A Large-scale Evaluation of a Model for the Evaluation of Games for Teaching Software Engineering

Giani Petri, Christiane Gresse von Wangenheim, Adriano Ferreti Borgatto
Graduate Program in Computer Science (PPGCC)
Department of Informatics and Statistics
Federal University of Santa Catarina (UFSC)
Florianópolis/SC, Brazil
giani.petri@posgrad.ufsc.br, c.wangenheim@ufsc.br, adriano.borgatto@ufsc.br

Abstract—In order to adopt games for Software Engineering (SE) education effectively it is essential to obtain sound evidence on their quality. A prominent evaluation model is MEEGA (Model for the Evaluation of Educational Games), which provides a systematic support to evaluate the game's quality in terms of motivation, user experience and learning. To facilitate its application, the model provides a questionnaire for collecting data on the perception of the students after they played an educational game in a case study with a one-shot post-test design. However, in order to assure a valid feedback on the game's quality an important issue is the reliability and validity of the questionnaire. In this respect, this article presents a large-scale evaluation of the MEEGA questionnaire in terms of reliability and construct validity. The analysis is based on data collected in 43 case studies, evaluating 20 different SE games, involving a population of 723 students. Our analysis indicates that the MEEGA questionnaire can be considered reliable (Cronbach's alpha α=.915). In terms of construct validity, there exists evidence of convergent validity through an acceptable degree of correlation of almost all item pairs within each dimension. Yet, we identified a need for the re-grouping of items based on the results of a factor analysis, mainly with respect to items related to motivation and user experience. These results allow SE researchers and instructors to rely on the MEEGA questionnaire in order to evaluate SE games and, thus, contribute to their improvement and to direct an effective and efficient adoption for SE education.

Large-scale evaluation; questionnaire; SE games (key words)

I. INTRODUCTION

One of the main challenges when teaching Software Engineering (SE) is to give students sufficient hands-on experience in building software [1]. In addition, software professionals are not only expected to successfully cope with technical challenges, but also to deal with non-technical issues, including management, communication and team work [2]. And, although, students are, typically, skilled in programming, they often do not have experience in applying SE practices and methods to complex problems in an engineering environment.

One of the reasons for this problem can be the way in which SE is taught. Traditional lectures are still the dominant instructional technique [3]. While they are adequate to present abstract and factual information, they are not the most suitable for higher-cognitive objectives aiming at the application and transfer of knowledge to real-life situations [3]. However, practical course constraints usually limit the exposure of students to realistic scenarios, which may hinder them to learn on how to apply the concepts.

As a solution, educational games have been introduced as instructional strategy for SE education in order to achieve learning on higher levels more effectively [1, 4, 5, 6]. Games for SE education provide a practical experience to the students in a safe and controlled environment [8], for example through simulation games (e.g., SimSE [9], SCRUMIA [9]).

Educational games are supposed to be an effective and efficient instructional strategy for teaching and learning [1, 10]. However, these claims seem to be questionable [1, 4, 5]. Often games seem to lack either the expected learning impact and/or the engagement they promise [1, 4, 5]. Thus, these claims seem not rigorously established as most evaluations of SE games are performed in an ad-hoc manner in terms of research design, measurement, data collection & analysis [1, 4, 5, 6].

So far there have been only few attempts to provide a more systematic support for the evaluation of educational games [11]. An exception is MEEGA (Model for the Evaluation of Educational Games) [12] a well-defined model developed for the evaluation of educational games. The model measures the reaction of students after they played the game by applying a standardized questionnaire answered by the students. It has been developed by using the GQM (Goal/Question/Metric) approach [13] to explicitly define a measurement program for evaluating three quality factors of educational games: motivation, user experience, and learning [14]. Currently, MEEGA seems to be one of the most widely used evaluation models in practice [4, 5, 11].

However, although, having been developed systematically, a question that remains is if the MEEGA model allows to evaluate the quality of SE games in a reliable and valid manner. An initial evaluation of the MEEGA questionnaire [12, 14] based on four case studies with a sample size of 169 students on four educational games [14], indicated a satisfactory reliability (Cronbach's alpha coefficient α=0.8). However, in terms of validity, a clear correlation within the quality factors motivation and user experience was not identified. No factor analysis was performed due to the small sample size and, thus, it was not possible to determine empirically how many factors underlie the MEEGA questionnaire [12]. These results indicate the need for further analysis with a larger data set to obtain more significant results.

Therefore, we conduct a large-scale evaluation of the MEEGA questionnaire with respect to its reliability and construct validity. Reliability and construct validity are fundamental issues with respect to measurement instruments such as questionnaires [15, 16]. In this context, reliability refers to the degree of consistency or stability of the instrument items on the same quality factor. Internal consistency reliability is estimated in order to assess the consistency of results across items within a questionnaire [15, 17] through Cronbach's alpha coefficient [18]. Construct validity of an instrument is defined as its ability to actually measure what it purports to measure, involving convergent and discriminant validity, which is obtained through the degree of correlation between the instrument items [15, 17]. In order to evaluate the MEEGA questionnaire, we follow the procedure proposed by DeVellis [16] using a large-scale data set from 43 case studies, evaluating 20 different SE games in different higher computing education institutions and professional IT training, involving a total population of 723 students.

II. Background: Evaluation Model MEEGA

MEEGA is a model specifically developed for the evaluation of educational games [12]. The model focuses on the evaluation of educational games (including digital as well as non-digital games such as card or board games). It enables a quick evaluation with minimal interruption to the instructional unit providing useful feedback without requiring detailed knowledge of educational theory, measurement or statistics from the instructor. To facilitate its application, the model provides a questionnaire for collecting data on the reaction of the students after they played an educational game.

