The NEON Evaluation Framework for Educational Technologies

Michael Leitner[1(✉)], Philipp Hann[2], and Michael D. Kickmeier-Rust[2,3]

[1] CREATE 21st Century, Vienna, Austria
michael.leitner@create.at
[2] University of Technology, Graz, Austria
philipp.hann@tugraz.at
[3] University of Teacher Education, St. Gallen, Switzerland
michael.kickmeier@phsg.chõ

Abstract. The evaluation of educational technology and concrete training measures is an important task to identify strengths and weaknesses, to elucidate the applicability for specific educational goals, and to make judgements about its effectiveness. This is a non-trivial task and unusually it requires lengthy inquiries with potential end users and clients. In many cases, the evaluation procedures are too much focused on usability-like criteria and superficial aspects of effectives. In this paper, we present a holistic framework based on four distinct dimensions which serves as the conceptual starting point for the set-up of evaluation activities. A special focus of the framework lies in the mutual dependence between evaluation dimensions and evaluation procedures. In order to reduce the efforts required for evaluation procedures we developed very short (10 item) instruments and compared the outcomes with those of a standard test battery of in total 231 items. The results of an exploratory study indicate that the short versions may provide sufficiently valid and reliable results which are not significantly different from the results of the long versions.

Keywords: Evaluation · eLearning · Training · NEON framework
Short scales

1 Introduction

The evaluation of educational technology is an important task to identify strengths and weaknesses, to elucidate the applicability for specific educational goals, and to make judgements about its effectiveness [5, 9]. This is a non-trivial task and usually it requires lengthy inquiries with potential end users. A comprehensive and scientifically sound evaluation is costly on the one hand, and on the other hand, not realizable for many scenarios. Thus, an approach is required that provides a short yet valid and reliable survey to evaluate an educational technology. In addition, the evaluation of technology for learning and teaching is oftentimes too focused on usability aspects, In many cases, the evaluation of educational software is reduced to 'conventional' usability and technology acceptance studies, at best it covers a superficial learning

P. Zaphiris and A. Ioannou (Eds.): LCT 2018, LNCS 10924, pp. 256–265, 2018.
https://doi.org/10.1007/978-3-319-91743-6_20

performance dimension. This is particularly true when it comes to settings of workplace learning, distance learning, or continuing qualification. Furthermore, psychological, sociological and ethical factors influence the "learner experience" too. In practice, such factors are often ignored, or evaluators are unaware of these factors' impact on the use and experience of educational software. Education as such, however, is a very complex field and a multitude of significant variables influence the quality of a product [2]. Thus, an approach is required that builds on a holistic and comprehensive view of learning scenarios.

In this paper, we introduce the NEON Evaluation Framework that has been developed in the context of an applied research project. We present the framework's dimensions, as well as a short eLearning evaluation questionnaire that has been designed and tested with a large-scale online study. We introduce a short eLearning evaluation questionnaire with 10 items. For this questionnaire, we condensed 19 evaluation tools (UX and usability questionnaires, eLearning questionnaires, etc.). We conducted a large-scale online study comparing the results of the full-scale evaluation questionnaire with a short version. The full-scale questionnaire has about 250 items covering all 19 evaluation tools. The short questionnaire has only 10 items.

In general, we experience many evaluation instruments and scales as too time consuming [7]. They are composed of too many items and questions. Due to time and budget constraints, they can hardly be applied in commercial or practical settings. Alternatively, these instruments consist of only a few superficial items, which do not reflect the characteristics of the evaluated technology.

A solid evaluation of educational technologies is a crucial, however non-trivial task. We aim to make evaluation "easier" and more "effective" without the loss of validity. Here are some of the problems we experience: Not only the quality and user-friendliness of an eLearning tool is important, also the educational effectiveness and validity must be assured. Educational effectiveness is often miss-evaluated, meaning that the effects of a single tool are not seen in a holistic educational context and thus the eLearning software's effects are either over or underestimated.

2 NEON Evaluation Framework for eLearning Technologies

As a result of our research, we introduce the NEON evaluation framework which summarizes four major dimensions for the evaluation process: the *medium of an educational goal* (this refers to the software and the hardware), the *quality of the educational contents* in itself, the *quality of the pedagogical approach*, and the aspects of the *context* within which the educational goal is to be reached. These dimensions are the result of research into existing evaluation tools and frameworks.

With these dimensions we aim to cover all relevant aspects and dimensions of educational technologies. The major dimensions are broken down into detailed sub-aspects for which we provide a catalogue of instruments, methods, scales, and items. This supports evaluators to assemble the right amount of items to keep the evaluation process short enough for real world settings. However, the selection of items aims to produce evaluation results valid enough to gain the necessary insights into the strong and weak spots of an eLearning software. The following figure gives an overview of the framework (Fig. 1).

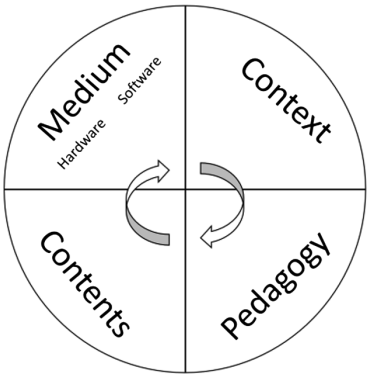

Fig. 1. Sketch of the NEON framework

We argue that a holistic approach to understanding and assessing educational measures requires building upon aspects far beyond characteristics such as usability or effectiveness. This is specifically true when evaluating with a formative approach to evaluation and technology improvement – in contrast to summative measurements.

The **first dimension** is *the medium* with which a training is presented; this refers to the hardware as such (e.g., tablets, smartphones, laptops or large screens) as well as the software (e.g., a learning management system such as Moodle, an app, or a software like *cBook* – a tool to present multimedia eLearning programs). Both hard and software must suffice the quality criteria; more importantly, these components are in a mutual dependence with the educational goal and the context conditions. Thus, statements about the quality or adequacy of the medium can only be made in the light of the other dimensions of the framework.

The **second dimension** is the *context* within which a training occurs. To develop the right educational approach, it makes a clear difference whether the learning occurs at the workplace, via mobile apps, or in form of a face-to-face workshop. More importantly, context includes the particular characteristics of the target audience, the previous knowledge, the competencies and backgrounds, as well as motivational aspects (e.g., highly motived learners do require different approaches than an audience that is not particularly interested in a training). Finally, context also includes certain constraints and limitations, e.g. in terms of course or learning time.

The **third dimension** refers to the *learning contents* as such. We argue that the contents must be seen independent from the concrete manifestation in form of a medium. The dimension refers to the correctness of the contents, the adequacy for specific learners, the syllabus (curriculum) along which it is composed, and the alignment with the educational goals. Although this seems trivial, in many trainings we could identify a gap between the actual training intentions of a customer und the concrete contents in a training. Once the right content is identified, suitable to reach the defined educational goals, these contents can be translated into the right media and the right sequence of media.

Finally, the **fourth dimension** refers to the *pedagogical (or didactic) approach*. In dependence of goals, characteristics of the target audience, certain context conditions

and constraints, various pedagogical approaches may vary significantly in their effectiveness. In certain situations exploratory approaches might be useful, in others game-based/gamified approaches, social approaches, or even "talk and chalk" teaching might be the approach of choice.

3 An Exploratory Study

3.1 Aim

When it comes to evaluation of a training or an educational measure, it is evident that a complex approach - like the described framework - results in a massive battery of evaluation items. Looking at research literature, we find complete and well-elaborated, partially standardized, test instruments for all framework dimensions. To cover all NEON framework dimensions using standardized and existing tools, test subject would be presented with 231 test items. However, it is unrealistic to ask subjects to fill in hundreds of questionnaire items. This is specifically true when focusing on real-world situations, where customers wants to deploy certain educational measures and evaluate their quality and effectiveness.

The aim of this exploratory study was to investigate whether a very short questionnaire with only 15 well-chosen items could deliver valid results – comparing it with the full and extended questionnaire comprising 231 test items.

The full scales have been compiled from the following dimensions. eLearning readiness [1], user experience [8], application related aspects [10], usability [3, 6, 13, 14], esthetics [11], acceptance [12], and educational quality [4].

3.2 Study Setup

Participants. A total of 54 subjects participated in the online study. Participants were between 19 und 55 years old (M = 30.07, SD = 9.4). 38 participants were female, 16 participants were male. Regarding their use of computers 19 participants stated that they use the computer very often, 14 use it often and 21 use it on average (M = 2.04, SD = 0.87). Regarding their use of social media 17 participants stated that they use social media very often, 15 use it often, 12 use it on average and 10 use it sometimes (M = 2.28, SD = 1.11). Participants were recruited at the University of Graz and a college of education.

The Tested Self-learning Program. For the study we designed and produced an interactive 10 min self-learning program, using a technology called the "cBook". The program introduced the concept of "gamification" to participants, designed as sequential charts. Participants could browse through the charts themselves. The program included text charts, an audio speaker guiding through parts of the self-learning program, a video, a voting chart as well as an interactive video in which participants interact with a fictional character. Participants could browse through the charts (Figs. 2 and 3).

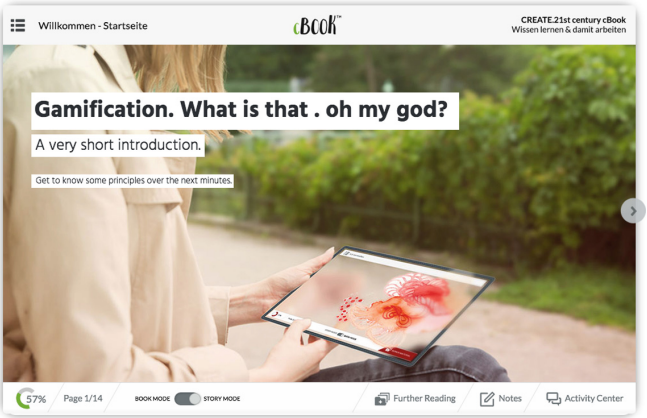

Fig. 2. Start page of the self-learning program. The program was designed on a chart-based layout. Participants could browse through the charts themselves (see arrow on the right side of the chart).

Fig. 3. As part of the self-learning program participants were presented an interactive video, in which they interacted with a fictional character called Mr. Fish (in German language).

Procedure. The first step for participants was to state their subject code, their age, their gender and their usage of computer and their usage of social media. The usage of computer and the usage of social media was measured with a 6-point Likert scale from "very often" to "never". Next the participants had to work through a short online learning course, which was created using the software "cBook" (see above). It took participants about 10 to 15 min to browse through the learning course. After finishing the online course participants were instructed to complete an evaluation questionnaire which consisted of 231 questions. It took participants about 45 min to an hour to complete the questionnaire, which contained 45 questions about the setting of the eLearning Software, 130 questions about the supporting medium, 41 questions about

the contents of the eLearning Software and 15 questions about the pedagogical concept behind the eLearning Software. The answers were given with a 7-point Likert scale from "exactly" to "strongly disagree".

15 questions out of the 231 questions were part of the long questionnaire (Table 1), but also part of a shorter questionnaire, which contained seven questions about the *medium,* three questions about the *context*, two question each about the *contents* and the *pedagogical approach*. The remaining three questions were selected out of a pool of questions that covered more than one dimension. Two questions were selected to cover the dimensions *medium and contents* and one questions was selected to cover the dimensions *medium and pedagogical concept.*

Table 1. 15 questions selected out of 231. This short questionnaire was tested against the long questionnaire version (Translated from German).

Nr	Question
1	I know why I am part of this training course
2	I have the feeling that this eLearning software is following an overall concept
3	In general, I understand what eLearning is
4	During the eLearning course I sometimes did not know what to do next
5	It was a pleasure working with the eLearning software
6	The software's structure and its elements is well thought through
7	The software is visually attractive
8	The software seems to be overloaded
9	The software includes new and innovative elements
10	Working with this software is easy, considering the know-how and the resources that are required to use it
11	I have difficulties explaining why this software is useful/useless
12	The presented text content was short and precise
13	The content was presented without any errors (grammar, spelling, etc.)
14	Most of what I have learned did not seem to be related to each other
15	The course allows me to obtain new skills in realistic situations

The aim of the study was to compare the results of the longer questionnaire with the results of the shorter questionnaire in hopes that both questionnaires would evaluate a similar total score and similar scores across the four previously mentioned dimensions: medium, context, contents and pedagogical concept.

3.3 Study Results

In a first exploratory study, we compared the 4 major dimensions in long and short versions, in detail we distinguish between educational setting, pedagogical concept, contents, and the medium (technology). In addition, we separately analysed the complete short and long instruments. Table 1 shows the descriptive results for the compared tests, that is, long and short versions for all dimensions.

A t-test was used to determine whether the total score and score of each of the four dimensions of the long questionnaire matches the total score and score of each of the four dimensions of the new short questionnaire. The total score did not differ significantly between the long (M = 2.715, SD = .568) and the short questionnaire (M = 2.729, SD = .901; t(53) = −.235, p = .815). See Table 2.

Table 2. Descriptive Statistics

	N	Min	Max	Mean	SE	SD	Var
Total long	54	1,39	4,13	2,7146	,07732	,56817	,323
Context long	54	1,64	4,20	2,6522	,07299	,53636	,288
Medium long	54	1,19	4,00	2,7604	,08354	,61392	,377
Contents long	54	1,71	4,49	2,6143	,08415	,61835	,382
Pedagogy long	54	1,33	5,13	2,5506	,09599	,70535	,498
Total short	54	1,00	4,80	2,7293	,12261	,90098	,812
Context short	53	1,00	5,00	2,0472	,12678	,92298	,852
Medium short	54	1,00	5,50	3,0101	,14514	1,06657	1,138
Contents short	54	1,00	7,00	3,0710	,21270	1,56300	2,443
Pedagogy short	47	1,00	6,00	2,2979	,19455	1,33376	1,779

The same was found with the dimension pedagogical concept (Mlong = 2.521, SDlong = .704, Mshort = 2.298, SDshort = 1.334; t(46) = 1.673, p = .101), while the other three dimensions again differed significantly between the long and the short questionnaire as followed: setting (Mlong = 2.640, SDlong = .534, Mshort = 2.047, SDshort = .923; t(52) = 6.838, p = .000), medium (Mlong = 2.760, SDlong = .614, Mshort = 3.010, SDshort = 1.067; t(53) = −2.945, p = .005) and contents (Mlong = 2.614, SDlong = .618, Mshort = 3.071, SDshort = 1.563; t(53) = −2.317, p = .024).

The t-test clearly revealed that in general the mean results in short and long versions did not differ. In order to investigate the prediction quality of short version on the item basis instead of the means, we applied a linear regression model. Table 3 lists the results of the regression analysis.

The total score of the new short questionnaire significantly predicted the total score of the long questionnaire (β = .902, t = 15.063, p = .000). The regression model explained 81% of the variance of the criteria total score of the long questionnaire (R^2 = .814). Figure 4 illustrates the regression model.

In addition, we investigated the individual correlation of scales (for layout reasons we do not present the full correlation matrix). Except for the inter-correlations between the total scores and the scores of the four dimensions from both the long and the short questionnaire, there were three notable significant correlations. First, there was a *positive correlation* between the age of the subjects and the score of the dimension *contents* of the short questionnaire, which means the older the participants were the better they rated the contents of the eLearning software, but only if you look at the short questionnaire. (r(52) = .519, p = .000). Second there was a negative correlation between the age of the subjects and the score of the dimension *pedagogical approach* of the short questionnaire, which means the older the participants were the poorer they

Table 3. Regression analysis for long and short versions.

Mod. Sum.	R	R^2	Adjusted R^2	SE	Durbin-Watson
	.902	.814	.810	.24768	1.917
ANOVA	SS	df	Mean Square	F	Sig.
Regression	13.920	1	13.920	226.899	.000
Residual	3.190	52	.061		
Total	17.110	53			
Residual statistics	Min	Max	Mean	SD	N
Prediction	1,7310	3,8924	2,7146	,51248	54
Residual	−,5384	,60261	,00000	,24533	54
SPV	−1,919	2,298	,000	1,000	54
Std. Res.	−2,174	2,433	,000	,991	54

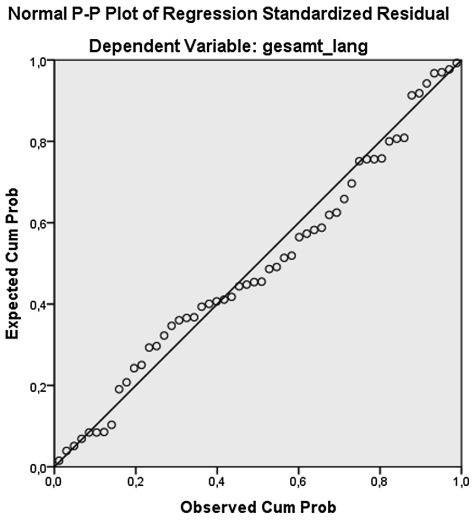

Fig. 4. Scatterplot of the linear regression between total score of the long questionnaire and total score of the new short questionnaire.

rated the pedagogical approach of the eLearning software, but also only if you look at the short questionnaire ($r(45) = -.310$, $p = .034$). Third there was a positive correlation between the usage of computers and the usage of social media, which means the more often participants use their computer the more often they use social media ($r(52) = .716$, $p = .000$).

There were no other significant correlations between the total scores and the scores of the four dimensions from both the long and the short questionnaire, the age of the participants, their gender and their computer and social media usage. Therefore, it can be assumed that both the long questionnaire and the short questionnaire can be properly used with people regardless of their age, their gender or their computer and social media usage.

4 Discussion

The aim of the NEON framework is to provide a scientifically robust, valid, and reliable approach for an evaluation of learning software and learning media that is applicable in practice. The NEON framework emphasizes the importance of taking all facets of an digital learning medium into account. It's short-sighted to believe the aspects and characteristics of the software as a carrier medium are enough for gauging its quality. A digital learning medium occurs always within certain context conditions, e.g., where and when the learning/training sessions occur, and limitations, e.g., the technology and time available. There is also a massive interaction between the content that is to be conveyed and the technological medium that is transporting the contents. With poor contents, the medium cannot be good enough to result in good learning performance and user satisfaction. The NEON framework is built around that considerations and brings together all the well-elaborated and proven tests and instruments.

Clearly it is not enough to arbitrarily select a handful of items to compose short test questionnaires, as it is for example done with the famous SUS test. As this study yielded, the short version proposed by the NEON framework showed a considerable good result. The evidence-based revision of the short questionnaire versions resulted in concise and practical instruments that meet the criteria of the original long versions.

From the application perspective, this is an important result. On the one hand, we can provide a framework that allows planning and evaluating training measures in a holistic way, encompassing all relevant characteristics and specifically their mutual dependencies. On the other hand, we can provide a methodology and concrete scales to have a handy and usable evaluation of training measures. Based on our results we argue that the short versions provide good and robust results while requiring only a minimum of time and efforts from the participants. Specifically in business-oriented learning settings this can be a decisive advantage. Certainly there is a cost-quality trade-off - in the sense that the more one invests in the evaluation process, the more detailed and reliable the obtained results.

References

1. Aydin, C.H., Tasci, D.: Measuring readiness for E-Learning: reflections from an emerging country. Educ. Technol. Soc. **8**(4), 244–257 (2005)
2. Berger, T., Rockmann, U.: Quality of e-learning products. In: Handbook on Quality and Standardisation in E-Learning, pp. 143–155. Springer, Berlin (2006). https://doi.org/10.1007/3-540-32788-6_10
3. Brooke, J.: SUS: a quick and dirty usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.) Usability Evaluation in Industry. Taylor & Francis, London (1996)
4. Entwistle, N.: Experiences of Teaching & Learning Questionnaire. ETL Project, The School of Education, University of Edinburgh (2002). www.ed.ac.uk/etl/project.html
5. Flagg, B.: Formative Evaluation for Educational Technologies. Routledge, Oxford (1990)
6. Kirakowski, J., Corbett, M.: SUMI: the software usability measurement inventory. Br. J. Edu. Technol. **24**(3), 210–212 (1993)

7. Kühl, S.: Das Evaluations-Dilemma der Beratung: Evaluation zwischen Ansprüchen von Lernen und Legitimation (2009). http://www.uni-bielefeld.de/soz/personen/kuehl/pdf/Das-Evaluations-Dilemma4-Kap-4-02042008.pdf
8. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (ed.) USAB 2008. LNCS, vol. 5298, pp. 63–76. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89350-9_6
9. Oliver, M.: An introduction to the evaluation of learning technology. Educ. Technol. Soc. **3** (4), 20–30 (2000)
10. Oztekin, A., Kong, Z.J., Uysal, O.: UseLearn: a novel checklist and usability evaluation method for eLearning systems by criticality metric analysis. Int. J. Ind. Ergon. **40**(4), 455–469 (2010)
11. Thielsch, M.: Ästhetik von Websites. Verlagshaus Monsenstein & Vannerdat, Münster (2008)
12. Venkatesh, V., Bala, H.: Technology acceptance model 3 and a research agenda on interventions. Decis. Sci. **39**(2), 273–315 (2008)
13. Willumeit, H., Gediga, G., Hamborg, K.-C.: IsoMetricsL: Ein Verfahren zur formativen Evaluation von Software nach ISO 9241/10. Ergonomie Informatik **27**, 5–12 (1996)
14. Zaharias, P.: A usability evaluation method for e-learning courses. Unpublished PhD Dissertation, Department of Management Science and Technology, Athens University of Economics and Business (2004)