Application of Learning Curves for Didactic Model Evaluation: Case Studies

http://dx.doi.org/10.3991/ijet.v8iS1.2357

F. Mödritscher[1], M. Andergassen[1], E.L.-C. Law[2] and V.M. García-Barrios[3]
[1] Vienna University of Economics and Business, Vienna, Austria
[2] University of Leicester, Leicester, United Kingdom
[3] Carinthia University of Applied Sciences, Villach, Austria

***Abstract*—The success of (online) courses depends, among other factors, on the underlying didactical models which have always been evaluated with qualitative and quantitative research methods. Several new evaluation techniques have been developed and established in the last years. One of them is 'learning curves', which aim at measuring error rates of users when they interact with adaptive educational systems, thereby enabling the underlying models to be evaluated and improved. In this paper, we report how we have applied this new method to two case studies to show that learning curves are useful to evaluate didactical models and their implementation in educational platforms. Results show that the error rates follow a power law distribution with each additional attempt if the didactical model of an instructional unit is valid. Furthermore, the initial error rate, the slope of the curve and the goodness of fit of the curve are valid indicators for the difficulty level of a course and the quality of its didactical model. As a conclusion, the idea of applying learning curves for evaluating didactical model on the basis of usage data is considered to be valuable for supporting teachers and learning content providers in improving their online courses.**

***Index Terms*—distance learning; didactical model; self-assessment quizzes; feedback; learning curves; case study.**

I. INTRODUCTION

Assessment is one of the crucial aspects of designing learning environments [1], because it can provide opportunities for feedback and revision, rendering learning processes to be congruent with learning goals. The two main types of assessment are: *formative* - the evaluation of learning and the provision of feedback in terms diagnostic information so as to improve teaching and learning (e.g. teachers' comments on the work in progress), and *summative* - the assessment of learning in order to determine what students have learned at the end of an educational unit (e.g. exams). Besides grading learners due to certification purposes, the assessment of learning is highly related to evaluating and improving educational technology and didactical models.

In this paper, we attempt to evaluate learning outcomes as well as the underlying didactical models in a summative way and on the basis of data generated by users while interacting with the learning materials provided by a Learning Management System (LMS). Specifically, we make use of 'learning curves', as introduced by Martin et al [2], to measure the effectiveness of adaptive educational technologies (and models), and to evaluate distance learning methods which are based on self-assessment quizzes. Overall, we aim at showing that learning curves can be applied to assess the performance of didactical models implemented with educational technologies.

In the following section we present an overview of the evaluation of educational technology, briefly explaining the idea of learning curves and sketching the related work from the field of Human Computer Interaction (HCI). Then, in Section III, we describe a didactical model of a distance learning method, namely the 'Quiz World Cup' [3], and summarize the findings on its realization in practice. In Section IV we revisit this didactical model and analyze it on the basis of learning curve plots created from the data-set of a former study. Furthermore, in Section V we report another study which was conducted to examine the effects of different types of feedback in self-assessment tasks. Section VI outlines possible application areas and discusses problematic aspects of our research. Finally, in Section VII we discuss the findings and conclude with implications for future work.

II. EVALUATION METHODS, LEARNING CURVES AND RELATED WORK

In order to evaluate personalized environments and adaptive learning models, Iqbal and colleagues [4] classify adequate methods for Intelligent Tutoring Systems (ITS) along two dimensions: (1) the applicability for internal and external evaluation, i.e. testing components of an ITS vs. considering the whole system, and (2) methods for experimental research (i.e. methods for systematically varying independent variable(s) and measuring the dependent ones) and exploratory research (i.e. methods involving in-depth studies of an ITS in a natural context and using multiple sources of data). Basically, Iqbal et al [4] argue that measuring the effectiveness of ITSs requires determining if the whole system or only components should be evaluated and if it is possible to systematically manipulate variables and measure the effects on the users.

Drawing conclusions for the **evaluation of didactical models**, the aforementioned set of ITS evaluation methods can also be applied to measure the effectiveness of educational software. As quizzes (e.g. multiple-choice questions, matching questions, true/false questions, short answer questions etc) enable assessing the learning progress, data on user performance in online exams and self-assessment exercises can be seen as a valuable source for validating the effectiveness of instructions. Methodologically, this kind of **measurement** requires methods for the evaluation of components of a learning platform as well as for exploratory research.

However, making use of the data-sets which result from learners interacting with educational platforms requires defining what to measure and how to interpret these measurements. In mathematical logic, **interpretability** is a relation between formal theories that expresses the possibility of translating one into the other [5]. In mundane usage, to interpret is to explain the meaning of (information or actions) [6]. Data interpretation can be highly objective such as body temperature or highly subjective such as a piece of artwork. Nonetheless, such subjectivity (or objectivity) is not all-or-none but a continuum.

Referring to self-assessment tests, the performance of learners can be measured on the basis of the scores received in the quizzes. Although the nature of this data seems to be very objective it can be highly dependent on the circumstances and situation in which learners conduct the quizzes and produce the interaction data. On the other hand and with respect to the variety and dynamics of interactions in learning environments, the focus of measurements and interpretability could be set to the **error rates in learning tasks**, i.e. the discrepancy between the possible and the current achievements in self-assessment activities (e.g. maximum score vs. result on the latest attempt). Hereby, quizzes enable the measurement of scores and errors of learners.

In this context, Martin et al [2] proposed to use so-called '**learning curves**' to analyze and improve systems and models for personalized adaptive learning. A learning curve is a plot of the performance on a task versus the number of opportunities to practice either on an individual or a group basis (see Figure 1). Accordingly, Martin and his colleagues measure the errors on performing tasks in relation to the number of interactions with a system and state that, with reference to the 'Power Law of Practice' [7], this error rate should follow a power law distribution on repeating a task. The learning curve of a learner group is generated by calculating the mean of all students' error rate for each attempt.

Consequently, this approach can be used to compare different implementations of adaptive behavior in learning systems and tasks addressing different skills. In those studies [2], the focus is set on the shapes of the learning curves, i.e. the slope of the curve, the y-axis intercept (error rate for the first attempt), and the fit of the curve (measurement of the deviation from a power law distribution). Going beyond the measurement of adaptive educational technology, we believe that the concept of learning curves could be useful for evaluating pedagogical and didactical models which are realized in the form of software solutions and can be used by learners.

Similar approaches can be identified in the field of HCI in which recent shifts from usability to user experience (UX) and from emphasizing pragmatic goals to hedonic goals have been observed [8]. For measuring usability, HCI professionals often apply three well-established metrics (i.e. effectiveness, efficiency, satisfaction; ISO 9241-110:2010). In contrast, it is proved more challenging in quantifying UX, given the fuzziness of affective quality attributes such as fun, happiness and boredom. As such quality attributes are highly dynamic [9] and context-dependent, researchers posit that their changes over time can more reliably and validly be represented by curves. The tools *iScale* [10] and its variant *UX curve* [11] are used to capture users' evolving experiences with an interactive product/service over a longer-term duration (cf. episodic usability evaluation). Nonetheless, these so-called experience curve approaches are challenged for their validity and reliability, because of the inherent subjective nature of self-reporting. More debatable is the use of the Day Reconstruction Method [12], which is susceptible to memory lapse and even fabrication.

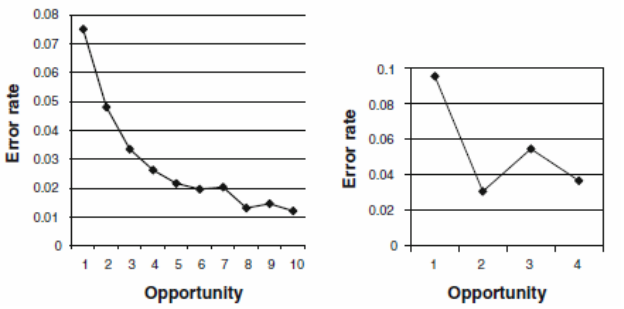

Figure 1. Two sample learning curves [2], one for an individual student (right-hand side) and the other one for a group of learners (left-hand side)

While experience curves, as the name suggests, focus on the emotional and affective aspects of interaction, learning curves address the performance and behavioral aspects, which are more quantifiable and conducive to being represented graphically. However, a crucial issue of learning curves is the definition of error rate and the collection of a sufficiently large data-set in order to plot learning curves and derive valid findings from them. Thus, we briefly revisited a distance learning approach developed a few years ago and then re-evaluated this method and the underlying didactical model through learning curves.

III. MOTIVATION FOR AND DIDACTICAL MODEL OF THE QUIZ WORLD CUP

In 2005 and 2006, courses at the FH Campus 02 (in Graz, Austria) were didactically enriched through the introduction of blended learning elements. Among others, Mödritscher reported on the development of the 'Quiz World Cup' [3], a teaching strategy for virtual classes "*aiming at mediating cognitive subject matter, but also addressing motivational aspects by means of interactive learning content and a game-based element, namely a competition*" ([3], p. 4306]. Traditionally, online courses on cognitive subject matter are prone to a number of potential problems, including (a) a focus on (low-level) cognitive learning objectives leading to rote memory on the part of the learner, (b) the lack of diversity of teaching methods, (c) the lack of support for meta-cognitive skills (such as self-directed learning and social competences, etc.), and (d) passive learning.

A study by Mödritscher [3] describes the implementation of the Quiz World Cup as part of an online course within the open-source e-learning platform Moodle. Accordingly, the distance learning unit consisted of three phases with each addressing a specific set of educational objectives and containing high-level learning materials, self-assessment quizzes and a competition including a bonus point system.

Table 1 characterizes the three phases of the online course according to the educational objectives classified by the Bloom Taxonomy [13]. The objectives indicate that the online course deals with the basics of information

technology (i.e. on the topic 'document formats'), as most of the goals are from the cognitive domain. Amongst others, students have to learn about fundamentals of digital images, audio formats, videos as well as textual documents. Next to the cognitive educational objectives (marked with K1 to K3 in Table I), the online course requires mastering two hand-on skills (e.g. applying compression algorithms on arbitrary data) as well as one attitude (i.e. the consideration of given citation rules).

