SELF-EVALUATION OF OPEN ANSWERS AS A BASIS FOR ADAPTIVE LEARNING SYSTEMS

Egon Werlen and Per Bergamin
Institute for Research in Open-, Distance- and eLearning (IFeL)
Swiss Distance University of Applied Sciences (FFHS), Switzerland

ABSTRACT

The continuous measurement of learning behavior or learning outcome by appropriate sensors is fundamental for the implementation of technology-based adaptive learning courses. An adaptive system needs such learning data to adapt the instruction to the needs of students. Depending on the learning objective, the sensors use information generated within tasks as basis for instructional adaptation, such as closed clear answers to learning tasks or more ambiguous open answers. In the second case, the use of self-evaluation concepts is one possibility. However, the automation and corresponding generation of learning data for adaptive systems is difficult because the answers are not clearly correct or incorrect. In a preliminary study we constructed a corresponding self-evaluation scaffold in the context of a Blended-Learning course in health psychology with 60 adult students. The goal of the study was to analyze if self-evaluation delivers valid data for sensors of an intended adaptive learning system, and what influences the accuracy of the self-evaluation. Therefore, we conducted an external evaluation of the students' answers. The correlation of the self-evaluation with the external evaluation was moderate ($r=.50$; $p < .001$), and there was a large overlap of 65.9% between self- and external evaluation. The difference between self-evaluation and external evaluation can partly be explained by the task and subtask used, the difficulty of the task and the quality of the answers. In this respect, these results provide us with initial insights into what must be taken into account when designing corresponding sensors with an open response format. The analyses have also encouraged the development of a technology-based adaptive course on the basis of corresponding sensors.

KEYWORDS

Self-Evaluation, Validity, Adaptive Learning, Sensors, Measurement

1. INTRODUCTION

Technology-based adaptive learning systems require a continuous flow of learning data in order to continuously and effectively adapt learning objects or instructions to the current needs of learners. For this purpose, sensors (measuring methods) must be found which, for example, provide the system with regular data on whether and how well a learner has solved certain tasks. In learning modules that aim at factual knowledge and its reproduction and in which clear answers can be determined, as is often the case in basic courses in natural sciences or vocabulary training in language courses, sensors usually deal with numbers or precisely defined terms. However, at all levels of education, especially university level or professional education, there are often no clear correct or incorrect answers to certain tasks. In such cases the solution can be the use of open answers. Such open answers cause additional challenges for the implementation of adaptive systems.

One way to measure the correctness of open answers is to compare the learners' self-evaluation with given standard answers. On this basis, we have developed a scaffold in the context of processing case vignettes in a module of health psychology, in which students compare their answers and evaluation of counselling procedures and activities with a standard answer and comment on it.

2. THEORETICAL BACKGROUND AND RESEARCH QUESTION

Modern adaptive learning systems are characterized by various adaptive instructional interventions in a dynamic technology-based learning environment. Learning content, navigation and learning support continually change during the learning process that are adapted to the individual requirements of the learners (Wauters, Desmet, & van den Noortgate, 2003).

Wauters et al. (2003) distinguishes three basic factors for adaptation: 1) personal characteristics (stable, situational), 2) content-specific characteristics, and 3) context-based characteristics. Information from all these factors may be used to adapt the learning process. They have to be measured and stored. For the development and implementation of our adaptive system on the basis of open questions and corresponding answers, we rely on a model of Zimmermann, Specht and Lorenz (2005). In this approach the measurements are described as Sensors. The data of the sensors are analyzed and categorized in the Analyzer. It determines the basis of the adaptation (e.g. person, content, context, …). The Controller implements the adaptation mechanism and determines what threshold and in which way the learning process is adapted, i.e. it defines the object of adaptation (content, prompts, form, etc.). Based on the decision of the Controller, the Presenter displays the concrete objects of adaptation.