Following the empirical study process [19] and the guide for the development of measurement scales [16], MEEGA was developed by systematically decomposing quality factors using the GQM (Goal/Question/Metric) approach [13]. Then, the quality factors were refined into a set of dimensions from which the questionnaire items are derived.

Here we assume that the quality of a game is achieved if it provides a positive learning effect, motivates students to study and provides a pleasant and engaging learning experience. Accordingly, the goal of MEEGA is to evaluate an educational game with respect to motivation, user experience and learning from the viewpoint of the learners in the context of an instructional unit. The quality factor motivation was decomposed based on the ARCS model [20], a well-known model of motivation that has also been used in several studies to assess the motivation of students utilizing educational resources. The ARCS model decomposes motivation into four dimensions: attention, relevance, confidence and satisfaction. Attention refers to students' cognitive responses to instructional stimuli. Relevance refers to the students' need to realize that the educational proposal is consistent with their goals and that they can link content with their professional future. Confidence means to enable students to make progress in the study of educational content through their effort and ability (e.g., through exercises with increasing level of difficulty). Satisfaction means that the students feel that the dedicated effort results in learning.

The quality factor user experience was decomposed into [21, 22, 23]: immersion, challenge, social interaction, fun, and competence/control. Immersion allows the player to have an experience of deep involvement within the game, creating a challenge with real-world focus, so that they forget about the outside world during gameplay. Challenge means that a game needs to be sufficiently challenging with respect to the players' competency level. Social interaction refers to the creation of a feeling of shared environment and being connected with others in activities of cooperation or competition. Games should also provide feelings of fun, enjoyment, pleasure, relaxation, recreation and satisfaction. When playing becomes something special to the player, it will provide a strong positive experience, accompanied by the desire to rejoin the game and recommend it to others. The game should also allow the player to have a sense of control over the game interactions, which should be easy to learn and allow them to explore the game freely and at their own pace. To provide a good experience, games should support the development and mastery of competencies so it is possible to overcome the challenges of the game [21, 22, 23].

The quality factor learning is measured in relation to the first three levels of the revised version of Bloom's taxonomy (remembering, understanding and applying) [24] including two dimensions with respect to short-term and long-term learning based on the assessment model by Sindre and Moody [25]. The evaluation of short-term learning is based on the more immediate educational goals, whereas long-term learning focuses on the contribution to the professional life of the individual. Each of these quality factors is further refined in dimensions as presented in Fig. 1.

Based on the defined quality factors a questionnaire has been developed customizing and unifying existing standardized questionnaires [20, 21, 22, 23, 25, 26, 27]. The questionnaire items are presented in Table VI. The response format for each of these standardized items is based on a 5-point Likert scale ranging from strongly disagree to strongly agree [16].

In addition to the standardized items, in order to capture feedback on the learning effect related to three levels of the revised version of Bloom's taxonomy (remembering, understanding and applying) [24], the questionnaire also collects data on the learner's perception on the achievement of specific instructional objective(s) of the game through customized items for each game. The learners rate their perceived level of knowledge before and after the game for each of the concepts/methods to be learned on a 5-point interval scale ranging from 1.0 to 5.0.

MEEGA also provides a spreadsheet for the analysis of the collected data. The spreadsheet assists in the organization of information and automatic generation of graphics showing the results of the evaluations.

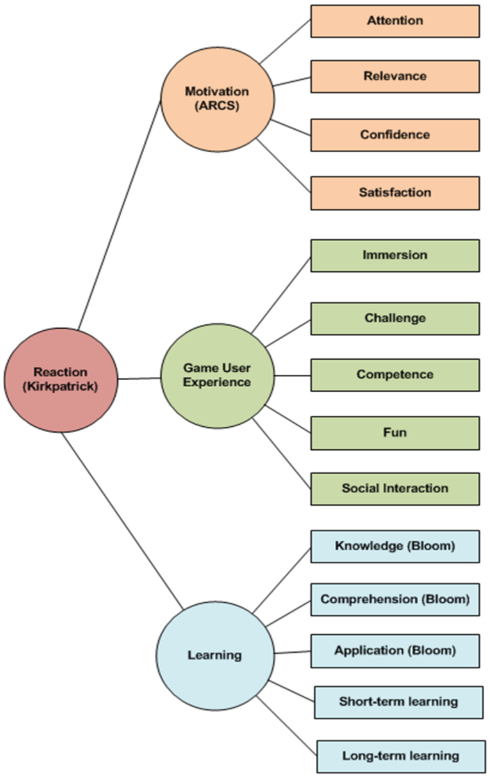
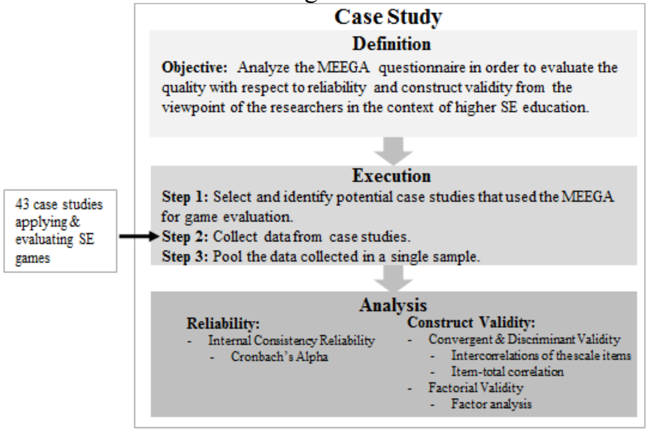


Figure 1. Decomposition of the quality factors [12].

The MEEGA model is intended to be used in case studies with a one-shot post-test only design, in which the case study begins with the application of the treatment (educational game) and after the game play the MEEGA questionnaire is answered by the learners in order to collect the respective data.

III. RESEARCH METHOD

In order to perform a large-scale evaluation of the MEEGA questionnaire, we conduct a case study [19, 28] structured as illustrated in Fig. 2.



Case Study

Definition

Objective: Analyze the MEEGA questionnaire in order to evaluate the quality with respect to reliability and construct validity from the viewpoint of the researchers in the context of higher SE education.

Execution

Step 1: Select and identify potential case studies that used the MEEGA for game evaluation.
Step 2: Collect data from case studies.
Step 3: Pool the data collected in a single sample.

43 case studies applying & evaluating SE games

Analysis

Reliability:
- Internal Consistency Reliability
  - Cronbach's Alpha

Construct Validity:
- Convergent & Discriminant Validity
  - Intercorrelations of the scale items
  - Item-total correlation
- Factorial Validity
  - Factor analysis

Figure 2. Research method.

Following the GQM approach [13], the study objective is decomposed into quality aspects and analysis questions to be analyzed based on the data collected in the game evaluations [16, 17].

The execution phase was organized in three steps. First, we identified and selected potential case studies by searching via Google and Google Scholar for articles that reported the usage of the MEEGA model for the evaluation of SE games. As result, we identified 43 case studies (Table I). Then, we contacted (via email) the authors requesting the collected data. In step 3, we pooled the data collected in a single sample for data analysis.

In the analysis phase, we performed a statistical evaluation in order to answer the analysis questions, following the procedure proposed by Trochim and Donnelly [17] and the scale development guide proposed by DeVellis [16]. In terms of reliability, we measured internal consistency reliability based on the correlations between different items on the same questionnaire [15, 17]. We measured internal consistency through Cronbach's alpha, a popular method to assess the reliability of the questionnaire [15].

In terms of construct validity, convergent and discriminant validity are the two subtypes of validity that make up construct validity [17]. Convergent validity refers to the degree to which two items of quality factors that theoretically should be related, are in fact related. In contrast, discriminant validity tests whether concepts or measurements that are supposed to be unrelated are in fact unrelated [17]. In order to analyze the convergent and discriminant validity of the MEEGA questionnaire, the intercorrelations of the items and item-total correlation is calculated [16]. Intercorrelation refers to the degree of correlation between the items of a measurement instrument [15, 16]. The higher the correlations among items that measure the same quality factor, the higher the validity of individual items and, hence, the validity of the instrument as a whole. Item-total correlation is analyzed in order to check if any item in the questionnaire is inconsistent with the averaged correlation of the others, and thus, can be discarded [15, 16].

In addition, factor analysis was used to determinate how many factors underlie the set of items of the MEEGA questionnaire, following the analysis process proposed by Brown [29]. Each factor is defined by those items that are more highly correlated with each other than with other items. A statistical indication of the extent to which each item is correlated with each factor is given by the factor loading. Thus, the higher the factor loading, the more the particular item contributes to the given factor. Thus, factor analysis also explicitly takes into consideration the fact that the items measure a factor unequally [15].

IV. DEFINITION AND EXECUTION OF STUDY

The objective of this study is analyze the MEEGA questionnaire in order to evaluate its quality in terms of reliability and construct validity from the viewpoint of the researchers in the context of higher SE education and professional IT training.

From this objective, we derive the following analysis questions:

Reliability

AQ1: Is there evidence for internal consistency of the MEEGA questionnaire?

Construct Validity

AQ2: Is there evidence of convergent and discriminant validity of the MEEGA questionnaire?

AQ3: How do underlying factors influence the responses on the items of the MEEGA questionnaire?

In order to maximize the sample size, we collected data from case studies that evaluated SE games in computing courses in higher education and professional IT trainings using MEEGA.

As a result, we obtained data from 43 case studies, with responses from a total of 723 students in 6 different contexts/institutions as summarized in Table I.

TABLE I. SUMMARY OF CASE STUDIES.

Game	Game type	SE Knowledge [36]	Course/Semester	Institution/ Country	Sample size
Dealing with difficult people	Non-digital	SE Management	Project planning and management/2013-1	UFSC/ Brazil	14
			Project management/2013-1		28
			Project planning and management/2015-2		23
DELIVER!	Non-digital	SE Management	Project planning and management/2010-2	UFSC/ Brazil	15
			Project management/2010-2		13
DOJO	Digital	SE Management	Project management/2013-1	UDESC/ Brazil	19
EAReqGame	Digital	Software Requirements	Software Engineering/2014-2	UFSM/ Brazil	14
Paper Tower	Non-digital	SE Management	Project management/2013-1	UDESC/ Brazil	4
PERT-CPM Game	Non-digital	SE Management	Project management	UNISUL /Brazil	5
PizzaMia	Non-digital	SE Management	Project management/2013-1	UDESC/ Brazil	17
			Project management/2014-1		19
			Project management/2015-1		13
PMMaster	Non-digital	SE Management	Project planning and management/2010-2	UFSC/ Brazil	7
			Project management/2010-2		16
			Project planning and management/2012-1		21
			Project management/2012-1		33
			Project planning and management/2015-1		17
			Project planning and management/2015-2		12
PMQuiz	Digital	SE Management	Project planning and management/2015-1	UFSC/ Brazil	20
			Project management/2015-1		13
			Project planning and management/2015-2		18
			Project management/2015-2		20
Project Detective	Non-digital	SE Management	Project planning and management/2011-2	UFSC/ Brazil	18
			Project management/2011-2		31
			Project planning and management/2013-1		13
Risk Game	Non-digital	SE Management	Project management/2013-1	UDESC/ Brazil	15
Risk Management Game	Non-digital	SE Management	Project planning and management/2015-2	UFSC/ Brazil	18
Schedule and Risk Game	Non-digital	SE Management	Project management/2014-1	UDESC/ Brazil	5
SCRUM'ed	Digital	SE Management	Project planning and management/2015-1	UFSC/ Brazil	23
SCRUMIA	Non-digital	SE Management	Project planning and management/2010-2	UFSC/ Brazil	16
			Project management/2010-2		12
			Project planning and management/2011-1		15
			Project management/2011-1		30
			Project planning and management/2015-1		13
			Project planning and management/2015-2		18
			Agile Methods/2013-1	UDESC/ Brazil	23
SCRUM-SCAPE	Digital	SE Management	Project planning and management/2013-2	UFSC/Brazil	17
ThatPMGame	Digital	SE Management	Project management/2013-1	UDESC/ Brazil	6
			Project management/2013-1		13
TRIVIAL PURSUIT – IFPUG FPA	Non-digital	SE Management	Training course on IFPUG FPA v4.2	Engineering.IT/ Italy	14
			Training course on IFPUG FPA v4.2		5
UsabiliCity	Digital	Software Design	Human-Computer Interaction/2014	Uninorte /Brazil	37
XPEnigma	Non-digital	SE Management	Project management/2013-1	UDESC/ Brazil	20

Data collected in the selected case studies (Table I) were pooled in a single sample, thus, used them cumulatively only in order to validate the model MEEGA (and no specific game). The pooling of data was possible due to the similarity of the selected case studies and standardization of the data collection. The similarity and standardization in terms of definitions, methods, and measurements are essential aspects for the pooling of data [30]. In this respect, the selected studies are similar in terms of definition (with the objective to evaluate an educational SE game with respect to motivation, user experience and learning), research design (case studies), and context (higher SE education and professional training). In addition, all selected case studies are standardized in terms of measures (quality factors/dimension), data collection method (MEEGA questionnaire), and response format (5-point Likert and 5-point interval).

V. ANALYSIS

We analyze each of the analysis questions as defined in the research methodology. In addition to the analysis results here, details on the data analysis can be found in a separate report [37].

A. Reliability

AQ1: Is there evidence for internal consistency of the MEEGA questionnaire?

We measured the internal consistency of the MEEGA questionnaire through Cronbach's alpha coefficient [20]. It indicates indirectly the degree to which a set of items measures a single quality factor. Thus, we want to know whether the MEEGA questionnaire measures the same quality factor, the reaction of students after they played and educational game. Typically, values of Cronbach's alpha, ranging from 0.70 to 0.95 are reported as acceptable [16], indicating internal consistency.

The MEEGA questionnaire is composed of two parts: the first part including the standardized items using a Likert scale and the second part including items to be customized to the specific instructional objective of a particular educational game. Due to this characteristic of the instrument, we analyzed the standardized and customized items separately.

Evaluation of the standardized items. Analyzing the 29 standardized items of the MEEGA questionnaire (Table VI), the value of Cronbach's alpha is satisfactory (α=.915). Results of a detailed analysis per quality factor (Table II) show that Cronbach's alpha is also acceptable with respect to each quality factor separately.

TABLE II. CRONBACH'S ALPHA PER QUALITY FACTOR

Quality factor	Cronbach's alpha
Motivation	.802
User Experience	.862
Learning	.797
Total	**.915**

This indicates that there exists an acceptable internal consistency, not only in terms of the reaction of the students to the educational game, but that there exists also an internal consistency of the items related to each of the quality factors (motivation, user experience, and learning). Thus, we can conclude that responses between the items are consistent and precise, indicating the reliability of the standardized items of the MEEGA questionnaire.

Evaluation of the Customized Items. Results of the analysis of the reliability of the customized items related to the learning objectives of particular educational game shows that Cronbach's alpha is high for the customized items (Cronbach's alpha α=.966). This indicates a high reliability also of this part of the MEEGA questionnaire.

B. Construct Validity

AQ2: Is there evidence of convergent and discriminant validity of the MEEGA questionnaire?

In order to obtain evidence of convergent and discriminant validity of the standardized items of MEEGA questionnaire, the intercorrelations of the items and correlation item-total are calculated [17].

Intercorrelations of the standardized items. In order to obtain evidence of convergent validity it is expected that the items of the same quality factor (e.g., motivation, user experience and learning) and same dimension (e.g., attention, satisfaction, fun, immersion, etc.) have a higher correlation [15]. On the other hand, to obtain evidence of discriminant validity is expected that the items of different quality factors (e.g., motivation and user experience) or dimensions have a low correlation [17], as in theory, the items are measuring different quality factors.

In order to analyze the intercorrelations between the standardized items of the same quality factor, we used the nonparametric Spearman correlation matrices for each quality factor (Table III to V). A complete matrix including all quality factors is presented in Table 6 of the supplementary material [37]. The matrices show the Spearman correlation coefficient, indicating the degree of correlation between two items (item pairs). We used this correlation coefficient as it is the most appropriate for Likert scales [16]. The correlation coefficients between the items within the same dimension are colored. In accordance to Cohen [31], a correlation between items is considered satisfactorily, if the correlation coefficient is greater than 0.29, indicating that there is a medium or high correlation between the items. Satisfactory correlations are marked in bold. The numbers of the items relate to the specification presented in Table VI.

TABLE III. SPEARMAN CORRELATION COEFFICIENT OF QUALITY FACTOR: MOTIVATION.

No. Item	1	2	3	4	5	6	7	8	9	10
	Attention			Relevance			Confidence		Satisfaction	
1	1.00									
2	**.367**	1.00								
3	**.339**	**.458**	1.00							
4	.247	.289	**.372**	1.00						
5	.269	.274	**.380**	**.404**	1.00					
6	.140	.240	.254	.272	**.304**	1.00				
7	.212	.172	.152	.209	.230	.230	1.00			
8	.203	**.298**	**.397**	**.322**	**.472**	.251	.268	1.00		
9	.231	**.335**	**.431**	**.426**	**.392**	.275	.198	**.472**	1.00	
10	.152	.255	.276	.233	.246	.272	.192	.267	**.374**	1.00

Analyzing the intercorrelations of the quality factor motivation (Table III), we can observe that 17 item pairs are correlated. This indicates that the motivation is fragmented into its various dimensions more independently (attention, relevance, confidence and satisfaction). Yet, items that belong to the same dimension show an acceptable degree of correlation in almost all item pairs, as neither negative values have been detected nor a group of items that consistently did not present a correlation with respect to all evaluated games. We, therefore, can observe a tendency for the items of one dimension to be correlated, thus, indicating evidence of convergent validity. On the other hand, some item pairs (e.g., 3-9, 5-8, 8-9), which do not belong to the same dimension also show an acceptable degree of correlation. Therefore, we cannot establish discriminant validity. However, in this case, the lack of an indication of discriminant validity is acceptable, as, although the subscale is fragmented into its various dimensions, all dimensions refer to a single quality factor (motivation).

With respect to the quality factor user experience, 69 item pairs are correlated (Table IV). This also indicates that the user experience is fragmented in its individual dimensions (immersion, social interaction, challenge, fun, competence, and digital game).

TABLE IV. SPEARMAN CORRELATION COEFFICIENT OF QUALITY FACTOR: USER EXPERIENCE.

No. Item/ Dimension	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
	Immersion			**Social Interaction**			**Challenge**		**Fun**				**Competence**		**Digital Game**	
11	1.000															
12	**.625**	1.000														
13	**.598**	**.637**	1.000													
14	.264	.239	.253	1.000												
15	**.398**	**.393**	**.364**	**.641**	1.000											
16	**.332**	**.354**	**.319**	**.556**	**.586**	1.000										
17	.287	**.291**	.260	.238	.279	**.303**	1.000									
18	**.377**	**.411**	**.420**	.287	**.396**	**.331**	**.455**	1.000								
19	**.412**	**.471**	**.416**	**.316**	**.548**	**.421**	**.359**	**.565**	1.000							
20	**.303**	**.400**	**.400**	.113	.265	.170	.265	**.335**	**.387**	1.000						
21	**.360**	**.382**	**.396**	.203	**.324**	.274	**.404**	**.482**	**.530**	**.454**	1.000					
22	**.302**	**.396**	**.372**	.157	.276	.236	**.349**	**.418**	**.461**	**.489**	**.680**	1.000				
23	**.301**	.310	**.352**	.142	.168	.210	**.292**	**.363**	**.332**	.279	**.369**	**.343**	1.000			
24	**.370**	**.376**	**.401**	.168	**.307**	.283	**.401**	**.445**	**.465**	**.374**	**.466**	**.457**	**.496**	1.000		
25	-142	-.101	-.113	-.220	-.128	-.118	-.173	-.208	-.078	.053	-.079	-.037	-.011	-.060	1.000	
26	-.148	-.104	-.128	-.200	-.119	-.086	-.153	-.210	-.054	.037	-.086	-.042	-.023	-.097	**.784**	1.000

Again, items belonging to the same dimension show a satisfactory correlation with respect to all pairs of items. However, items 25 and 26 (digital game dimension) present negative correlation values, showing that these items do not have a correlation with other dimensions, and, therefore, seem not to be measuring user experience. Although, within the dimension (digital game), the items showed a satisfactory correlation. Thus, indicating that these items are measuring the characteristics of digital games, yet, they seem not to be related to the quality factor user experience.

The items belonging to the same dimension showed a high degree of correlation, thus, we can establish a convergent validity. However, several item pairs of different dimensions also showed a high degree of correlation. Thus, again we cannot establish discriminant validity.

TABLE V. SPEARMAN CORRELATION COEFFICIENT OF QUALITY FACTOR: LEARNING.

	27	**28**	**29**
	Short-term Learning		**Long-term Learning**
27	1.000		
28	**.620**	1.000	
29	**.528**	**.460**	1.000

Analyzing the intercorrelations of the quality factor learning, all three item pairs of the questionnaire are correlated (Table V). This quality factor demonstrated a greater correlation between items than the other two quality factors.

Summarizing, we can observe that in general, there exists a correlation between items within a single dimension with respect to all three quality factors. This indicates that, considering the dimensions, a convergent validity can be established. On the other hand, few item pairs present lower or negative intercorrelations, which indicate that these items need be revised or re-grouped to other quality factors or dimensions. In general, items of different dimensions within a quality factor also present a medium or high correlation, thus, no discriminant validity could be established. This also can be observed when analyzing the matrix that shows the intercorrelations between all quality factors (Table 6 of the supplementary material [37]). However, the lack of an indication of discriminant validity is acceptable, as all dimensions at the end refer to a single quality factor (the reaction of students after they played and educational game).

Item-total correlation. This test is complementary to the previous one in order to evaluate the correlation with all the other items. Each item of the instrument should have a medium or high correlation with all the other items [16] as this indicates that the items present consistency in comparison to the other items. On the other hand, a low item-total correlation of an item undermines the validity of the scale, and, therefore, should be eliminated.

For this analysis we used the method of corrected item-total correlation, which compares one item with every other one of the instrument, excluding itself. Reference values for the analysis are the same as presented in the previous section following Cohen [31], considering a correlation satisfactorily, if the correlation coefficient is greater than 0.29. In addition, we analyze the Cronbach's alpha if an item was deleted expecting that no item should cause a substantial decrease in the Cronbach's alpha [18].

In general, the correlations are medium to high considering reference values as defined by Cohen [31]. Only item 25 ("The controls to perform actions in the game responded well", item-total correlation ρ= -.098) and 26 ("It's easy to learn how to use the interface and game controls", item-total correlation ρ= -.124) (digital game dimension) present a low item-total correlation. These results show that these items do not have a satisfactory correlation with the others dimensions and, thus, indicates that they are not measuring user experience. In addition, the value of Cronbach's alpha if the items were deleted shows a small increase. Consequently, these items need to be revised and possibly excluded. All other items demonstrated sufficient item-total correlation and satisfactory value of Cronbach's alpha if the item was deleted, thus, indicating the validity of the measured quality factors. The detailed results of the item-total correlation of all items are presented in Table 7 of the supplementary material [37].

The degree of correlation between the items shows whether the items measure (or not) the same quality factor/dimension, thus indicating evidence of convergent and discriminant validity. However, these results do not determine how many quality factors underlie the set of the MEEGA questionnaire. With this objective, we performed a factor analysis, answering analysis question AQ3.

AQ3: How do underlying factors influence the responses on the items of the MEEGA questionnaire?

In order to identify the number of factors (quality factors or dimensions) that represents the responses of the set of the 29 standardized items of the MEEGA questionnaire we performed a factor analysis. Based on the original definition of the MEEGA questionnaire we assume that it is influenced by three underlying factors (motivation, user experience and learning).

In order to analyze whether the items of the MEEGA questionnaire can be submitted to a factor analysis process [29], we used the Kaiser-Meyer-Olkin (KMO) index and Bartlett's test of sphericity being the most commonly used ones [29]. The KMO index measures the sampling adequacy with values between 0 and 1. An index value near 1.0 supports a factor analysis and anything less than 0.5 is probably not amenable to useful factor analysis [29]. Bartlett's sphericity test also indicates whether factor analysis is appropriate with values of a significance level < 0.05 are considered acceptable [29]. Analyzing the set of items of the MEEGA questionnaire, we obtained a KMO index of .914 and a significance level of 0.000, indicating that factor analysis is appropriate in this case.

Applying factor analysis, the number of factors retained in the analysis is decided [29]. Here we used the Kaiser-Guttman criterion [29] for this decision as the most commonly used method. This method states that the number of factors is equal to the number of eigenvalues greater than 1 [29]. The eigenvalue refers to the value of the variance of the all the items which is explained by a factor [29]. Following the Kaiser-Guttman criterion, our results show that 6 factors should be retained, explaining 59.65% of the data. The scree plot (Fig. 3) shows the eigenvalue for each factor number (representing each item).

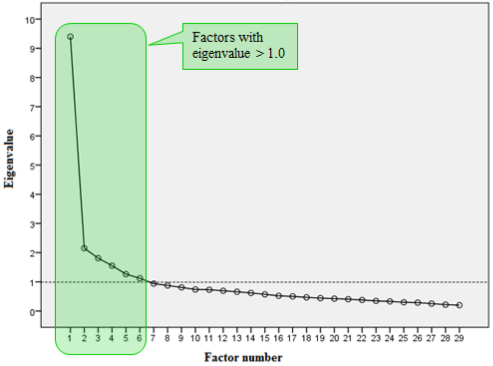


Figure 3. Scree Plot

Thus, the results of the factorial analysis indicate a decomposition into 6 factors, different from than the assumed one into three quality factors (motivation, user experience and learning) as proposed in the original MEEGA model.

Once identified the number of underlying factors, another issue is to determine which items are loaded into which factor. In order to identify the factor loadings of the items we used the Varimax with Kaiser Normalization rotation method being the most widely accepted and used rotation method [31]. Table VI shows the factor loadings of the items associated with the 6 retained factors. The highest factor loading of each item, indicating to which factor the item is most related, is marked in bold.

TABLE VI. FACTOR LOADINGS

Quality factor	Dimension	No.	Description	Factor					
				1	2	3	4	5	6
Motivation	Attention	1	The game design is attractive	.045	.082	**.668**	.092	.037	.101
		2	There was something interesting at the beginning of the game that captured my attention	.108	.224	**.625**	.118	.163	-.119
		3	The variation (form, content or activities) helped me to keep attention to the game	.260	.288	**.461**	.226	.181	-.116
	Relevance	4	The game content is relevant to my interests	**.484**	-.028	.418	.086	.152	.025
		5	The way the game works suits my way of learning	**.554**	.093	.274	.015	.298	-.099
		6	The game content is connected to other knowledge I already had	.155	-.015	.253	-.041	**.559**	-.085
	Confidence	7	It was easy to understand the game and start using it as study material	.098	-.117	**.368**	.221	.393	.205
		8	Passing through the game, I felt confident that I was learning	**.557**	.182	.247	.057	.337	-.147
	Satisfaction	9	I am satisfied because I know I will have opportunities to use in practice things I learned playing this game	**.586**	.082	.238	.259	.300	-.008
		10	It is due to my personal effort that I manage to advance in the game	.117	.209	-.059	.203	**.727**	.084
User Experience	Immersion	11	Temporarily I forgot about my daily; I have been fully concentrated on the game	.106	**.790**	.088	.170	.130	-.085
		12	I did not notice the time pass while playing; when I saw the game had already ended	.166	**.819**	.168	.140	.078	-.036
		13	I felt myself more in the game context than real life, forgetting what was around me	.186	**.768**	.176	.129	.155	-.055
	Social Interaction	14	I was able to interact with others during the game	.081	.049	.082	**.846**	.091	-.180
		15	I had fun with other people	.149	.250	.190	**.821**	.019	-.050
		16	The game promotes cooperation and/or competition among the players	.147	.145	.146	**.778**	.094	-.019
	Challenge	17	This game is appropriately challenging for me, the tasks are not too easy nor too difficult	**.345**	.125	.285	.207	.311	-.175

		18	The game progresses at an adequate pace and does not become monotonous - offers new obstacles, situations or variations in its tasks	.277	.351	**.382**	.225	.240	-.222
	Fun	**19**	I had fun with the game	.269	.363	**.487**	.345	.115	.007
		20	When interrupted at the end of the class, I was disappointed that the game was over	.263	.467	**.400**	.004	.063	.129
		21	I would recommend this game to my colleagues	.508	.248	**.574**	.087	.081	.003
		22	I would like to play this game again	**.506**	.273	.496	.085	.024	.047
	Competence	**23**	I achieved the goals of the game applying my knowledge	.231	.298	.073	.011	**.685**	.019
		24	I had positive feelings on the efficiency of this game	**.427**	.361	.306	.012	.349	-.020
	Digital game	**25**	The controls to perform actions in the game responded well	-.040	-.022	-.001	-.124	.003	**.905**
		26	It's easy to learn how to use the interface and game controls	-.115	-.070	.040	-.083	.019	**.908**
Learning	**Short-term learning**	**27**	The game contributed to my learning in this course	**.820**	.095	.077	.104	.120	-.017
		28	The game was efficient for my learning, comparing it with other activities of the course	**.768**	.175	-.055	.084	.094	-.037
	Long-term learning	**29**	The experience with the game will contribute to my professional performance in practice	**.726**	.189	.139	.119	.002	-.071

Analyzing the factor loadings of the items (Table VI), we can observe that, the first factor (factor 1), includes a set of 10 items (4, 5, 8, 9, 17, 22, 24, 27, 28 and 29), including items from three original quality factors (motivation, user experience, and learning). However, this seems to suggest that these items are related to the quality factor learning. Although the items 4, 5, 8, 9, 17, 22 and 24 are originally related to other dimensions (relevance, confidence, satisfaction, challenge, fun, and competence), the description of the most of these items also involves learning and, thus, seems justifying this relation.

With respect to factor 3, it includes a set of 8 items (1, 2, 3, 7, 17, 18, 19, and 20). These items refer to four different dimensions of the original questionnaire (attention, confidence, fun, challenge), measuring two different quality factors (motivation and user experience). Similarly, factor 5, includes a set of 3 items (6, 10, and 23). Originally the items of factor 5 are related to three different dimensions (relevance, satisfaction, and competence), and two different quality factors (motivation and user experience). Thus, the results of factor 3 and 5 suggest that the original classification of MEEGA model may not be the most appropriate and needs to be revised/redesigned.

Analyzing the results of factor 2, we can observe that this factor is composed by a set of three items (11, 12, 13), all three of the same dimension (immersion) as proposed by the original MEEGA model. The same can be observed with respect to factor 4 being mainly composed by a set of three items (14, 15 and 16), all of which are referring to the original dimension of social interaction. Factor 6 also is composed of a set of two items (25, 26) both related to the dimension of digital game, as originally proposed in MEEGA model.

In summary, the results of the factor analysis partially confirm the structure of the original MEEGA model clearly indicating a quality factor related to learning (yet, involving more items than defined originally). On the level of dimensions, the factor analysis also confirmed the composition of the dimensions of immersion, social interaction and digital game. The results however also clearly indicate a need for restructuring not only in terms of the number of underlying quality factors, but also with respect to the grouping of the items with respect to the dimensions and quality factors. Especially items related to the quality factors motivation and user experience need to be revised, as the results of the factor analysis indicate that they may not be related to a different quality factors as proposed in the original model. Thus, these two quality factors (motivation and user experience) seems to be integrated into one experience factor.

VI. Discussion

The obtained results show sufficient evidence to consider the reliability and construct validity of MEEGA acceptable as a model for the evaluation of SE games. In terms of reliability (AQ1), the results of the analysis indicate a satisfactory Cronbach's alpha for all quality factors (Cronbach's alpha α=.915), indicating the internal consistency of the MEEGA questionnaire. Thus, showing that the items of MEEGA questionnaire are consistent and precise with respect to the evaluation of SE games.

In terms of construct validity, with respect to convergent validity (AQ2), we identified that the items belonging to the same dimension, in general, presented a higher correlation. This indicates that the quality factors motivation and user experience are complex and are not constituted by a single trait, but are actually fragmented into individual dimensions. With respect to the quality factor learning, it was also possible to identify a correlation between all items. Thus, we can conclude that, considering the dimensions, there exists evidence of convergent validity. This means, that regarding the dimensions, the items seem to be measuring what the questionnaire purports to measure (e.g., attention, immersion, social interaction, etc.). However, few item pairs present low or negative intercorrelations (e.g., items pairs 4-6, 5-6 and 7-8), which indicates that these items need be revised or re-grouped to other quality factors or dimensions.

With respect to discriminant validity (AQ2), in general, items belonging to different quality factors and items of different dimensions within a quality factor also present a medium or high correlation. Thus, no discriminant validity could be established. However, we consider that the lack of an indication of discriminant validity is acceptable, as all dimensions in the end refer to a single quality factor measuring the reaction of students after they played an educational game as proposed in the MEEGA.

Analyzing the item-total correlation, the results also indicate a sufficient item-total correlation of most items of the MEEGA questionnaire. Only item 25 and 26, corresponding to the dimension on digital game, showed a negative item-total correlation. In addition, the value of Cronbach's alpha if these items were deleted also indicated that they are not measuring the same quality factor (user experience). Thus, these items need be revised or excluded from the MEEGA questionnaire.

Based on the factor analysis (AQ3), we identified that 59.65% of the data variability is explained by 6 factors. Three factors (2, 4 and 6 (Table VI)) showed, as a result, the suggestion of a quality factor/dimension (e.g., learning, social interaction, immersion, and digital game) as originally proposed in MEEGA model. In these cases, the items of dimensions of social interaction, immersion, and digital game demonstrate a high factor loading with respect to the same factor. Thus, indicating that these items adequately measure the corresponding dimension. Yet, another three factors (1, 3, and 5, (Table VI)) do not clearly suggest a quality factor or dimension. Although, factor 1 seems to be related to learning, as items 27, 28, and 29 (originally related to the quality factor learning) have a high factor loading in this factor (factor 1). In addition, items 4, 5, 8 and 9 (related to motivation in the MEEGA model) and items 17, 22, and 24 (related to user experience) seem to belong rather to a different quality factor (learning) than original proposed (motivation, user experience). Thus, these items need to be either re-grouped or re-formulated, as another possible reason for the observed behavior may be the free translation to Brazilian Portuguese from the original ARCS measurement instrument [32].

Analyzing the items related to factor 3 and 5, we also identified as potential explanation for the allocation of the items to different factors the free translation (to Brazilian Portuguese) and adaptation of the original items to the context of educational games. E.g., The item ('The wording of feedback after the exercises, or of other comments in this lesson, helped me feel rewarded for my effort') from the original ARCS questionnaire [32] is represented through item 10 in the MEEGA questionnaire ('It is due to my personal effort that I manage to advance in the game'). Another possible indication may in fact be the need for a re-grouping of these items as also pointed out by other empirical studies evaluating the standardized questionnaires used in the development of MEEGA [33, 34]. However, based on the results, the items of the quality factors motivation and user experience seems to overlap conceptually and, thus, need to be revised with respect to their wording and classification to a quality factor/dimension.

Threats to validity. Due to the characteristics of this type of research, this work is subject to various threats to validity. We, therefore, identified potential threats and applied mitigation strategies in order to minimize their impact on our research.

Construct validity. Some threats are related to the design of the study [19]. In order to mitigate this threat, we defined and documented a systematic methodology for our study using the GQM approach [13]. Another risk is related to the omission of existing data sets related to the evaluation of the MEEGA model. In order to mitigate this risk, we searched for existing evaluation studies using the MEEGA model via Google and Google Scholar representing broad search engines. We included data sets from all studies we encountered and for which we received the collected data. Another risk refers to the quality of the data pooled into a single sample, in terms of standardization of data (response format) collected and adequacy to MEEGA model. As our study is limited exclusively to evaluations that used the MEEGA model the risk is minimized as in all studies the same data collection instrument has been used. Another issue refers to the pooled data from different contexts. To mitigate this threat we selected studies considering only the context of higher education and professional training with respect to only one knowledge area: Software engineering.

External validity. In terms of external validity, a threat to the possibility to generalize the results is related to the sample size and diversity of the data used for the evaluation. In respect to sample size, our evaluation used data collected from 43 case studies evaluating 20 different digital and non-digital SE games, involving a population of 723 students. In terms of statistical significance, this is a satisfactory sample size allowing the generation of significant results. The data has been obtained from game applications in 6 different institutions/contexts. However, as the data collection was restricted to evaluation that used the MEEGA questionnaire for data collection, the majority of the data came from Brazil where it is used more prominently, with only one application from an organization in Italy.

Reliability. In terms of reliability, a threat refers to what extent the data and the analysis are dependent on the specific researchers. In order to mitigate this threat, we documented a systematic methodology, defining clearly the study objective, the process of data collection, and the statistics methods used for data analysis. Another issue refers to the correct choice of statistical tests for data analysis. To minimize this threat we performed a statistical evaluation based on the approach for the construction of measurement scales as proposed by DeVellis [16], which is aligned with procedures for the evaluation of internal consistency and construct validity of measurement instruments [17].

VII. CONCLUSIONS

The results from our large-scale evaluation of the MEEGA model for the evaluation of SE games with respect to its reliability and construct validity indicate that the MEEGA questionnaire is acceptable in terms of reliability and construct validity. A Cronbach's alpha α=.915 indicates an acceptable internal consistency, which means that the responses between the items are consistent and precise. Our analysis also indicates convergent validity through an acceptable degree of correlation found between almost all items regarding the dimensions of the quality factors. Thus, indicating that the MEEGA model can be a reliable and valid measurement instrument for evaluating SE games. However, as a result of the factor analysis, we also found that 6 underlying factors influence the responses on the items of MEEGA questionnaire different from the original structure defining three factors (motivation, user experience and learning). This indicates a need for the redesign of the model or re-grouping of items within the instrument mainly with respect to the quality factors motivation and user experience. Continuing this research, we are currently re-designing the MEEGA model

[35] in order to improve the feedback on the quality of games for SE education.

Acknowledgments

We would like to thank all researchers, who shared their data on game evaluations.

This work was supported by the CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico* – www.cnpq.br), an entity of the Brazilian government focused on scientific and technological development.

References

[1] C. Gresse von Wangenheim and F. Shull, F. "To game or not to game?", Software, IEEE, vol. 26, no. 2, 2009, pp. 92-94.

[2] Y. Sedelmaier and D. Landes. "Active and Inductive Learning in Software Engineering Education", Proc. of the 37th IEEE International Conference on Software Engineering, 2015, pp. 418-427. Florence, Italy.

[3] P. Parsons. "Preparing computer science graduates for the 21st Century", Teaching Innovation Projects, vol. 1, n. 1, article 8, 2011.

[4] A. Calderón and M. Ruiz. "A systematic literature review on serious games evaluation: an application to software project management", Computers & Education, vol. 87, 2015, pp. 396-422.

[5] G. Petri and C. Gresse von Wangenheim. "How games for computing education are evaluated? A systematic literature review", Computers & Education, vol. 107, 2017, pp. 68-90.

[6] P. Battistella and C. Gresse von Wangenheim. "Games for teaching computing in higher education – a systematic review", IEEE Technology and Engineering Education Journal, vol. 9, no. 1, 2016, pp. 8-30.

[7] A. Baker, E. O. Navarro, and A. Van der Hoek, A. "An experimental card game for teaching software engineering processes*", