Developing a Measurement Scale for the Evaluation of AR-Based Educational Systems

A. Balog, C. Pribeanu

National Institute for Research and Development in Informatics – ICI Bucharest,

Bd. Mareşal Averescu Nr. 8-10, 011455 Bucureşti, Romania

Abstract: Educational systems based on the augmented reality (AR) technology are creating a new kind of user learning experience by bringing real life objects into a computer environment. The mix of the real and virtual world requires the design of new interaction techniques which have to be tested with users early in the development process. For these new e-learning systems to be effective traditional usability evaluation is not enough. The adoption of AR-based e-learning systems in schools also requires an investigation into the perceived usefulness and increase in students' motivation. This paper presents a measurement model for the usability evaluation of AR-based e-learning systems that is targeting the educational and motivational values. The model was developed during a European research project and is inspired from the technology acceptance theories. The scale development was carried on in a methodological approach starting with the definition of a conceptual model from which an initial scale of 28 items was generated. The evaluation of the measurement model which is based on a confirmatory factor analysis resulted in a reliable scale with 19 items organized into five constructs.

Keywords: technology acceptance models, AR, e-learning, usability evaluation, user experience

Alexandru Balog received his PhD degree in Economic Informatics from the Academy of Economic Studies in Bucharest, Romania, in 1994. Since 1990 he has been a Senior Researcher in ICI Bucharest. His research interests include quality evaluation, software quality, e-services quality, quantitative methods, and information systems.

Costin Pribeanu received his PhD degree in Economic Informatics from the Academy of Economic Studies in Bucharest, Romania, in 1997. Since 1990 he has been a Senior Researcher in ICI Bucharest. Costin Pribeanu is also the Chair of the Romanian CHI group (RoCHI – SIGCHI Romania) since 2001. His research interests include usability evaluation, task analysis, and user interface design.

1. Introduction

The proliferation of AR (Augmented Reality) technologies is creating a strong opportunity for teachers to apply new teaching methods in schools. Desktop AR configurations are bringing real life objects into a computing environment thus making the e-learning process more natural and enjoyable for young students. Instead of interacting with representations of real objects and processes displayed onto a computer monitor, the user is manipulating a real object (e.g. a torso of the human body used in Biology lessons) which is observable on a see-through screen where computer generated images are superimposed. From a pedagogical point of view, AR systems have a great potential to support a learning-by-doing approach to education.

AR systems are expensive since a lot of research and design effort is needed to develop visualization and rendering software. On another hand, the mix of the real and the virtual requires appropriate interaction techniques, which have to be tested with users early in the development process in order to avoid usability problems. While the usability of interaction techniques is critical for a good user experience the successful adoption of AR technologies in school require to investigate also the educational and motivational values of a given application.

The main objective of the ARiSE (Augmented Reality for School Environments) project is to test the pedagogical effectiveness of introducing AR (Augmented Reality) in schools and creating remote collaboration between classes around AR display systems. The project has developed a new technology, the Augmented Reality Teaching Platform (ARTP), by adapting an existing augmented reality system for museums to the needs of students in primary and secondary classes.

The specific objectives of the project are: (1) To adapt the AR technology for the specific needs of schools; (2) To develop interaction scenarios to promote collaborative work between students; (3) To develop tools for easy use of the ARTP by teachers; (4) To demonstrate the pedagogical effectiveness of activities using the AR platform; (5) To build on a design framework able to support the usability of interactive systems.

To address the project's objectives, we developed a usability questionnaire that goes beyond the traditional usability evaluation approaches, by targeting the educational and motivational value. The measurement instrument (scale) is based on a conceptual model inspired by the technology acceptance theories and consists of five constructs: ergonomics of the platform, perceived ease of using the application, perceived usefulness, perceived enjoyment and intention to use. By addressing issues like perceived enjoyment and perceived usefulness, the usability evaluation results could be better integrated with the pedagogical evaluation results.

The questionnaire was firstly administered during and after a summer school in 2007, with the aim of improving the usability of the implemented learning scenarios. This way, it served as an instrument for the user-centered formative evaluation of the ARTP. After the installation of the improved version of the software, it was administered again to 278 students. The results were used to evaluate the measurement model.

The methodological approach to the development and validation of the measurement scale consists in five main steps: (1) conceptualization of constructs and item generation, (2) pilot test and preliminary item analysis, (3) data analysis and processing, (4) preliminary scale evaluation, and (5) model testing and validation. In this paper we will present the first four steps undertaken to define a reliable and sensitive scale as a pre-requisite for validating the measurement model (cf. Gediga et al., 1999). The preliminary evaluation of the measurement scale was performed by carrying on a exploratory factor analysis. The final result is a scale with 19 items organized into 5 constructs.