A first central component are the sensors, their measurement quality strongly influences both the way the system functions and the quality of the entire system. To measure the correctness of the students' open format answers we implemented the above-mentioned self-evaluation concept in form of a scaffold guiding comparison process for the students. The scaffold with the name OPeL (Online Prompting e-Learning) was realized in the standard Learning Management System of our university (Moodle; version 1.9 and 3.2) by using its different question types following a stepwise sequence: a) problem/situation (vignette) presentation b) question, c) student answer, d) standard answer presentation e) students' comparison of the solutions f) self-reflection about differences. For different vignettes, a combination of several of these sequences was possible. Such combinations were mainly arranged to evaluate whole episodes of entire behavioral changes. In the development of the entire adaptive system, steps d), e) and f) provide the learning data for the sensor.Bourke (2014) cites Boud's (1991) definition of self-evaluation as "the involvement of students in identifying standards and/or criteria to apply to their work and making judgements about the extent to which they met these criteria and standards". It's primarily the second part of this definition that fits to our work. The students judge their own answers in comparison to a sample answer using their own standards and criteria for the comparison. The students were not trained to identify their standards or criteria. Research shows rather mixed findings to such self-evaluation approaches. There are authors reporting good results of self-evaluation (Plant, Corden, Mourad, O'Brien, & van Schaik, 2013 for pediatric crisis resource management skills; Bachman, & Palmer, 1989 for communicative language ability), others found an inappropriate validity for self-assessment (e.g. Maderick, 2013 for digital competences). Ross (2006) reports that the validity of self-assessments (i.e. comparison with teacher assessment or tests) gives mixed results. He cites the meta-analysis of Steven Ross (1998) who found a mean correlation of r=.63 within a range of r=.15 to r=.80 (k=60).

The goal of the present investigation is whether the self-evaluation within the implemented scaffold delivers valid data for sensors of the intended adaptive learning system. Further, we want to know what influences the accuracy of the self-evaluation of the answers (under-, and overestimation).

3. METHODS AND HYPOTHESES

In the course Fundamentals of Health Psychology at our university the answers from fife case vignettes used in the years 2014 to 2017 were evaluated. The mean age of all course participants was 39 in a range of 24 to 52 (and one participant 74). 36 students never used any of the proposed tasks. Of the 60 students performing at least one of the tasks 87% were women. 285 processed case vignettes with a total of 791 evaluable answers and the associated self-evaluation were evaluated. In these analyses, we included only the first attempt, if a student executed the same case vignette several times. The self-evaluation was encouraged with the request to "Compare your answer with the sample answer". The students rated their agreement on a ten-point Likert-scale from "complete disagreement" to "complete agreement", with no other indications. Three questions on a nine-point Likert-scale included in the tasks were to measure the cognitive load (Brunken,

Plass, & Leutner, 2003) of the students while completing the tasks. The intrinsic load was measured asking for the difficulty of each subtask ("How high do you estimate the difficulty of the task you just completed?“; "extremely easy" to "extremely difficult"), the extraneous load was asked for with the influence of the learning environment ("How much did the learning environment (presentation, question type, etc.) help you to solve this task?“; "did not help at all" to „was very helpful"), and the germane load was asked for with the mental effort ("How high do you estimate your mental effort to solve the task you have just solved?“; „very little effort“ to „very large effort").

To evaluate the students' self-evaluation of their answers we conducted an external rating by the teacher of all students' answers with the same scale as the students did. Additionally, the quality of the answer (formulation, errors), the students' engagement (elaboration, effort to give correct answer) and the insight contained in the self-reflexion related to the differences between students' answer and sample answer (recognizes differences, understands differences). Further ratings by other collaborators are planned. Quality, engagement, and insight were measured on a three-point scale with the possibility to give half points.

In order to evaluate the self-assessment of students' answers, we conducted on the same scale an external rating of all students' answers by the teacher. In addition, the quality of the answer (formulation, mistakes), the students' engagement (elaboration, effort to give correct answer) and the insight into self-reflection with regard to the differences between the students' answer and the model answer (recognizes differences, understands differences) were checked. Quality, commitment and insight were each measured on a three-point scale with the possibility of awarding points. The following three hypotheses formed part of our explorative analysis: 1) There is a significant correlation between self-assessment and external assessment; 2) The percentage of under- and overestimation does not exceed 50% of students' answers; 3) The difference between self- and external assessment is predicted by the task, the cognitive burden of the task and the characteristics of the students' answers (quality, commitment). Statistical analyses were conducted with R (3.3.4; R-Core-Team, 2017). For the generalized linear mixed-model, we used the lme4 pack- age of Bates et al. (2014), the p-values were calculated by means of Satterthwaite's approximation with lmerTest (Kuznetsova et al., 2017), and the pseudo- R^2 of the fixed effects with MuMln by Nakagawa and Schielzeth (2013).

4. RESULTS

The mean value of all self-evaluations was 5.81 (SD = 3.12). The average value of all external evaluations was 6.34 (SD = 2.86). The correlation between self- and external evaluation was r = .50. The agreement of the self- and external evaluations with a difference of 0 to 2 points was 65.9%. The complete match was 18.2%. There were 15.4% underestimations (including 4.0% with 6 to 9 points lower self-evaluations; high underestimation) and 18.7% overestimations (including 3.2% with 6 to 9 points higher self-evaluations; high overestimation).

We conducted a regression analyses looking for variables that influence the difference between self- and external evaluation. The table 1 shows the results of the regression analysis predicting the difference between self- and external evaluation. The explained variance of the fixed effects is 18.9%. The model is significantly better than the intercept only model (chi^2=150.45, df=9, p<.001). This difference is influenced by the task, the subtask, the difficulty of the subtasks, and the quality of the students' answers. There were five tasks, one of them was applicated only the first year (task 5). The regression analysis predicted for the first, the third (both: case vignettes following the Transtheoretical Model) and the fourth task (Health Action Process Approach) a larger difference (positive values indicating an overestimation) compared to the second task (Health Action Process Approach). For the fifth task (Social Cognitive Theory) a smaller difference was predicted (negative values indicating an underestimation). Looking at the percentages of under- and overestimation confirmed this analysis: The second task (Health Action Process Approach) and the fifth task (Social Cognitive Theory) had with 20% and 32% the highest percentages of underestimation and the lowest overestimation (17% and 7%). The first and the third task (both: Transtheoretical Model) had the lowest underestimation (9% and 5%) and with 27% and 25% the highest overestimation. The fourth task (Health Action Process Approach) was in between (underestimating 11%; overestimation 17%).

Table 1. Full Model (fixed effects) of the prediction of the difference between self-evaluation and external evaluation by task, subtask, difficulty of the subtask and quality of the answer

Predictor	**β**	**CI (95%)**	**P**	
Intercept	2.59	2.10 ; 3.08	< .001	***
Task 1	1.09	0.79 ; 1.38	< .001	***
Task 3	1.21	0.93 ; 1.49	< .001	***
Task 4	0.77	0.48 ; 1.05	< .001	***
Task 5	-1.41	-1.75 ; -1.07	< .008	**
Subtask 2	0.47	0.25 ; 0.70	.036	*
Subtask 3	0.96	0.72 ; 1.19	< .001	***
Subtask 4	0.26	-0.27 ; 0.79	.625	
Difficulty of subtask (self)	-0.26	-0.31 ; -0.21	< .001	***
Quality of answer (external)	-0.88	-1.04 ; -0.73	< .001	***
Observations	703			
N_{VP}	60			

After an introducing text with the case vignette, all but the second task had three subtasks. Over all five tasks, the second and the third subtask had a significantly more positive difference between self- and external evaluation, indicating higher overestimations. The first subtask had the highest underestimation (19% vs. 11% and 12%), the third subtask hat the highest overestimation (26% vs. 17% and 16%). The fourth subtask (existing only on the second task) hat a low underestimation (7%) and a high overestimation 29%. That corresponds to less underestimation and more overestimation at the latter subtasks. The predicted values for each subtask are shown in figure 1.

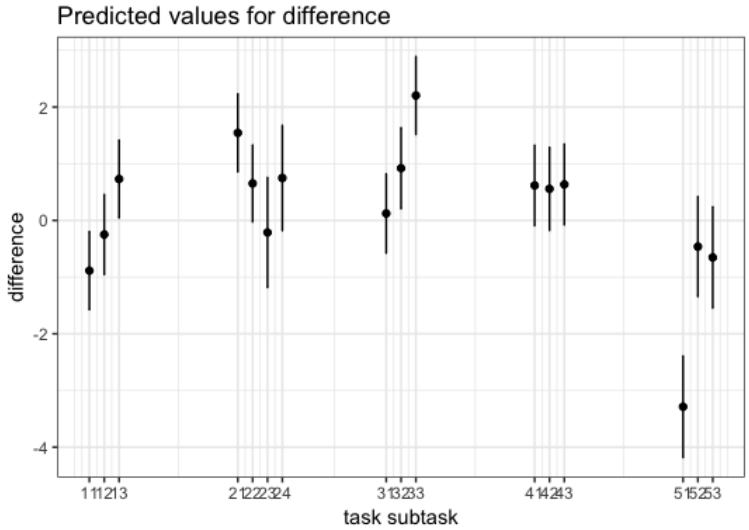

Figure 1. Predicted values for the difference between self-evaluation and external evaluation for all tasks and subtasks (task subtask: e.g. 11 = task 1, subtask 1)