According to the literature (e.g. [14]), online examinations (e.g. quizzes) are useful for assessing the students' performance on low-level educational objectives. For plotting learning curves, however, this distance learning unit can be considered as a valid setting because each phase represents an item clearly defined by the given objective(s), thereby providing a unified assessment method that can be practiced by students multiple times.

To enhance variation within the self-assessment quizzes, different scoring modes within the three phases were implemented and each defined educational objective was assessed through one question randomly assigned from a question pool for this objective. As a motivational element, the 32 students participating in the study were told that the top-10 performers in the self-assessment quizzes (in terms of scores earned and time spent) would receive a bonus. As will be seen in the next section, the competitive element has advantages for plotting learning curves because students are encouraged to practice each item several times, which was observed in the study described in [3].

The three online learning phases were run over a period of two months. Each phase differed in terms of learning objectives and the mode and scoring in the self-assessment quizzes (see Table I and lines 1 to 4 in Table II). The intermediate results of the self-assessment quizzes were regularly posted in the Moodle course. In addition, the distance learning experiment was evaluated in various ways, such as a post-questionnaire, a final examination or the analysis of Moodle data (log-file, course statistics).

Table II (lines 5 to 10) gives an overview of the students' achievements in the three phases of virtual, self-directed learning. A detailed analysis of these results is given in [3]. As a conclusion, we can state that the 'Quiz World Cup' worked well in the context of the case study, which was a higher education course on the basics of ICT (Information and Communication Technologies), and that the didactical model can be considered valid and efficient.

TABLE I.
CHARACERIZATION OF THE ONLINE COURSE BASED ON THE BLOOM TAXONOMY

Educational objective*	Domain, Level	Phase
1. Overview about scientific working	K1	1
2. Valuing given citation rules	A3	1
3. Comparing layout- and structure-oriented formats	K2	1
4. Overview about text-oriented formats	K1	1
5. Reasoning facts of text-oriented formats	K3	1
6. Explaining color models	K2	2
7. Overview about halftone images	K1	2
8. Explaining compression algorithms	K2	2
9. Applying compression algorithms	S3	2
10. Comparing graphical formats	K2	2
11. Overview about digital audio	K1	3
12. Overview about digital video	K1	3
13. Designing an information system for different document formats	S3	3
14. Reasoning the application of document formats in information systems	K3	3

Note: These objectives are classified as *cognitive* domain aka (**k**)nowledge, *psychomotor* domain aka (**s**)kill and *affective* domain aka (**a**)ttitude, followed by a level

TABLE II.
CHARACTERISTICS AND STATISTICS OF THE EXAMS IN THE THREE ONLINE LEARNING PHASES

Characteristics of exams	Phase 1	Phase 2	Phase 3
Duration of exam	10 min.	12 min.	6 minutes
Difficulty of exam	medium	hard	easy
Scoring mode of exam	best attempt	avg. score	avg. score, 3 attempts
Score for overall course	10%	12%	8%
Statistics of exams	**Phase 1**	**Phase 2**	**Phase 3**
Overall attempts	200	173	52
Max. attempts (by one student)	42	49	3
Min. time (best student)	0:48	6:04	2:02
No. attempts [$\bar{x}$ /σ]	6.3/8.0	5.4/8.7	1.6/0.8
Score (all attempts) [$\bar{x}$ /σ]	8.6/1.7	9.9/2.0	7.4/0.9
Score (rated attempts) [$\bar{x}$ /σ]	9.5/0.8	9.5/1.4	7.6/0.6

Note: $\bar{x}$ is the mean value, σ the standard deviation.

IV. RE-EVALUATION OF THE 'QUIZ WORLD CUP' USING LEARNING CURVES

In this section, we revisit the experiment described in the previous section and apply the concept of learning curves which was introduced by Martin et al [2] to measure and improve adaptive behavior in learning systems. Nonetheless, in our case, we use learning curves to evaluate the didactical model of a distance learning method. The 'Quiz World Cup' is a good example of users interacting with learning technology in different ways and several times in order to foster competence development.

As shown in Table II, the three online learning phases differ in terms of learning objectives, the difficulty level of content, the scoring mode of the self-assessment quizzes, and the score to reach. Thus, in the first and second phases, students had many interactions with the online exam whereas in the third phase they were restricted to a maximum of three attempts due to the examination mode. The learning objectives seem to have an impact on the minimum time and the results of the exams while the difficulty level and the examination mode influence the scoring (see the difference between the average score of all attempts and the average score of the attempts which were rated according to the examination mode).

In order to plot learning curves, we calculated the error rate of a user interacting with one exam in the following way:

$$error\ rate = \frac{score\ to\ reach - actual\ score}{score\ to\ reach}$$

Moreover, the error rate for the n[th] attempt (opportunity) is defined as the mean of the error rates of the n[th] attempt of all learners:

$$error\ rate_n = \frac{\sum_{i=1,\dots,m} error\ rate_{learner_i, attempt_n}}{no.attempts_n}$$

Given the data from the case study conducted in the year 2006, Figure 2 shows the plots of the arithmetic series of the error rates for the three online learning phases, whereby the curve for the third phase is restricted to three attempts due to the examination mode. As shown by the plots the three curves seem to follow a power law distribution, thus evidencing the existence of the 'Power Law of Practice' manifested in [7].

For the approximation of the power law distributions we used the error rates of all attempts. As also defined in [2], the formula for a power law is:

$$T(x) = Bx^{-\propto}$$

Consequently, we received the parameter *B* (y-axis intercept; error rate at *x=1*) and estimated the power law slope α of the three curves according to a maximum likelihood estimation and by using R, an open source software for statistical computing and graphics (cf. http://cran.r-project.org; *power.law.fit* function of the 'igraph' package). Thereby, the maximum likelihood estimator (MLE) which is implemented in the igraph package is based on a function to calculate the negative log-likelihood.

In order to summarize the discrepancy between the observed values and our assumption of the validity of the 'Power Law of Practice', we tried to formalize the goodness of fit of the curve. With respect to regression analysis we calculated the coefficient of determination (R^2; termed goodness of fit from now on) as an indicator for the reliability of the approximation:

$$R^2 = 1 - \frac{\sum_{i=1,\dots,n} (error_i - T(i))^2}{\sum_{i=1,\dots,n} (error_i - \overline{error})^2} \text{ where}$$
$$\overline{error} = mean\ of\ error\ rates \text{ and } n = no.attempts$$

For the learning curves of the three learning phases we estimated the parameters and goodness of fit for the power law approximation as given in Table III.

The characteristics of the curves (see Figure 3 and Figure 4) allow deriving inferences on or a comparison of the characteristics of didactical models of educational units. For instance, the difficulty levels of the three learning phases, which were estimated by the teacher, have been validated by the initial error rates in the three learning curves. For a complex learning content and a hard exam, the learning curve starts with a high error rate (phase 2: B_2=*34%)* while easier exams have lower initial error rates (phase 1: B_2=*28%*, phase 3: B_3=*9%*).

The steepness (α_1, α_2, α_3) and length of the three learning curves are clearly influenced by the examination mode. Ideally, the error rate should tend towards zero if the didactical model of an instructional unit is well designed. While the first phase (best attempt; see Figure 3) shows that the curve approaches the x-axis with various fluctuations in the error rate, the minimal-error scoring made in the second phase (the average score over all

TABLE III.
APPROXIMATION OF THE LEARNING CURVES FOR EACH DISTANCE LEARNING PHASE

Parameter	Phase 1	Phase 2	Phase 3
Power law slope (α)	0.4621	0.4469	1.3120
Error rate at x=1 (B)	0.28	0.34	0.09
Goodness of fit (R^2)	0.5794	0.6070	0.7816

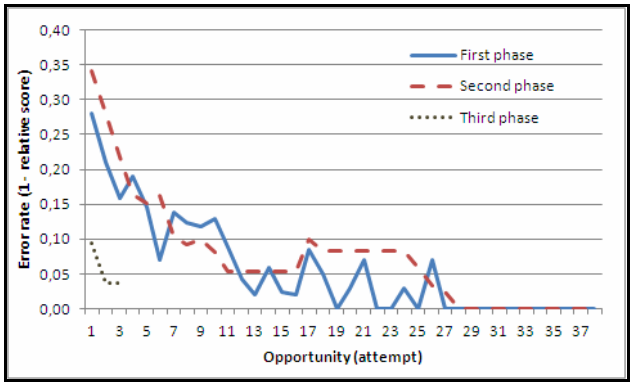

Figure 2. Learning curves for the three 'Quiz World Cup' phases

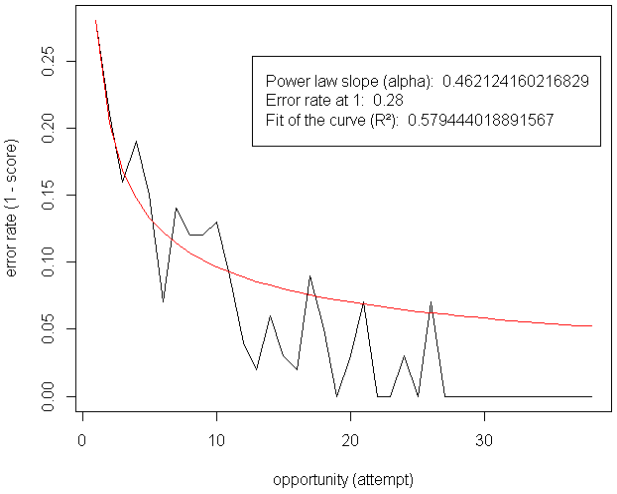

Figure 3. Learning curve and power law approximation for the first distance learning phase

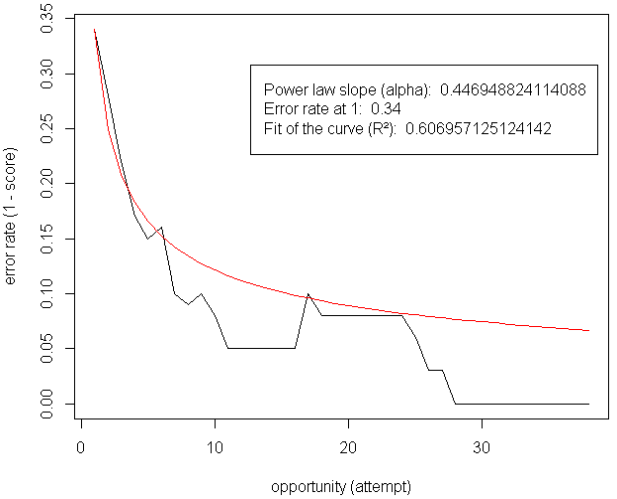

Figure 4. Learning curve and power law approximation for the second distance learning phase

attempts) has only one relapse in the error rate before the curve slowly declines down close to the x-axis. The examination mode of the third phase (the average score of the first three attempts) motivates students to minimize wrong answers and the exam duration at the same time.

Due to the large amount of attempts, the first two curves (Figure 3 and Figure 4) are smoother and not that steep, whereby the learning curve of the first distance learning phase has many unexpected leaps in later attempts due to the possibility to practice by trial-and-error. The learning curve generated for the third learning phase is steeper despite starting with a lower initial error rate. However, this curve only consists of three attempts due to the examination mode, so it is not possible to derive valid conclusions from this curve.

An analysis of the goodness of fit (cf. Table III) shows that the power law approximation of the error rates increases with each learning phase. This observation is backed up by the examination mode again, because the students dared to have more attempts in the first phase (trial-and-error) whereas they were very careful to avoid getting low scores in the second self-assessment quiz, because this would have required higher efforts to correct their results. Nevertheless, only a few students completed the online exam more than 10 times longer (in terms of minutes) than they did in the first two phases. In the third learning phase there were many interactions for each of the three attempts recorded, so the approximation fits the error rates well. Basically, we observed a positive correlation between the number of available scores (error rates) per attempt and the fit of a power law distribution (cf. goodness of fit in Table III) – the more the interactions are available the better the quality of the approximation.

In summary, it can be stated that the didactical model of the 'Quiz World Cup' was efficient because the students' performance in the self-assessment exams improved with each additional interaction (opportunity) over time, even though the teacher did not influence or drive the three distance learning phases. Important design decisions on such online learning activities seem to be the difficulty level of the content and the examination mode. A good setup for quizzes is the approach to use the average score of all attempts, because this can motivate students to prepare more carefully to avoid low scores at the beginning. Finally, the inclusion of a competitive element in online learning phases is a clear enabler for increasing the number of interactions with the online learning material (cf. [3]) as well as the quality of an approximated learning curve, which is also shown by comparing these findings to another case study presented in the following section.

V. THE EFFECTS OF FEEDBACK ON SELF-ASSESSMENT PERFORMANCE

In addition to evaluating the 'Quiz World Cup', we revisited a study which was conducted at the Vienna University of Economics and Business in the summer term 2012 [15]. The goal of this study was to examine the effect of different types of feedback, namely comprehensive textual feedback versus simple true/false feedback in online self-assessments. In contrast to the 'Quiz World Cup', the study was not conducted within a course setting, but followed an experimental setup. A quiz for self-assessment in the domain of Business Law was developed; it consisted of 18 multiple-choice questions and addressed 6 specific knowledge areas with each being assessed by three questions, whereby a specific subset of questions is related to one of the areas (e.g., the questions 1, 7 and 13 to area 1, the questions 2, 8, 14 to area 2).

The self-assessments were conducted in the traditional format of paper-and-pencil. After the self-assessments, two free-text questions were given to each student in order to evaluate if the learning content was understood. Moreover, the study ended with a questionnaire to gather the information about the participants. Table IV summarizes the characteristics and statistics about the study. In total, 28 students participated in the experiment. The students were randomly split into two groups (Group 1 and Group 2) with each group being treated with a specific intervention. While Group 1 received comprehensive textual feedback, Group 2 got true/false feedback after submitting an answer to a question.

Similar to the analysis done in the last section, we plotted the error rates of the 18 questions (see Figure 5 and Figure 6) and estimated the parameters for the learning curves fitting these plots by using the R framework and the *power.law.fit* function. Table V includes the power law slope, the y-axis intercept and the goodness of fit for these two approximations.

Comparing Figures 5 to Figure 6 shows that both feedback groups (comprehensive versus true/false feedback) had decreasing error rates within the total attempts. Both curves are characterized by a high error rate at the beginning (*B1=89%; B2=94%*) and a low power law slope ($\alpha1=\alpha2=0.3003$). The first aspect indicates that the students of both groups had low background knowledge in this domain or were not used to this kind of self-assessment (multiple-choice questions on paper). The initial error rate of Group 2 was slightly higher than the one of Group 1. This might be due to the fact that more students (n=6 vs. n=4) lacked domain-specific background knowledge in Group 2 than in Group 1. The misfit, then, was a consequence of the random assignment of the students to the two groups.

TABLE IV.
CHARACTERISTICS AND STATISTICS OF THE EXPERIMENT IN THE DOMAIN OF BUSINESS LAW

Characteristics of experiment	**Group 1**	**Group 2**
Type of feedback	Comprehensive feedback	True/false feedback
Statistics of experiment	**Group 1**	**Group 2**
No. students	14	14
No. students without background knowledge	4 (28.57%)	6 (42.86%)
Avg. length of examination session	28.5 minutes	16 minutes
No. students answered the two open questions properly	7 (50%)	5 (35.71%)

TABLE V.
TABLE V. APPROXIMATION OF THE LEARNING CURVES FOR DIFFERENT KINDS OF FEEDBACK IN AN ONLINE SELF-ASSESSMENT

Parameter	**Phase 1**	**Phase 2**	**Phase 3**
Power law slope (α)	0.4621	0.4469	1.3120
Error rate at x=1 (B)	0.28	0.34	0.09
Goodness of fit (R^2)	0.5794	0.6070	0.7816

The low slope can be explained by the low progress of learning. In contrast to the case study on the 'Quiz World Cup', the self-assessment for this study was not held as part of the course at the university. Thus, efforts to succeed might have been lower but the achievements of the students were still notable (i.e. error rates below 0.1 at the end of the self-assessment). The goodness of fit (R^2) is higher in the group with comprehensive feedback compared to the one with true/false feedback ($R1^2=0.2811$ vs. $R2^2=0.2085$). This indicates that the didactical model using comprehensive feedback slightly outweighed the one with true/false feedback. This conclusion is supported by the results of the two open questions presented in Table IV, posed to the students at the end of each self-assessment. While 50% of students of the comprehensive feedback group were able to respond correctly to the open questions, only 35% of students of the true/false group were able to address these questions properly.

However, the goodness of fit (R^2) is very low for both curves. Moreover, both curves include various strong relapses in the error rates. This indicates flaws in the didactical model. One can conclude that neither the true/false nor the comprehensive feedback sufficiently supported the learner in improving his or her domain knowledge. Besides the drawbacks in the didactical model, we also identified flaws in the experimental setup. According to Martin et al [2], a learning curve focuses on the error rate for practicing one specific item. For every six questions one of six specific knowledge areas, which are designated with A to F, is addressed. With three questions per knowledge area, the total number of questions is 18. Thus, the student could practice knowledge area A in questions #1, #7 and #13, knowledge area B in questions #2, #8 and #14, and so on and so forth. In fact, when splitting up the plot of Figure 5 or Figure 6 into six plots with three questions each for the related knowledge areas, this leads to curves with a higher goodness of fit (shown in [15]).

In the plots of Figures 5 and 6, one can say that the curves show tendencies towards a power law distribution but that the goodness of fit is very low. Generally it can be said that a flaw in the didactical model can be identified where the discrepancy between the arithmetic series of error rates and the approximated curve starts to increase. For the first scenario (Figure 5) this would be the case starting with question 5 while the second learning curve (Figure 6) would point teachers to the questions 10 and above.

Summing up, this second study shows an approach of using learning curves to evaluate different forms of feedback in online self-assessment. The results indicate that there were only small differences between giving feedback in the form of comprehensive text or true/false. For instance, the curve in Figure 6 fluctuates more intensely. Having a look at the other characteristics of the two feedback types, Table IV shows that self-assessments with comprehensive feedback require students to spend more efforts and time on learning and that they could comprehend the subject matter better and were thus able to answer open questions more easily. Consequently, we tend to conclude that the didactical model including comprehensive feedback is better than the one including only true/false feedback, albeit both didactical models are moderate concerning their effectiveness.

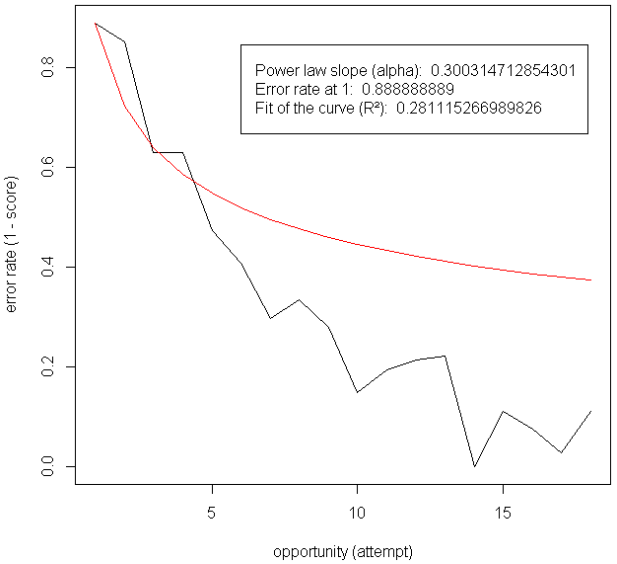

Figure 5. Learning curve and power law approximation for a multiple-choice based online quiz which gives comprehensive feedback to learners

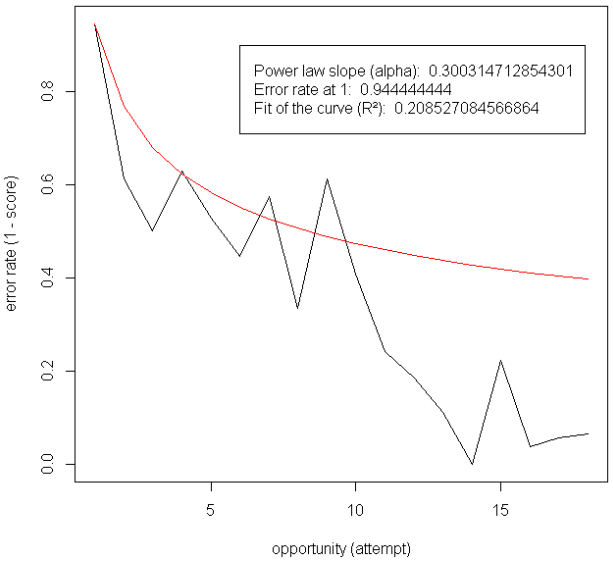

Figure 6. Learning curve and power law approximation for a multiple-choice based online quiz which gives true/false feedback to learners

Lessons learned from the experimental setup include that learning materials – e.g. multiple-choice questions for self-assessment tests – need to be carefully chosen according to the knowledge area they cover and that they should be ordered following a sound didactical model so that learning is efficient.