2. Related Work

2.1 Usability of software systems

In an early version of the ISO 9126:1991 standard, usability was defined as a software quality attribute that bears on the capability of being easy to understand, learn and operate with. Later on, the ISO standard 9241-11:1994 took a broader perspective on usability as the extent to which a product can be used by specified users to achieve specified goals effectively, efficiently and with satisfaction in a specified context of use. In this standard, the context of use has four main components: user, tasks, platform and environment. These definitions were revised and integrated in the new version of ISO 9126:2001 standard on quality of software product as follows:

- Usability is the capability of a software product to be understood, learned, used and attractive to the user, when used under specified conditions.
- Quality in use is the capability of software product to enable specified users to achieve specified goals with effectiveness, productivity, safety and satisfaction in specified contexts of use.

How to measure and improve the usability of interactive systems is a key research concern in HCI and has led to guidelines for improving usability as well as methods and techniques to test the extent to which it has been achieved. There are several approaches to usability evaluation and, consequently many usability evaluation methods (Hornbaek, 2006). In the last decade, many usability studies compared the effectiveness of various usability evaluation methods. As Law and Hvannberg (2002) pointed out, the trend is to delineate the trade-offs and to find ways to take advantage of the complementarities between different methods.

Formative usability testing is performed in an iterative development cycle and aims at finding and fixing usability problems as early as possible (Teofanos and Quesenbery, 2005). This kind of usability evaluation is called "formative" in order to distinguish it from "summative" evaluation which is usually performed after a system or some component has been developed (Scriven, 1991). Formative usability evaluation can be carried on by conducting an expert-based usability evaluation (sometimes termed as heuristic evaluation) and / or by conducting user testing with a small number of users. In this last case, the evaluation is said to be user-centered, as opposite to expert-based formative evaluation.

2.2 Technology acceptance

Significant progress has been made in explaining and predicting user acceptance of

information technology. There have been several theoretical models employed to study user acceptance and usage behavior of emerging information technologies. The Technology Acceptance Model (TAM) (Davis 1989, Davis et al., 1989) is the most widely applied model of user acceptance and usage.

TAM suggests that use is influenced by the user's attitude towards the technology, which in turn is influenced by two specific beliefs: perceived ease of use and perceived usefulness. As shown in Figure 1, TAM proposes six constructs (Davis, 1989): actual behavior (actual system use), behavioral intention to use, attitude toward behavior, perceived usefulness, perceived ease of use and external variables.

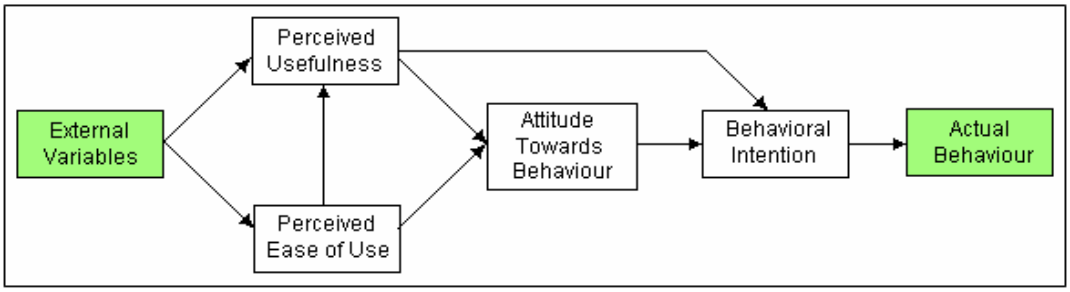

Figure 1. Technology Acceptance Model (Davis, 1989)

Two constructs, namely external variables and actual behavior, were introduced to encapsulate observable components of technology adoption. External variables refer to all the external characteristics of a system ranging from menus, icons to output produced by the system (Davis, 1989). Actual system use refers to the potential adopter's system usage behavior. TAM explains how the external characteristics of the system affect the potential adopter's attitudes and perceptions leading to actual use of the system based on the theory of reasoned action.

Perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance" (Davis, 1989). Perceived usefulness explains the user's perception of the extent to which the technology will improve the user's workplace performance, such as decreasing the time for doing the job, more efficiency and accuracy. A potential adopter's perception of usefulness is directly affected by the degree to which they perceive that the external characteristics of a system will aid them in performing a task or a set of tasks.

Perceived usefulness is also considered to have a positive direct effect on behavioral intention and on attitude towards using a system. When potential adopters observe that the system delivers positive outcomes this will positively increase their affect with regards to using the system and their intention to use it.

Perceived ease of use is defined as "the degree to which a person believes that using a particular system would be free of effort" (Davis, 1989). The complexity of the external characteristics of the system has a direct effect on perceived ease of use.

Perceived ease of use is considered to have a positive direct effect on attitude; for example, if an individual considers that using a system will be fairly free of effort, their affect with regards to using the system will increase positively.

Attitude towards using a technology was omitted in final model (Davis et al., 1989)

because of partial mediation of the impact of beliefs on intention by attitude, a weak direct link between perceived usefulness and attitude, and a strong direct link between perceived usefulness and intention.

In another acceptance model based on TAM, Venkatesh (2000) has introduced the perceived enjoyment, a conceptualization of intrinsic motivation that is system-specific. It is defined as "the extent to which the activity

of using a specific system is perceived to be enjoyable in its own rights, aside from any performance consequences resulting from system use"(Venkatesh, 2000). This factor is actually relating to the intrinsic motivation which is an important feature of modern educational systems, especially when these are devised for young learners.

TAM has been tested to explain or predict behavioral intention on a variety of information technologies and systems, such as: word processors, email, voicemail, graphics software, net conferencing software, Internet, online shopping, online learning, Internet banking and so on. Thus the TAM has been shown to be valid over a variety of commercially available technologies that are primarily used in an office or educational environment (Legris et al., 2003; Silva, 2007, Venkatesh et al, 2003).

The quality in use is closely related with four characteristics of the software quality: functionality, usability, reliability and efficiency. As such, it is more comprehensive and requires a broader view on the design and evaluation of interactive systems, by targeting usefulness and user attitudes towards the system.

Subjective user perceptions of an interface can directly mediate perceptions of the system usability. Research has shown that user perceptions of a system's user interface are strongly related to perceived usability and may significantly affect overall system acceptability. Measuring perceived usability can be accomplished through the observation of actual users interacting with the system and collecting objective and subjective data measuring the users` satisfaction with the system.

3. Research Model and Hypotheses

Our research model is presented in Figure 2. We hypothesized that the usability of the ARTP defined by two factors (ergonomics and perceived ease of use) is influencing two factors of motivation to adopt a technology (perceived enjoyment and perceived usefulness), which in turn are influencing the intention to use AR in schools. The relationships between factors are labeled with the number of the corresponding hypothesis.

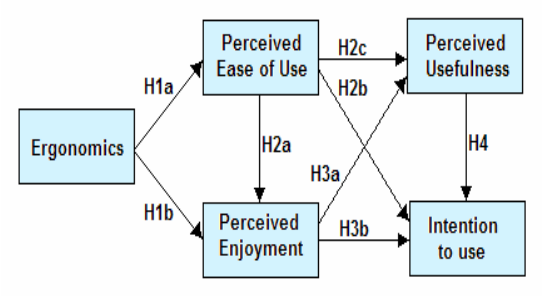

Figure 2. The research model

The hypotheses in this research model are summarized bellow:

H1a: The ergonomics of the platform has a positive effect on the perceived ease of use

H1b: The ergonomics of the platform has a positive effect on the perceived enjoyment

H2a: The perceived ease of use has a positive effect on the perceived enjoyment

H2b: The perceived ease of use has a positive effect on the intention to use

H2c: The perceived ease of use has a positive effect on the perceived usefulness

H3a: The perceived enjoyment has a positive effect on the perceived usefulness

H3b: The perceived enjoyment has a positive effect on the intention to use

H4: The perceived usefulness has a positive effect on the intention to use.

The ergonomics of the platform refers to the ease of use of the specific ARTP hardware and accessories: see-through screen, stereo glasses and headphones. This construct does not depend on a specific learning scenario.

The perceived ease of use refers to the ease of use of a particular learning application implemented onto the ARTP. This construct is targeting several usability aspects such as: ease to understand, ease to learn how to operate with, ease to remember how to operate with, and ease to operate.

The perceived usefulness refers to the pedagogical value of the ARTP. The construct is targeting specific pedagogical aspects, such as faster understanding, support for learning and general usefulness of ARTP for the learning process.

The perceived enjoyment refers to the motivational value (intrinsic motivation) of

the ARTP. The construct is targeting various aspects that create an enjoyable learning experience: interesting way of learning, captivating exercises, attractive technology, manipulation of real objects, enjoyable and exciting way of learning.

4. Scale Development

This work was carried on in a methodological framework for scale development and validation which is grounded on the Churchill (1979) paradigm and a set of the updates and improvements (Hair et al., 2006; Worthington & Whittaker, 2006).

4.1 Conceptualization of constructs and generation of variables

The first step in scale development effort is to decide what is being measured. More specifically, this step is a review of the extant literature to specify the domain of the construct (i.e., prepare an exact definition of the construct based on literature that delineates the boundaries of the construct domain). This is one of the most critical steps of the scale development process because the definition generated in this step will serve as the basis for constructing items early in the process and will be important in determining validity later in the process.

Three experts, whose expertise is related to the areas of interactive systems usability, software quality, and scale development, generated prospective items for the usability measurement scale. The experts initially developed 41 scale items based on the relevant literature review. Most items were adopted from the usability questionnaires (e.g., QUIS, SUMI, QSUQ, PUTQ) and prior related empirical studies (e.g., Davis et al., 1989; Venkatesh et al., 2003) and were adapted to reflect the context of AR technology.

In order to minimize bias in this research, the activities were focused on three issues when designing the questionnaire: (a) the wording of the questions (statements); (b) planning of issues of how the variables will be categorized and scaled, and (c) the general appearance of the questionnaire.

An internal pretest of the questionnaire was conducted to assess the face validity of measurement scales. Face validity can be evaluated by a panel of persons, sometimes experts, who judge whether a scale is logically appears to reflect accurately what it supposed to measure. The three experts reviewed the questionnaire draft and items were screened for ambiguity, double barreled wording, overlay, and redundancy. Also, the statements were limited to positively worded items due the potential confusion created by negatively worded items.

Based on their suggestions, eighteen scale items were reworded because they had ambiguous terms and thirteen were deleted because they were not directly related to AR systems usability. The length of the questionnaire to be administrated to students of 13-16 years old was another consideration in excluding scale items.

Table 1 presents the initial list of 28 variables grouped onto the 5 constructs of the measurement model.

4.2 Initial pilot test and preliminary item analysis

The usability questionnaire was administrated to a pilot sample (N=124) in October-November 2007. A five-point Likert scale ranging from (1) “strongly disagree” to (5) “strongly agree” was used. The purpose of this first round of data collection was to examine the initial structure of the scale items and to begin purification (removing) of the items on the basis of their psychometric properties.

The procedures utilized in this study included data cleaning, data analysis, item analysis and preliminary reliability tests. Data were analyzed by means of the statistical software package SPSS 15 for Windows. The main purpose of the item analysis and reliability analysis of the data is to determine whether the data is trustworthy. Reliability refers to the instrument’s consistency and is defined as “assessment of the degree of consistency between multiple measurements of a variable” (Hair et al., 2006).

With this purpose, the questionnaire data were subjected to the analysis using the measure of internal consistency. Cronbach’s alpha for the scale was 0.940, which is high

Table 1. Constructs and variables in the measurement model

Constructs	Items	Variables
Ergonomy of the ARTP (ERG)	ERG1	Adjusting the "see-through" screen is easy
	ERG2	Adjusting the stereo glasses is easy
	ERG3	Adjusting the headphones is easy
	ERG4	The work place is comfortable
	ERG5	Observing through the screen is clear
Perceived ease of use (PEOU)	PEOU1	Understanding how to operate with ARTP is easy
	PEOU2	The superposition between projection and the real object is clear
	PEOU3	Learning to operate with ARTP is easy
	PEOU4	Remembering how to operate with ARTP is easy
	PEOU5	Understanding the vocal explanations is easy
	PEOU6	Reading the information on the screen is easy
	PEOU7	Selecting a menu item is easy
	PEOU8	Correcting the mistakes is easy
	PEOU9	Collaborating with colleagues is easy
	PEOU10	Overall, I find the system easy to use
Perceived usefulness (PU)	PU1	Using ARTP helps to understand the lesson more quickly
	PU2	After using ARTP I will get better results at tests
	PU3	After using ARTP I will know more on this topic
	PU4	Overall, I find the system useful for learning
Perceived Enjoyment (PE)	PE1	The system makes learning more interesting
	PE2	Working in group with colleagues is stimulating
	PE3	I like interacting with real objects
	PE4	Performing the exercises is captivating
	PE5	Overall, I enjoy learning with the system
	PE6	Overall, I find the system exciting
Intention to use (INT)	INT1	I would like to have this system in school
	INT2	I intend to use this system for learning
	INT3	I will recommend to other colleagues to use ARTP

and acceptable. The conclusion of this analysis was that none of the items would substantially affect reliability if they were deleted. Therefore no item was removed.

5. Data Analysis

To develop and validate the measurement scale, the current research employed methods for multivariate analysis of interdependency are used (factor analysis). Before applying these methods we analyzed the data and verified the adequacy of application. The procedures, verification criteria and critical values are based on recommendations from the literature (Hair et al., 2006; Field, 2006; Tabachnick & Fidell, 2007).

5.1 Data analysis

Data are available from a sample with 278 observations. The sample size is within the recommended values and is acceptable for multivariate analysis. The filled in questionnaires were checked for completeness and we didn't find missing values.

In order to identify univariate outliers we computed the z-scores for each variable. According to the sample size (N=278), observations with z-score over ±3,29 ($p < 0,001$, two-tailed test) are potential univariate outliers. We identified 21 distinct univariate outliers in 12 variables observed with z-scores varying from 3,34 to 4,09.

The multivariate outliers were identified by computing the Mahalanobis distance. Using the criterion α=.001 with DF=28, the critical value is χ^2=56.892 we identified 19 distinct observations which present multivariate outliers, from which 4 were also in the list of univariate outliers.

Then we analyzed two aspects of normality: the distribution shape and the sample size.

As regarding the distribution shape, we computed the skewness and kurtosis and we applied several methods for the verification of normality hypothesis: histograms, and normal probability plot, statistical tests (z statistics for skewness and kurtosis). We found a moderate negative skewness for all 28 variables, from which 19 are bellow -1. The kurtosis has positive values for 26 variables and negative values for 2. For 13 variables the kurtosis value is over +1.

As regarding the sample size, the effects of non-normality of the data could be diminished in samples of 200 observations and higher (cf. Hair et al., 2006).

5.2 Data processing

The outliers were analyzed together with the deviation from normality. We tested the possibility of improving the data quality by subsequent elimination of observations with univariate and multivariate outliers (36 observations). Results shown a slight improvement of the normality but not significant enough as reported to the number of removed observations.

In order to comply with the statistical hypotheses for multivariate analysis methods, we carried on an iterative procedure based on the recommendations of Field (2005) and Tabachnick & Fidel (2007):

- Data transformation by variable reflection and square root extraction.
- Identification and analysis of univariate outliers by computing *z-scores.*
- Successive elimination of observations with univariate outliers and reiteration of univariate outliers identification and normality tests.
- Identification and analysis of multivariate outliers by computing the Mahalanobis statistics.
- Successive elimination of observations with multivariate outliers and reiteration of multivariate outliers identification and normality tests
- Analysis of results.

Data transformation resulted in a substantial improvement of normality criteria by a successive elimination of 24 observations with univariate and multivariate outliers. The final sample (N=254) had a moderated deviation from normality and was further used to verify the supplementary conditions required to apply multivariate analysis methods.

6. Preliminary Scale Evaluation

6.1 Methods, procedures and criteria

In this phase we applied Exploratory Factor Analysis (EFA) at construct level (sub-scale) and scale level. EFA is an adequate procedure for a preliminary evaluation and refinement of measurement scales. The approach is justified for at least two reasons:

- The development of a new scale (constructs and variables) in the AR context of AR-based applications.
- The employment of some constructs from TAM (Davis et al., 1989; Venkatesh, 2000) and the adaptation of the variables semantics to the context of AR systems.

We applied EFA at each level in two steps: the analysis of unidimensionality and the evaluation of internal consistency (scale reliability). The procedures and criteria have been selected based on the recommendations from Fabriger et al. (1999), Costello & Osborne (2000) and Tabachnick & Fidel (2007). In the current research, the applications of EFA were carried out using SPSS 15 for Windows.

Unidimensionaliy is defined as the existence of one construct underlying a set of items. It is the degree to which a set of items represent one and only one underlying latent construct. The test for unidimensional scales is important before undertaking reliability tests because reliability such as Cronbach alpha does not ensure unidimensionality but instead assumes it exists (Hair et al. (2006).

The unidimensionality of scale was evaluated with the Factor Analysis procedure from SPSS. As factor extraction method we selected Principal Axis Factoring since the objective is to explore latent constructs represented in the original variables. Also, based on recommendations from Fabriger et al., (1999), we selected the oblique rotation

method PROMAX since the variables within constructs and the constructs within the model are correlated from a theoretical point of view.

The main purpose of the first step is to see whether the scale for each construct under investigation is unidimensional (i.e. first-order construct) or multidimensional (i.e. second-order construct). For a scale to be empirically unidimensional, the factor analysis must result in only one factor extracted.

eliminated. As a standard for this preliminary assessment, the scale for each construct must achieve a minimum Cronbach alpha of 0.70.

In the second step, Principal Axis Factoring was performed on all items of all constructs put together to have a preliminary asesssment of unidimensionality. Given the results of step one where each item loads highly on the factor representing its underlying construct, this factor analysis allows all items to correlate

Table 2. Unidimensionality and reliability test results

Construct / Item	Factor Loading	Communalities	Cumulative % Variance Explained	Eigenvalue	Corrected Item-Total Correlation	Cronbach's Alpha
Perceived Usefulness (PU)			64.95	2.598		.819
PU1	.690	.477			.615	
PU2	.693	.480			.618	
PU3	.694	.481			.613	
PU4	.843	.711			.722	
Intention to use (INT)			72.46	2.174		.809
INT1	.654	.428			.588	
INT2	.794	.630			.674	
INT3	.854	.730			.710	

This is necessary because all latent constructs in the theoretical model are operationalized as unidimensional constructs.

In the first step, Principal Axis Factoring was applied to each of the five constructs under investigation. The number of factors to be retained was decided based on Eigenvalue ≥1.0, scree plot test and explained variance ≥ 60%. Moreover, variables (items) with low factor loading (< 0.60) were eliminated because they do not converge properly with the latent constuct they were designed to measure. Also, variables with low communalities (< 0.40) were eliminated based on theoretical justification and values obtained for other criteria.

Then, realiability analysis (Cronbach alpha) was applied to each set of items to assess and refine the measurement items. Scale reliability evaluation has been done with the Reliability Analysis procedure from SPSS. Items having low corrected item-to-total correlation coefficients (< 0.40) were

with every factor without being constrained to correlate only with its underlying factor.

Consequenly, it allows the investigation of the general corelation pattern of the measurement items (Fabriger et al., 1999).

6.2 Results

Following the procedure and criteria presented above, the factor analysis results show that out of the total five scales, two were immediately acceptable ("Perceived Usefulness" and "Intention to Use") while three scales needes some refinements.

The scales that did not require any modification are shown in Table 2.

Using the *Eigenvalue* ≥ 1.0 criterion, the results show that only one factor was extracted for each of these scales. The variance explained by the extracted factor is 64.95% for "Perceived Usefulness" and 72.46% for "Intention to Use", while the factor loadings are all above the threshold of 0.60.

Table 3. Unidimensionality and reliability test results – refined scales

Construct/ Item	original scale				refined scale			
	Factor loading	Commu-nalities	Correc-ted Item-Total Correla-tion	Cronbach's Alpha if Item Deleted	Factor loading	Commu-nalities	Correc-ted Item-Total Correla-tion	Cronbach's Alpha if Item Deleted
Ergonomy of the ARTP (ERG)								
ERG1	,737	,543	,586	,624	,774	,599	,658	,720
ERG2	,823	,677	,682	,584	,741	,550	,638	,741
ERG3	,362	,131	,315	,728	removed			
ERG4	,327	,107	,282	,746	removed			
ERG5	,712	,507	,554	,635	,761	,580	,650	,729
% Variance Explained	49,04%				71,74%			
Cronbach's Alpha	0,720				0,803			
Perceived Ease Of Use (PEOU)								
PEOU01	,727	,528	,686	,893	,739	,546	,698	,885
PEOU02	,662	,439	,622	,897	,682	,465	,642	,891
PEOU03	,706	,499	,673	,894	,680	,462	,640	,890
PEOU04	,591	,350	,565	,901	removed			
PEOU05	,580	,337	,552	,901	removed			
PEOU06	,711	,506	,675	,893	,709	,502	,671	,888
PEOU07	,800	,640	,755	,889	,797	,635	.749	,881
PEOU08	,755	,569	,709	,891	,761	,580	,715	,884
PEOU09	,680	,463	,644	,895	,662	,439	,622	,892
PEOU10	,778	,606	,733	,891	,797	,635	,751	,881
% Variance Explained	54,26%				59,02%			
Cronbach's Alpha	0,905				0,900			
Perceived Enjoyment (PE)								
PE1	,768	,590	,675	,770	,775	,601	,686	,770
PE2	,506	,256	,465	,817	removed			
PE3	,490	,241	,451	,817	removed			
PE4	,705	,496	,626	,781	,694	,481	,619	,801
PE5	,743	,552	,655	,776	,751	,564	,666	,781
PE6	,737	,543	,645	,777	,745	,556	,655	,785
% Variance Explained	53,20%				66,21%			
Cronbach's Alpha	0,820				0,830			

After establishing that these scales are unidimensional, the reliability of scales was assessed. As shown in Table 2, the Cronbach alpha of these scales are all above the threshold of 0.70 (0.819 and 0.808, respectively). The item-to-total correlation values, which range from 0.588 to 0.722, are also all above the threshold of 0.40. All items comprising these two scales were therefore retained.

The scales that needed some refinements are shown in Table 3. All these scales yielded only one factor with variances explained ranging from 49.04% to 54.26%. The Cronbach alpha are all well above the threshold of 0.70. However, the factor loadings and item-to-total correlations values of six items in these three scales are below the acceptable thresholds (original scales in Table 3).

Items ERG3 and ERG4 have low factor loading (0.362 and 0.327, respectively) and low item-to-total correlations (0.137 and 0.107, respectively). These two items did not provide consistent results with the other three items in the scale. Thus, ERG3 and ERG4 are deleted.

Items PEOU4 and PEOU5 have factor loadings bellow the threshold of 0.60 (0.591 and 0.580, respectively) and low item-to-total correlations bellow the threshold of 0.40 (0.350 and 0.337, respectively). Thus, PEOU4 and PEOU5 are deleted.

Items PE2 and PE3 have factor loadings bellow the threshold of 0.60 (0.506 and 0.490, respectively) and low item-to-total correlations (0.256 and 0.241, respectively). Thus, PE2 and PE3 are deleted.

The modified scales, show satisfactory factor loadings (range is from 0.680 to 0.797) and explain 71.74% (ERG), 59.02% (PEOU), and 66.21% (PE) of the variance. The reliability tests show the Cronbach alpha above the threshold of 0.70 (0.802, 0.899 and, respectively, .830) and the item-to-total correlations are all above the 0.40 threshold (range is from 0.619 to 0.751).

In summary, 6 items were eliminated (ERG3, ERG4, PEOU4, PEOU5, PE2, PE3), and the remaining 22 items for the 5 scales were retained. All the five scales are now acceptable.

After the evaluation of unidimensionality and reliability for each sub-scale we re-iterated the procedures and tests for all 22 variables together. Variables PEOU3, PEOU9 and PU3 had a factor loading of 0.590, 0.525 and 0.566 (bellow the specified 0.60 cut-off value). These variables were successively eliminated and criteria evaluation was performed at each step. The remaining 19 items for the 5 scales were retained. The results of this procedure are shown in Table 4.

Table 4. Results of factor analysis for 5 scales

Items	Factor				
	1	2	3	4	5
ERG1					,661
ERG2					,699
ERG5					,777
PEOU1	,748				
PEOU2	,776				
PEOU6	,693				
PEOU7	,629				
PEOU8	,729				
PEOU10	,796				
PU1				,741	
PU2				,618	
PU4				,815	
PE1		,751			
PE4		,619			
PE5		,705			
PE6		,761			
INT1			,609		
INT2			,795		
INT3			,894		

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.
Rotation converged in 6 iterations

As shown in the Table 4, five factors were extracted which together explain 68.77% of the total variance. The factor loadings of each of the 19 items vary from 0.609 to 0.894 which are higher than the threshold of 0.60. No item load highly on more than one factor and no item load highly on a factor other than its designate factor representing its latent construct.

Regarding the issue of appropriateness, the result of the Bartlett's Test of Sphericity and KMO measure (Tabachnick & Fidel, 2007)

indicated that the degree of intercorrelations among the items was suitable for factor analysis procedure (Chi-square=2321.22, df=171 and sig.=0.000, KMO=0.897).

Reliability test resulted in a Cronbach's alpha of .910 (over the 0.70 cut-off value) and the correlations between each variable and total score are above the cut-off value of 0.40.

The correlation between extracted factors is presented in Table 5.

Table 5. Correlation between extraction factors

Factor	PEOU	PE	INT	PU	ERG
PEOU	1.000				
PE	,500	1.000			
INT	,355	,417	1.000		
PU	,533	,553	,421	1.000	
ERG	,695	,516	,386	,530	1.000

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.

7. Conclusion and Future Work

While the usability of a software system is a critical feature, usability evaluation per-se does not ensure a successful adoption of the underlying information technology. The evaluation of new systems should go beyond the traditional usability approach and investigate the user acceptance in order to understand the various factors that influence the intention to use.

In this paper we presented the development of a measurement scale that is intended to measure three core features of an AR-based teaching platform: usability, pedagogical and motivational value. Our evaluation strategy is based on the integration of formative and summative usability evaluation with the methods and procedures used in the technology acceptance theories.