The influence of the difficulty of the text and the quality of the responses was significantly predictive for the difference between self- and external evaluation of the answers. Concerning task difficulty, the easier a task as estimated by the students the higher the overestimation of the task. If the quality of the responses, rated by the teacher, was low, an overestimation was more probable.Other variables showed no influence on the differences between self- and external evaluation or showed high covariation and were therefore excluded.

5. DISCUSSION

In line with our first hypothesis, we found a correlation between self-assessment and external assessment of r=.50 (p<.001). This correlation falls in the range of correlations found in the meta-analysis by Ross (1998) with an average of r=.63 and a median of r=.49. With 65.9% most of the students estimated their answers with a difference of 0 to 2 points compared to the external evaluation confirming the second hypothesis. However, we have also found an underestimation of 15.4% and an overestimation of 18.7%. These values correspond to a formative self- and external assessment of foreign language courses (French, English) reported by Leonzini (2009). She found an underestimation of 13.3% and an overestimation of 20.2% in six classes within three years with a total of 110 students who were attending the degree course in Primary Education and in Social Services at the Faculty of Education of the University of Trieste. Poppi and Radighieri (2009) also report similar results from 20 students assessed in English reading comprehension. They found an underestimation of 15% and an overestimation of 25%.

It may also be noticed, however, that students with higher task difficulties increasingly underestimate the correctness of their answers. For example, Suzuki (2015) reports more misfits with difficult tasks (questions) compared to simpler tasks (but he gives no indication of the direction of the misfits). This contradicts somehow the findings of Leonzini (2015) that "students with a higher degree of competence were more inclined to under-estimation, while the less proficient students over-estimated themselves" (S. 123).

In this context, the main question to be resolved is how to deal with the 15% underestimates and 19% overestimates, especially the high underestimates (4%) and the high overestimates (3%). However, according to our hypothesis, the results of our regression analysis show that self-evaluation depends on the task (and sub-task), the difficulty of the task and the quality of the students' answers. This has an impact on the further development of adaptive learning environments with self-evaluation as sensors. In order to better explain the differences between self-assessments and external assessments of students' answers, we have to evaluate these differences with predefined criteria regarding possible causes for deviations in the next step. This should help to design tasks that reduce the proportion of under- and overestimates. In our first adaptive course with self-assessments as sensors we will divide case vignettes into several parts, stimulate different aspects and analyse the effects. The influence of the externally evaluated quality of the students' answers also points to differences in the motivation of the students. Therefore it is important to control the motivation of the students and to support them in their work.

Some restrictions of the study are still to be pointed out at present. The external evaluations were only made by the teacher of the course. This holds the danger of a bias. Future studies should include further raters. The framework (OPeL) from which the data were taken was not designed to test sensors for adaptive learning courses, but as a scaffold to promote meta-cognitive reflections when answering the tasks and checking the answers . In addition, the framework was implemented in two different Moodle versions (1.9, 3.2), which may have influenced several parameters (e.g. cognitive load, motivation).

6. CONCLUSION

The overlap of self-evaluation and external evaluation of the students' answers of our task with case vignettes is large but there remains a considerable amount of under- and overestimations. To get a good sensor for an adaptive learning system for open answers with a grey area of correctness we need to manage reasons for under- and overestimations. From our findings we got first insights to focus more on task-difficulty and motivation. Thus, to prevent under- and overestimations is to strive for a medium task difficulty and well-built tasks. Several studies mention the need of a training or guidance for the self-evaluation (e.g. Poppi et al., 2015). Bad quality of answers, i.e. students made little effort to formulate a well written answer, gives more overestimations. That may be due to poor motivation or uneasiness with self-evaluation. Therefore, an extended instruction e.g. by some advanced organizers on how to self-evaluate and introduce some relevant criteria might help to get more valid students self-explanations (e.g. Hsu, Lai, & Hsu, 2015; in the context of scaffolds for inquiry learning). Another way to get better matching of self- and external evaluation may be the use of a scale with fewer response options. Therefore, we transformed on a trial basis the data that were collected with a 10-point Likert-scale in a 4-point Likert-scale. The correlation of self- and external evaluation with this 4-level sensor remained nearly identical (r=.48; p<.001). Converted to a 4-point sensor, we estimate 14.5% of the answers with a larger deviation between self-evaluation and external evaluation

(more than 1-point difference). That is 3.4% (n=24) of the answers with an overestimation and 11.1% (n=78) of the answers with an underestimation. Thus, we estimate a 10% higher accuracy.

All in all, we got the impression that it is possible to use the presented self-evaluation scaffold as a sensor in an adaptive system. The study encouraged us to implement the proposed sensors on an adaptive learning system using open answers as a source of adaptation.

ACKNOWLEDGEMENT

Many thanks to Dr. Stéphanie McGarrity of the Institute for Research in Open-, Distance- and eLearning (IFeL) of the FFHS for proof reading and her valuable comments.

REFERENCES

Bachman, L. F. and Palmer, A. S., 1989. The construct validation of self-ratings of communicative language ability. In *Language Testing*, Vol. 6, No. 1, pp. 14-29.

Bannert, M., 2009. Promoting self-regulated learning through prompts. In *Zeitschrift für Pädagogische Psychologie*, Vol. 23, No. 2, pp. 139-145.

Bates, D. et al., 2014. Fitting Linear Mixed-Effects Models using lme4. In *Journal of Statistical Software*. Vol 67, No. 1, pp. 1-48.

Boud, D., 1991. *Implementing student self assessment*. Green guide, no. 5. Higher Education and Development Society of Australasia Incorporated (HERDSA), Sidney.

Brunken, R. et al., 2003. Direct measurement of cognitive load in multimedia learning. In *Educational psychologist*, Vol. 38, No. 1, pp. 53-61.

Maderick, J. A., 2013. *Validity of subjective self-assessment of digital competence among undergraduate preservice teachers.* Dissertation, University of Nevada, Las Vegas. http://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=2942&context=thesesdissertations

Friedrich, H.F. and Mandl, H., 1997. Analyse und Förderung selbstgesteuerten Lernens. In F.E. Weinert & H. Mandl (Eds.). *Enzyklopädie der Psychologie, DII/4, Psychologie der Erwachsenenbildung*. Hogrefe, Göttingen, pp. 237-293.

Hsu, Y. S. et al., 2015. A design model of distributed scaffolding for inquiry-based learning. *Research in Science Education*, Vol. 45, No. 2, pp. 241-273.

Kuznetsova, A. et al., 2017. lmerTest Package: Tests in Linear Mixed Effects Models. In *Journal of Statistical Software.* 82(13).

Leonzini, L., 2009. *Motivation and self-reflection: The ELP as a tool to raise learners' awareness and to enhance self-assessment*. In F. Gori, *Il Portfolio Europeo delle Lingue nell'Università italiana: studenti e autonomia*. EUT Edizioni Università di Trieste, pp. 109-125.

Nakagawa, S. and Schielzeth, H., 2013. A general and simple method for obtaining R^2 from Generalized Linear Mixed-effects Models. In *Methods in Ecology and Evolution*, Vol. 4, pp. 133-142.

Plant, J. L. et al., 2013. Understanding self-assessment as an informed process: residents' use of external information for self-assessment of performance in simulated resuscitations. In *Advances in Health Sciences Education*, Vol. 18, No. 2, pp. 181-192.

Poppi, F. and Radighieri, S., 2009. The role of ELP and self-assessment in effective language learning. In F. Gori, Il *Portfolio Europeo delle Lingue nell'Università italiana: studenti e autonomia.* EUT Edizioni Università di Trieste, pp. 83-89.

Ross, J. A., 2006. The reliability, validity, and utility of self-assessment. In *Practical Assessment Research & Evaluation*, Vol. 11, No. 10.

Ross, S., 1998. Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. In *Language testing*, Vol. 15, No. 1, pp. 1-20.

Suzuki, Y., 2015. Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. In *Language Testing*, Vol. 32, No. 1, pp. 63-81.

Wauters, K. et al., 2010. Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. In *Journal of Computer Assisted Learning*, Vol. 26, No. 6, pp. 549-562.

Zimmermann, A. et al., 2005. Personalization and context management. In *User Modeling and User-Adapted Interaction*, Vol. 15, No. 3, pp. 275-302.