VI. APPLICATION AREAS AND RESTRICTIONS

The application of learning curves for evaluating didactical models is considered to be useful for two important target groups.

Above and beyond, the measurement of error rates of students should support teachers and providers of learning content to improve their online courses. Similarly to the field of Learning Analytics [16] we see a potential for a new stream in technology-enhanced learning (TEL), which we call 'Didactic Analytics' and which aims at exploiting learner-produced data to analyze and improve

didactical models. In such a discipline, learning curves are one possible instrument next to the already existing functionality of learning software and platforms (e.g. reporting tools and visualizations for students' engagement in courses) and next to novel methods which are based on indicators from Web Analytics, charts and diagrams for analyzing dynamics and distributions in online course, network analysis, web usage mining, and the forth (cf. [17]).

As shown with the two case studies in this paper, a learning curve plot can indicate if the didactical model of an instruction is valid and efficient, how difficult the content is and if there are flaws in the didactical design. In case of such flaws, the plots can also point to them and support a teacher in improving the quality of the online course.

In addition to validating the effectiveness of (distance) learning methods and their underlying didactical models, learning curves can be also used for typical application areas from Learning Analytics [16]. Amongst others, such plots might be also utilized to predict student characteristics (like task performance or affective states) or to give visual feedback to learners, e.g. by telling how he or she is performing in comparison to other learners or by indicating that the upcoming instruction will be more difficult than the ones before.

Next to these interesting application areas we also have to outline problematic aspects and restrictions of our approach. One issue deals with data gathering in terms of relying on an adequate number of attempts and students. In the 'Quiz World Cup' study we have shown that learning curves does not make sense if there are too few attempts per self-assessment test. In the third distance learning phase of this study only three attempts were allowed, thus the learning curve plot was rather trivial and consisted of an arithmetic series with three measurements. However, it has to be noted that the error rate tended to fall according to a power law distribution even with a good quality (cf. goodness of fit). Regarding this problem we are aware of the fact that many online instructions are not designed in the way that learners are motivated to repeat self-assessment tests very often. Most course entities rather restrict the number of attempts in order to force students to sufficiently prepare themselves for a concluding quiz.

Next to the number of attempts, we have also experienced that there must be an adequate number of students per attempt to avoid volatility of the learning curve. In the 'Quiz World Cup' study the first two distance learning phases are characterized through more than 40 attempts. Yet, only very few students conducted the self-assessment tests that often. That is why a weak performance e.g. in the 40th attempt can lead to a biased learning curve and wrong feedback for users (i.e. teachers and learners). Here, we have to rely on the experience of the didactical expert or consider realizing an additional check on the arithmetic series of measurement points or the approximation algorithm.

VII. CONCLUSIONS AND OUTLOOK

In concluding this paper, it can be stated that the concept of learning curves which has originally been introduced for evaluating adaptive educational systems can be applied to analyze and improve didactical models provided that they include elements for assessing any kind of error rate in learning (e.g. performance of learners). The 'Quiz World Cup' was considered to be an efficient distance learning method as it motivated learners to improve the defined objectives of a course in a self-directed and competitive way. Due to fostering multiple interactions of students with the learning platform, it also led to valuable data which can be used, for instance, for analyzing the didactical model through learning curves. Learning through feedback-based self-assessments seems to be less efficient, whereby the learning curves exhibited that the didactical model of the second study (Section V) might not be fully valid.

Of course, the data-sets used in this paper are not large enough for generalizing our findings to all didactical models of distance learning methods. Thus, future work should address other courses – probably also from different domains – as well as other definitions of the error rates within learning processes. It is also necessary to validate the existence of the 'Power Law of Practice' by analyzing and interpreting further data-sets resulting from online learning experiments.

Furthermore, another challenge is to check whether the 'Power Law of Practice' is applicable to account for the change of affective responses over time, assuming that errors normally lead to negative emotions such as frustration and confusion. Experiential qualities, however, are known to be ephemeral; it will be intriguing to observe how they evolve with practice. Independently from the learner characteristics to measure (e.g., task performance, affective or motivational states), we consider learning curves as a valuable technique for further approaches in the field of Learning Analytics.

Primarily and as indicated in the two case studies, learning curve plots are a useful instrument for evaluating didactical models, thus being an interesting method for Didactic Analytics which aims at supporting teachers and providers of educational services through feedback on and predictions for (distance) learning method based on learner-generated data.