The preliminary scale evaluation leaded to a measurement model of 19 items grouped into 5 constructs: ergonomics of the ARTP, perceived ease of use, perceived usefulness, perceived enjoyment and intention to use. As such, the model is able to answer the research questions of the ARiSE project and makes it easier to integrate the usability evaluation results with the pedagogical evaluation results obtained with specific qualitative methods.

Nevertheless, the measurement model is a prerequisite for the final model evaluation and validation. The next step is the evaluation of the measurement model and the validation of the structural model. This work will be carried on with a structural equations modeling approach. Then base on the causal relationships between constructs we will be able to statistically accept / reject the research hypotheses.

ACKNOWLEDGEMENT

This work was supported by the ARiSE research project funded by EU under FP6 027039.

REFERENCES

1. CHURCHILL, G.A. (1979). **A Paradigm for Developing better Measures of Marketing Constructs**. Journal of Marketing Research, Vol. XVI, February 1979, pp. 64-73.
2. COSTELLO, A.B., OSBORNE, J.W. (2005). **Best Practices in Exploratory Factor Analysis**, Practical Assessment, Research and Evaluation vol. 10, no. 7, pp. 1-9.
3. DAVIS, F.D. (1989). **Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology**. MIS Quarterly vol.13, no.3, pp. 319-339.
4. DAVIS, F.D., BAGOZZI, R.P., WARSHAW, P.R. (1989). **User Acceptance of Computer Technology: A Comparison of Two Theoretical Models**. Management Science, Vol. 35, No. 8, pp. 982-1003.
5. FABRIGER, L. R., MACCALLUM, R. C., WEGENER, D. T., & STRAHAN, E. J. (1999). **Evaluating the use of exploratory factor analysis in psychological research**. Psychological Methods, no. 4, pp. 272-299.
6. FIELD, A. (2005). **Discovering Statistics Using SPSS**. Second edition. SAGE Publications Ltd.

7. GEDIGA, G., HAMBORG, K.-C. & DUNTSCH, I. (1999) **The IsoMetrics Usability Inventory - An operationalisation of ISO 9241-10**. Behavior and Information Technology, 18, pp. 151-164.

8. HAIR, J.F., BLACK, W.C., BABIN, B.J., ANDERSON, R.E., TATHAM, R.L. (2006). **Multivariate Data Analysis**. 6th ed., Prentice Hall, 2006.

9. HORNBAEK, K. (2006) **Current practice in measuring usability: Challenges to usability studies and research**. Int. J. Human Computer Studies. 64 (2006).pp. 79-102.

10. ISO/DIS 9241-11:1994 **Information Technology – Ergonomic requirements for office work with visual display terminal (VDTs)** - Guidance on usability.

11. ISO 9126-1:2001 **Software Engineering - Software product quality.** Part 1: Quality Model.

12. LAW, E., HVANNBERG, E.T., HASSENZAHL, M. (2006). U**ser Experience – Towards a unified view.** Procedings of UX Workshop NordiCHI 2006, pp. 1-3, ACM Press, Oslo.

13. LEGRIS, P., INGHAM, J., & COLLERETTE, P. (2003). **Why people use information technology? A critical review of technology acceptance model**. Information & Management, vol. 40, pp. 191-204.

14. SCRIVEN, M.: **Evaluation thesaurus.** 4th ed. Newbury Park, CA: Sage Publications (1991).

15. SILVA, L. (2007). **Post-positivist Review of Technology Acceptance Model**. Journal of the Association for Information Systems, vol. 8, issue 4, pp. 255-266.

16. THEOFANOS, M. & QUESENBERY, W. (2005). **Towards the Design of Effective Formative Test Reports. Journal of Usability Studies**, Issue 1, Vol. 1. pp. 27-45.

17. TABACHNICK, B. G., AND FIDELL, L. S. (2007). **Using Multivariate Statistics**, 5th ed. Boston: Allyn and Bacon.

18. VENKATESH, V. (2000). **Determinants of Perceived Ease of Use:** Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. Information Systems Research, Vol. 11, No. 4, pp. 342-365.

19. VENKATESH, V., DAVIS, F.D. (2000). **A Theoretical Extension of the Technology Acceptance Model**: Four Longitudinal Field Studies. Management Science Vol. 46, No. 2, pp. 186-204.

20. VENKATESH, V., MORRIS, M., DAVIS, G., DAVIS, F. (2003). **User Acceptance of Information Technology**: Toward a Unified View. MIS Quarterly Vol. 27, No. 3, pp. 425-478.

21. WORTHINGTON, R.L., WHITTAKER, T.A (2006). **Scale Development Research. A Content Analysis and Recommendations for Best Practices.** The Counseling Psychologist vol. 34, no. 6, pp. 806-838.