An Evaluation Framework for Adaptive and Intelligent Tutoring Systems

Tiina Lynch, Ioana Ghergulescu
Adaptemy, Ireland
{tiina.lynch, ioana.ghergulescu}@adaptemy.com

Abstract: Evaluation frameworks for adaptive and intelligent tutoring systems have largely focused on their prediction power or user experience. However, neither subjective or objective method alone is enough to assess all the properties of any given system, including effectiveness, efficiency and accuracy. This paper proposes an evaluation framework as well as evaluation recommendations for adaptive and intelligent learning systems. The evaluation framework incorporates objective and subjective measures in terms of learning effectiveness, learning efficiency, system accuracy, satisfaction, ease of use and learner engagement.

Introduction

With the continuous development of technology and exponential growth of the e-learning market, both private and public educational sectors have acknowledged the possibilities of adaptive and intelligent learning systems. These advances offer the chance to extend the learning experience to outside traditional classrooms and laboratories (Looi et al., 2014). Furthermore, these technologies can encourage research skills and interactive learning while taking into account students' individual learning styles. The majority of higher education authorities believe that personalised, adaptive learning could make a positive impact on the field of education. Preliminary research results have shown a link between a reduction in drop-out rates and utilizing e-learning software in education (Forbes, 2014).

While plenty of global interest going into developing new programs to realize technology's full potential, it is important not to get lost into the possibilities without fairly assessing the effectiveness of these learning methods. Adaptive intelligent learning is not aimed at replacing the teachers in the classrooms, but to empower them to teach at a deeper level instead of merely trying to get through the curriculum. Several studies look at the benefits of integrating technologies in the learning process, e.g. improvement in attitudes of both teachers and students and increase in skills (Looi et al., 2014; Song et al., 2012). Furthermore, adaptive intelligent systems can help bridge the gap between low and high achievers (Ghergulescu et al., 2015).

Special attention should be given to the evaluation of adaptive and intelligent system as neither subjective or objective method alone is enough to assess all the properties of any given system, including effectiveness, efficiency and accuracy. We propose combining subjective and objective methods to evaluate the system, its learning effectiveness, learning efficiency, system accuracy and user experience. In addition, the motivation for this paper stems from the potential for personalization through adaptation of intelligent tutoring systems, as they offer a real opportunity to prevent disabled students from missing out on education, which has a knock-on effect on their chances of participating in working life after school or higher education.

The rest of the paper is structured as follows: in section 2, we discuss previous work in terms of measures of learning evaluations previous evaluation frameworks, including qualitative and quantitative methods, and the benefits and fallbacks of objective and subjective data. We give an overview of previous research on the different methods of measuring learning effectiveness, efficiency and user experience. In section 3 we describe the proposed evaluation framework, and offer subjective and objective measures and recommendations for comparison evaluations. Section 4 concludes the paper.

Literature Review

Measures of Learning Evaluation

When we measure learning, we measure an improvement in skills, increase in knowledge or change in attitude (Kirkpatrick, 1998; Giorno et al., 2013). In addition to increase in knowledge and understanding of the topic, the time taken to learn is also a factor in measuring learning effectiveness. As one's learning is a complex construct and difficult to measure in an objective manner (Gosen & Washbush, 2004), it is essential that we clearly define what is it that the students need to learn, i.e. the learning outcome, and also how this will be measured. It is also essential to understand the students' level of knowledge prior to the learning experience and recognize that students have different learning styles and preferences (Ghergulescu et al., 2016). For the educator, the issues of cost and time are deciding factors when it comes to deciding the models of learning methodologies (Gosen & Washbush, 2004).

Quantitative and qualitative methods of measuring of learning effectiveness both have their strengths and weaknesses. Quantitative measures such as multiple choice questionnaires are easier to analyze for statistically significant differences due to their standardization, which can form the basis of change in teaching and learning methodologies. On the other hand, qualitative measures give a deeper insight into the processes and outcomes of a learning experience, but extracting patterns and relationship is more difficult. The same is true for objective and subjective data: objective such as analytics can provide results from statistical analyses, while subjective data, though more difficult to process, can offer further information how the learner or educator experienced the learning process.

Researchers have also used synthetic subjects i.e. simulated learners to study learning effectiveness. Problems arise, however, if simulated learners are over-fitted to the learning environment resulting in unrealistic predictions. Similarly, using volunteers is prone to bias: the sample size if often small, and the group might not reflect the underlying population (Greer & Mark, 2015). Same applies to studies on adaptive e-learning systems (AeLS) in schools: it is important to note students' and primarily teachers' attitudes towards technology overall, and specifically when used in learning.

Learning systems and solutions include different attributes, including learning effectiveness, learning efficiency, learner engagement. It should be considered that each of these parameters have a part to play in forming a proper insight into the learning experience and its usefulness. The problem with many online teaching aids is that even though they might be rated high by the students, they might not actually improve the quality, depth or speed of the learning process. Effectiveness and efficiency, and how they are measured, must be clearly defined, while measuring learning experience from the user perspective, including level of engagement, an important indicator on how likely the user is to use it again.

Evaluating AeLS has largely focused on evaluating the correctness of recommendation algorithms, which is one of several objective methods. While comparing algorithms is statistically accurate and straightforward to analyze, it assumes that the users' behavior during the experiment is not significantly different to another time user might carry out the experiment (Knijnenburg et al., 2012; Shani & Gunawardana, 2011). However, analyzing the correctness of algorithms is cost-efficient and allows large sample sizes to be analyzed offline, while user-centric studies require input from users, making large-scale studies challenging and costly (Shani & Gunawardana, 2011). Furthermore, to focus on user experience as a stand-alone parameter of the system's quality makes the assumption that the feeling of the experience alone is enough to measure its effectiveness and efficiency, which is more often not the case. The users' familiarity with the software and/or hardware is positively correlated with having a positive user experience (Jannach et al., 2016), resulting in potential bias. Other subjective methods of evaluation of the systems themselves include how users find the systems' reliability, security, efficiency and maintainability. Exponentially expanding use of mobile technology and the large number of hardware providers require any system to emphasize portability and compatibility, which are strongly connected to providing a positive user experience.

Existing Frameworks

Several frameworks have been suggested for the evaluation of AeLS. Table 1 summarizes these frameworks, with an overview on the evaluation direction (whether the focus of the evaluation is on the user or the system, learning and training, or usability), its primary attributes (including performance, effectiveness, satisfaction, accuracy, reliability, adaptivity) and whether the methodologies have been subjective or objective, or both. Sottilare et al. (2012) focused on user experience, and subjectively evaluated how users found the system's functional suitability, reliability, security, efficiency, maintainability and portability. Their suggested framework, The Generalized Intelligent Framework for Tutoring (GIFT), was designed to assist in military training in the field, where having a human tutor is unpractical, unsafe and sometimes impossible. It is a good fit for this purpose, with heavy emphasis on security and portability, but lacks objective evaluation methods. Knijnenburg et al. (2012) proposed a framework with a focus on the accuracy of prediction algorithms, incorporating the influence of personal and situational characteristics, and found that objective aspects of the system were subjectively perceived by the user. Castellar et al. (2015) studied the enjoyment and cognitive development of students by comparing a math game and traditional paper exercise. While both objective and subjective evaluation methods were used, the subjects of the study were chosen by volunteers registering their children via the Computer-Aided Registration Tool for Experiments (CORTEX). This poses an issue highly common to studies using volunteers and undermines its subjectivity, as it takes a certain type of person, usually with an interest in the topic of the experiment, to volunteer, not necessarily reflecting the true underlying population and their opinions (Greer & Mark, 2015).

Table 1. Review of existing evaluation frameworks

Reference	Framework	Evaluation Direction	Attributes	Methods
Manouselis et al., 2011	Evaluation of Educational Systems	Learning & training, Affective	Effectiveness, efficiency, satisfaction, drop-out rate, learner contributions	Objective and Subjective
Knijnenburg et al., 2012	Generic User-centric evaluation of recommender system	System, Usability/User Experience	Effectiveness, satisfaction, recommendations, interactions, presentation, capabilities, usability, appeal, trust, privacy concerns	Objective and subjective
Sottilare et al., 2012	Evaluation based on based on ISO 9126-1 and ISO 26010 software quality criteria -as spider points	System	Functional suitability, reliability, usability, security, efficiency, maintainability, portability, compatibility	Subjective
Sottilare et al., 2012	System evaluation	Learning & training	Learner effect, performance	Subjective
Tintarev and Mashoff, 2007	Evaluation of Explanations in Recommender systems	System, Usability/User Experience	Effectiveness, efficiency, persuasiveness, trust, scrutability, transparency, satisfaction	Subjective and Objective
Shani & Gunawardana, 2011	Evaluating recommendation systems	System, Usability/User Experience	Prediction accuracy, satisfaction, coverage, confidence, novelty, serendipity, diversity, utility, risk, robustness, privacy, adaptivity, scalability	Subjective and Objective
Lawless et al., 2010	Evaluation of Adaptive Personalized Information Retrieval	System, Usability/User Experience	Effectiveness, efficiency, satisfaction	Subjective and Objective
Mulwa et al., 2012	Evaluation of end user experience in adaptive technology enhanced learning	System, Learning & training, Affective	System functionality, reliability, usability, efficiency, maintainability, learner collaboration, knowledge acquisition, reflection, engagement	Subjective and Objective
Shi et al., 2013, Shi 2014	Evaluation of Adaptive Personalized Information Retrieval	System, Learning & training	System functionality, learning improvement, system prospect	Subjective
Weibelzahl, 2001	Evaluation of Adaptive Learning Systems	System, Learning	System functionality, learning improvement, knowledge acquisition, adaptivity	Subjective and Objective
Orfanou et al., 2016	Empirical Evaluation of the System Usability Scale	Usability evaluation	System usability scale	Subjective
Jannach et al., 2016	Evaluation of Familiarity as a User Satisfaction component	User experience	Satisfaction, trust	Subjective
All et al., 2016	Evaluation of game-based learning systems	System, Learning	System functionality, learning improvement,	Objective
Castellar et al., 2015	Evaluation of gamification in math vs traditional paper exercise	System, Learning, User experience	System functionality	Subjective

Proposed Evaluation Framework

We propose including several evaluation directions to obtain an insightful, overall framework, the components of which are presented in Table 2. The proposed framework incorporates the following evaluation directions: learning and training; system, user experience; and affective. Within learning and training, we suggest evaluating effectiveness by means of learning improvements, and amount of completed or studied content, and efficiency by means of measuring time it has taken to reach an improvement. The system's accuracy is evaluated looking into how accurate the user model is (i.e., how accurate is the system grating in comparison with standardized tests) and how accurate the recommendations algorithms are. User experience is evaluated by how easy the users find the system, and their level of satisfaction. Finally, affective-related evaluation will be performed through engagement and motivation evaluation; how engaged the learners are, both in and out of class.

Table 2. Proposed evaluation framework

Evaluation Direction	Attribute	Description	References
Learning and training	Effectiveness	Learning improvements (with and without revisions); Amount of completed, or studied content (in comparison with other learning instructions)	Manouselis et al., 2011; Kirckpatrick's 1959; Sottilare et al., 2012; Mulwa et al., 2012: Pane et al., 2014; Greer and Mark 2015; Huang et al., 2016
Learning and training	Efficiency	How efficient is the use of the time	Manouselis et al., 2011
System	Accuracy	How accurate the user model is (i.e., how accurate is the system grading in comparison with their tests/ exams) and how accurate the recommendation algorithms are (higher accuracy scores or lower predictive errors)	Shani & Gunawardana, 2011; Mulwa et al. 2011
Usability/User Experience	Ease of use and Satisfaction	How easy to is to use the system	Tintarev and Mashoff, 2007; Lawless et al., 2010; Knijnenburg et al., 2012; Manouselis et al., 2013
Affective	Engagement, motivation	How engaged are the learners both in class and out of class	Ghergulescu, 2013; Cocea and Weibelzahl, 2011

We suggest incorporating both objective and subjective methods for all evaluation directions: learning and training, system, user experience and affective. These are presented in Table 3, with their levels of attributes and specific metrics broken down.

Table 3. Recommendations for Subjective and Objective Measures

Evaluation Direction	Attribute	Metrics	References
Learning and training	Effectiveness	*Objective:* amount of completed, or studied content objects during a learning session, improvement of response quality, effect of adaptive strategies on performance phase; reduced numbers of learners that drop out during the learning phase; the knowledge gain; how the knowledge gained was applied/implemented in real life; expectations achievements (e.g. meeting expectation, below expectation, above expectation); knowledge acquisition; amount of requested materials	Manouselis et al., 2011, Sottilare et al., 2012; Mulwa et al., 2012
		Subjective: score of questionnaires that include questions regarding learning outcome improvement	Shi et al., 2013, Shi 2014
Learning and training	Efficiency	*Objective:* time needed to reach the learning phase, time needed to achieve the learning goal; duration of interaction, number of navigation steps, task success, response time	Manouselis et al., 2011
		Subjective: score of questionnaires that include questions regarding efficient use of time	
System	Accuracy	*Objective*: Normalized Distance-based Performance Measure; utility based ranking; prediction accuracy	Shani & Gunawardana, 2011; Mulwa et al. 2011
		Subjective: score of questionnaires that include questions regarding metrics (especially grades) accuracy	
Usability/User Experience	Ease of Use and Satisfaction	*Objective*: Navigation patterns	
		Subjective: questionnaire regarding Satisfaction, Experience, Ease of use, Familiarity, Quality, Useful	Knijnenburg et al., 2012; Manouselis et al., 2013;
Affective	Engagement	*Objective*: TimeOnTask, NumberRepeatTask (sameTask), NumberMistakes, NumberHelpRequest, navigation behavior	Ghergulescu, 2013; Cocea & Weibelzahl, 2011
		Subjective: score of questionnaires that include questions regarding engagement and motivation	Knijnenburg et al., 2012; Mulwa et al., 2012; Spector 2014; Ghergulescu, 2013

For learning and training effectiveness, the objective metrics will include measuring the amount of content studied or completed during a learning session; improvement of response quality, knowledge gain; reduction in the number of drop-outs; expectations achievements etc. The subjective metrics will involve questionnaires, which will tell us how the users themselves feel they have improved. For learning and training efficiency, the objective metrics will focus on the time spent in learning and task success, while subjective metrics will be based on questionnaires on the users' views on the efficient use of time.

The system's accuracy will be objectively evaluated by normalized distance-based performance measure, utility based ranking and prediction accuracy, based on algorithm performance. The subjective metrics include questionnaires to the users and how accurate do they find the grading system. User experience will be objectively evaluated by analyzing the navigation events of the user, as this provides information on the individual user's journey, and subjectively with questionnaires regarding satisfaction, how they found the experience; was the software easy to use and of good quality. The engagement of users will be evaluated for affective, and the objective metrics will be based on time spent on a specific task, how many times a user repeated the same task, the number of mistakes they made or requested help, and navigation behavior. Questionnaires for subjective metrics will include questions regarding engagement and motivation.

Recommendations for Comparison Evaluation

To show how adaptive, intelligent learning system can truly enhance learning, it is important to compare its results to learning using traditional methods. We suggest using a mixed model, where two groups with similar characteristics such as numbers, age, gender, technology-orientation and level of knowledge, are given the same subjective pre-test and post-test questionnaires, concepts, and subject-specific tests. Attention should be given so that the two groups of students will have similar demographic characteristics, similar motivation and perception about the subject and simila knowledge. In the first period, group 1 will be learning with AeLS, and group 2 will learn without; in period 2, group 2 will use AeLS and group 1 will learn without.

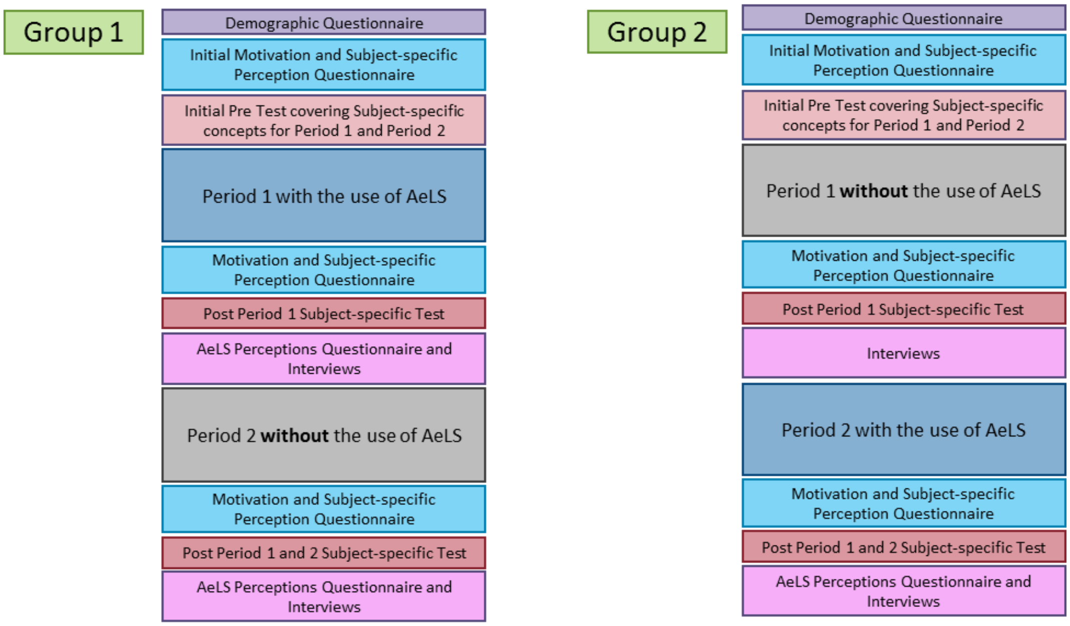
Group 1
Demographic Questionnaire
Initial Motivation and Subject-specific Perception Questionnaire
Initial Pre Test covering Subject-specific concepts for Period 1 and Period 2
Period 1 with the use of AeLS
Motivation and Subject-specific Perception Questionnaire
Post Period 1 Subject-specific Test
AeLS Perceptions Questionnaire and Interviews
Period 2 **without** the use of AeLS
Motivation and Subject-specific Perception Questionnaire
Post Period 1 and 2 Subject-specific Test
AeLS Perceptions Questionnaire and Interviews
Group 2
Demographic Questionnaire
Initial Motivation and Subject-specific Perception Questionnaire
Initial Pre Test covering Subject-specific concepts for Period 1 and Period 2
Period 1 **without** the use of AeLS
Motivation and Subject-specific Perception Questionnaire
Post Period 1 Subject-specific Test
Interviews
Period 2 with the use of AeLS
Motivation and Subject-specific Perception Questionnaire
Post Period 1 and 2 Subject-specific Test
AeLS Perceptions Questionnaire and Interviews

Figure 1. Comparison Evaluation model.

Conclusions

Technology enhanced learning is widely seen as the future of education, with personalised learning journeys for each student through adaptive software, taking into account the uniqueness of every student. These advances are not to replace the teacher in the classroom, but rather provide them with tools that will enable them to teach students at different levels, while keeping everyone engaged and up to speed. Several frameworks for evaluating technological learning systems are widely available but none seem completely comprehensive regarding both subjective and objective methodologies. Here, we propose an insightful, overall framework, that combines a variety of evaluation methods that have been found useful in previous research for the evaluation of adaptive and intelligent system.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement no. 688503.

References

All, A., Castellar, E.P.N., Van Looy, J. 2016. Assessing the effectiveness of digital game-based learning: best practices original research article. *Computers & Education* 92 90-103

Castellar, E. N., All, A., de Marez, L., Van Looy, J. 2015. Cognitive abilities, digital games and arithmetic performance enhancement: a study comparing the effects of a math game and paper exercises. *Computers & Education* 85 123-133

Cocea, M., Weibelzahl, S. 2011. Disengagement detection in online learning: validation studies and perspectives. *IEEE Transactions on Learning Technologies* 4 **2** 114-124.

Forbes. 2014. Rethinking Higher Ed: A Case for Adaptive Learning - Forbes. Retrieved from http://www.forbes.com/sites/ccap/2014/10/22/rethinking-higher-ed-a-case-for-adaptive-learning/

Ghergulescu, I. 2013. Automatic Non-Disturbing Motivation Monitoring in Game-based E-learning through Player Behaviour and EEG (Vol. PhD thesis).

Ghergulescu, I., Flynn, C., O'Sullivan, C. 2016. Learning effectiveness of adaptive learning in real world context. In Proceedings of EdMedia: World Conference on Educational Media and Technology 2016: 1391-1396. Association for the Advancement of Computing in Education (AACE).

Ghergulescu, I., Flynn, C. & O'Sullivan, C. 2015. Adaptemy science: adaptive learning for science for next generation classroom. In Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2015: 1477-1482. Chesapeake, VA: Association for the Advancement of Computing in Education (AACE).

Giorno, R., Wolf, W., Hindmarsh, P.L., Yule, J.V., Shultz, J. 2013. Using scientific abstracts to measure learning outcomes in the biological sciences. *Journal of Microbiology & Biology Education* 14 **2** 275-276.

Gosen, J., Washbush, J. 2004. A review of scholarship on assessing experiential learning effectiveness. *Simulation Gaming* 35 **2** 270-293.

Greer, J., Mark, M. 2015. Evaluation Methods for Intelligent Tutoring Systems Revisited. *International Journal of Artificial Intelligence in Education* 26 **1** 387–392

Huang, X., Craig, S. D., Xie, J., Graesser, A., Hu, X. 2016. Intelligent tutoring systems work as a math gap reducer in 6th grade after-school program. *Learning and Individual Differences* 47 258–265

Jannach, D., Lerche, L., Jugovac, M. 2016. Item familiarity as a possible confounding factor in user-centric recommender systems evaluation. Retrieved from https://pdfs.semanticscholar.org/4127/4d0e1c3b326bb8bfc4d4413ca68decce8270.pdf

Kirkpatrick, D.L. 1959. Techniques for evaluation training programs. *Journal of the American Society of Training Directors* 13 21-26

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22 **4-5** 441–504

Lawless, S., O'Connor, A., Mulwa, C. 2010. A proposal for the evaluation of adaptive personalized information retrieval. *Proceedings of the CIRSE 2010 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation held in conjunction with ECIR-2010 - European Conference on Information Retrieval.* England.

Looi, C-K., Wu., L., Seow, P.S.K., Soloway, E. 2014. Implementing mobile learning curricula in a grade level: Empirical study of learning effectiveness at scale. *Computers and Education* 77 101–115

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., Koper, R. 2011. Recommender systems in technology enhanced learning. In L. Rokach et al. (Eds.), *Recommender systems handbook: A complete guide for research scientists & practitioners* 387-415. Berlin, Germany. Springer.

Manouselis, N., Drachsler, H., Verbert, K., Duval, E. 2013. *Recommender Systems for Learning*. Springer.

Mulwa, C., Sharp, M., Wade, V. 2011. The evaluation of adaptive and personalised information retrieval systems: a review. International Journal of Knowledge and Web Intelligence 2 **2-3** 138–156

Mulwa, C., Lawless, S., O'Keeffe, I., Sharp, M., Wade, V. 2012. A recommender framework for the evaluation of end user experience in adaptive technology enhanced learning. *International Journal of Technology Enhanced Learning* 4 **1-2** 67-84

Orfanou, K., Tselios, N., Katsanos, C. 2016. Perceived usability evaluation of Learning Management Systems: empirical evaluation of the system usability scale. 16 **2** 227-246

Pane, J. F., Griffin, B. A., McCaffrey, D. F., Karam, R. 2014. Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis* 36 **2** 127-144

Shani, G., Gunawardana, A. 2011. Evaluating Recommendation Systems. *Recommender Systems Handbook*. 257-297. Springer.

Shi, L. (2014). Defining and evaluating learner experience for social adaptive e-learning. Retrieved from http://wrap.warwick.ac.uk/62796/

Shi, L., Awan, M. S. K., & Cristea, A. I. 2013. Evaluation of social personalized adaptive E-Learning environments: end-user point of view. Retrieved from http://wrap.warwick.ac.uk/56217/

Song, Y., Wong, L-H., Looi, C-K. 2012. Fostering personalized learning in science inquiry supported by mobile technologies. *Educational Technology Research and Development* 60 **4** 679–701

Sottilare, R.A., Brawner, K.W., Goldberg, B.S., Holden, H.K. 2012. The Generalized Intelligent Framework for Tutoring (GIFT). Concept paper released as part of GIFT software documentation. Orlando, FL: US Army Research Laboratory–Human Research & Engineering Directorate (ARL-HRED)

Spector, J.M. 2014. Conceptualizing the emerging field of smart learning environments. *Smart Learning Environments* 1 **2** 1-10

Tintarev, N., Mashoff, J. 2007. A survey of explanations in recommender systems. Data Engineering Workshop. *IEEE 23rd International Conference* 801-810. IEEE.