

[2] Ley, T., Kump, B., Albert, D.: A methodology for eliciting, modelling, and evaluating expert knowledge for an adaptive work-integrated learning system. International Journal of Human-Computer Studies. 68, 185-208 (2010).

[3] Lindstaedt, S., Kump, B., Beham, G., Pammer, V., Ley, T., Dotan, A., Hoog, R.D.: Providing varying degrees of guidance for work-integrated learning. In: Wolpers, M., Kirschner, P.A., Scheffel, M., Lindstaedt, S., and Dimitrova, V. (eds.) Sustaining TEL From Innovation to Learning and Practice 5th European Conference on Technology Enhanced Learning ECTEL 2010. pp. 213-228. Springer (2010)..

[4] Lindstaedt, S.N., Beham, G., Kump, B., Ley, T.: Getting to know your user - Unobtrusive user model maintenance within work-integrated learning environments. In: Cress, U., Dimitrova, V., and M., S. (eds.) Learning in the Synergy of Multiple Disciplines: Proceedings of ECTEL 2009, Nice, France, September/October 2009. pp. 73-87. Springer, Berlin, Heidelberg (2009).

[5] Kump, B., Seifert, C., Beham, G., Lindstaedt, S.N., Ley, T.: Seeing what the system thinks you know. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12. pp. 153-157. ACM Press, New York, New York, USA (2012)..

[6] Brusilovsky, P., Kobsa, A., Vassileva, J. (Eds. .: Adaptive hypertext and hypermedia. Kluwer Academic Publishers, Dordrecht (1998).

[7] Poulson, M.C., Richardson, J.J. (eds. .: Foundations of intelligent tutoring systems. Lawrence Erlbaum Associates, Hillsdale, New Jearsey (1988).

[8] Desmarais, M.C., Baker, R.S.J.: A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-adapted Interaction. 22, 9-38 (2011).

[9] Anderson, J.R., Boyle, C.F., Corbett, A.T., Lewis, M.W.: Cognitive modeling and intelligent tutoring. Artificial Intelligence. 42, 7-49 (1990).

[10] Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T.: Cognitive tutor: applied research in mathematics education. Psychonomic bulletin & review. 14, 249-255 (2007).
[11] Mitrovic, A.: Fifteen years of constraint-based tutors: what we have achieved and where we are going. User Modeling and User-Adapted Interaction. 22, 39-72 (2011).
[12] Heller, J., Steiner, C., Hockemeyer, C., Albert, D.: Competence-based knowledge structures for personalised learning. International Journal on E-Learning. 5, 75-88 (2006).
[13] Conati, C., Gertner, A.S., VanLehn, K.: Using Bayesian Networks to Manage Uncertainty in Student Modeling. User Modeling and User-Adapted Interaction. 12, 371-417 (2002).
[14] De Bra, P., Aroyo, L., Cristea, A.: Adaptive web-based educational hypermedia. Web Dynamics: Adapting to Change in Content, Size, Topology and Use. 387-410 (2004).
[15] Limongelli, C., Sciarrone, F., Temperini, M., Vaste, G.: Adaptive learning with the LS-Plan dystem: A field evaluation. IEEE Transactions on Learning Technologies. 2, 203-215 (2009).
[16] De Bra, P., Smits, D., Stash, N.: The Design of AHA! Proceedings of the ACM Conference on Hypertext and Hypermedia. p. 133. Odense, Denmark (2006).
[17] 1. Brusilovsky, P., Schwarz, E., Weber, G.: ELM-ART: An intelligent tutoring system on World Wide Web. In: Frasson, C., Gauthier, G., and Lesgold, A. (eds.) Intelligent Tutoring Systems (Lecture Notes in Computer Science, Vol. 1086). pp. 261-269. Springer, Berlin (1996).
[18] Brusilovsky, P.: KnowledgeTree: A distributed architecture for adaptive E-Learning. WWW 2004, May 17-22, 2004, New York, New York, USA. pp. 104-113 (2004).
[19] Kay, J., Kummerfeld, B., Lauder, P.: Personis: a server for user models. In: De Bra, P., Brusilovsky, P., and Conejo, R. (eds.) Proceedings of AH'2002. pp. 203-212. Springer, Berlin, Heidelberg (2002)..
[20] Augustin, T., Hockemeyer, C., Kickmeier-Rust, M.D., Albert, D.: Individualized skill assessment in digital learning games : Basic definitions and mathematical formalism. IEEE Transactions on Learning Technologies. 4, 138-148 (2011)..
[21] Linton, F., Schaefer, H.-P.: Recommender systems for learning: Building user and expert models through long-term observation of application use. User Modeling and UserAdapted Interaction. 10, 181-208 (2000).
[22] Happel, H.-J., Maalej, W.: Potentials and challenges of recommendation systems for software development. Proceedings of the 2008 international workshop on Recommendation systems for software engineering - RSSE '08, 11. ACM Press, New York (2008).
[23] 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, 31-47 (2008).
[24] Ley, T., Kump, B., Gerdenitsch, C.: Scaffolding Self-directed Learning with Personalized Learning Goal Recommendations. In: De Bra, P., Kobsa, A., and Chin, D. (eds.) Proceedings of the UMAP 2010. pp. 75-86. Springer, Berlin, Heidelberg (2010)..
[25] Brusilovsky, P., Cooper, D.W.: Domain, task, and user models for an adaptive hypermedia performance support system. Proceedings of the 7th international conference on Intelligent user interfaces - IUI '02, 23. ACM Press, New York (2002).
[26] Mislevy, R.J., Riconscente, M.M.: Evidence-centered assessment design. In: Downing, S.M. and Haladyna, T.M. (eds.) Handbook of Test Development. pp. 61-90. Lawrence Erlbaum Associates, Mahwah, New Jersey (2006).
[27] Lepsinger, R., Lucia, A.D.: The art and science of 360 degree feedback. John Wiley & Sons (2009).
[28] 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).

[29] Muellerbuchof, 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).
[30] Wild, F., Haley, D., Bülow, K.: Using latent-semantic analysis and network analysis for monitoring conceptual development. Journal for Language Technology and Computational Linguistics. 26, 9-21 (2011).
[31] Harris, M.M., Schaubroeck, J.: A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personell Psychology. 41, 43-62 (1988).
[32] Conway, J.M., Huffcutt, A.I.: Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. Human Performance. 10, 331-360 (1997).