ACKNOWLEDGMENT

This research has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 231396 (ROLE project). Furthermore, we would like to acknowledge the participation of the 32 students of a lecture held at the FH Campus02 in 2006 (first case study). The second case study was conducted by Ivan Grabic, Valentin Meiss and Gernot Reschenauer as part of a seminar work.

REFERENCES

[1] J.D. Bransford, A.L. Brown, and R.R. Cocking, *How people learn: Brain, mind, experience, and school*. Washington D.C.: National Academy Press, 2000.

[2] B. Martin, A. Mitrovic, K.R. Koedinger, and S. Mathan, "Evaluating and improving adaptive educational systems with learning curves," *User Modeling and User-Adapted Interaction*, vol. 21(3), pp.249-283, 2011. http://dx.doi.org/10.1007/s11257-010-9084-2

[3] F. Mödritscher, "The Quiz World Cup: A game-based distance learning method for teaching cognitive subject matter content," *Proc. of ED-Media*, Chesapeake, VA: AACE, 2008, pp. 4305-4311.

[4] A. Iqbal, R. Oppermann, A. Patel, and Kinshuk, "A classification of evaluation methods for intelligent tutoring systems," *Proc. of the Software-Ergonomie Fachtagung*, 1999, pp. 169-181.

[5] Wikipedia. 2012. *Interpretability*. Wikimedia Foundation. Retrieved from http://en.wikipedia.org/wiki/Interpretability (2012-11-14).

[6] Oxford Dictionaries. 2012. *Definition of interpret*. Oxford University Press. Retrieved from http://oxforddictionaries.com/definition/interpret (2011-11-14).

[7] A. Newell, and P.S. Rosenbloom, "Mechanisms of skill acquisition and the law of practice," in *Cognitive Skills and Their Acquisition*, J.R. Anderson, Ed. Hillsdale, NJ: Lawrence Erlbaum Assoc, 1981, pp. 1-56.

[8] M. Hassenzahl, "The interplay of beauty, goodness, and usability in interactive products," *Human Computer Interaction*, vol. 19(4), pp. 319-349, 2004. http://dx.doi.org/10.1207/s15327051hci1904_2

[9] E.L.-C. Law, V. Roto, M. Hassenzahl, A. Vermeeren, and J. Kort, "Understanding, scoping and defining user experience: a survey approach," *Proc. of CHI*, New York: ACM, 2009, pp. 719-728.

[10] E. Karapanos, J. Zimmerman, J. Forlizzi, and J.B. Martens, "Measuring the dynamics of remembered experience over time," *Interacting with Computers*, vol. 22 (5), pp. 328–335, 2010. http://dx.doi.org/10.1016/j.intcom.2010.04.003

[11] S. Kujala, V. Roto, K. Vaananen-Vainio-Mattila, E. Karapanos, and A. Sinnela, "UX Curve: A Method for Evaluating Long-Term User Experience," *Interacting with Computers*, vol. 23(5), pp. 473-483, 2011. http://dx.doi.org/10.1016/j.intcom.2011.06.005

[12] D. Kahneman, A.B. Krueger, D.A. Schkade, N. Schwarz, and A.A. Stone, "A survey method for characterizing daily life experience. The day reconstruction method," *Science*, vol. 306, pp. 1776-1780, 2004. http://dx.doi.org/10.1126/science.1103572

[13] B. Bloom, M. Engelhart, E. Furst, W. Hill, and D. Krathwohl, *Taxonomy of education objectives the classification of educational goals: Handbook I cognitive domain*. New York: David McKay, 1956.

[14] F. Mödritscher, and A. Sindler, "Quizzes are not enough to reach high-level learning objectives!" *Proc. of ED-Media*, Chesapeake, VA: AACE, 2005, pp. 3275-3278.

[15] I. Grabic, V. Meiss, and G. Reschenauer, *Evaluation von Lernsystemen mittels Learning Curves*. Technical Report, Vienna University of Economics and Business, 2012.

[16] G. Siemens, "What are Learning Analytics?", *Elearnspace.org*, 2010, http://www.elearnspace.org/blog/2010/08/25/what-are-learninganalytics/ (2012-02-15).

[17] F. Mödritscher, and B. Steiner, "Didactic Analytics: On exploiting intelligent, learner-produced data to analyze and improve didactical models," *Proc. of the LAK2013 Conference*, Leuven, 2013. (to appear)

AUTHORS

Felix Mödritscher is with the Institute for Information Systems and New Media at the Vienna University of Economics and Business, Vienna, Austria (e-mail: felix.moedritscher@wu.ac.at).

Monika Andergassen is with the Institute for Information Systems and New Media at the Vienna University of Economics and Business, Vienna, Austria (e-mail: monika.andergassen@wu.ac.at).

Effie Lai-Chong Law is with the Department of Computer Science at the University of Leicester, Leicester, United Kingdom (email: elaw@mcs.le.ac.uk).

Victor Manuel García-Barrios is with the Institute for Geoinformation & Spatial Information Management at the Carinthia University of Applied Sciences, Villach, Austria (email: v.garcia@cuas.at).

This article is an extended and modified version of a paper presented at the International Conference on Interactive Collaborative Learning (ICL2012), held 26 - 28 September 2012, in Villach, Austria. Received 15 November 2012. Published as resubmitted by the authors 3 December 2012.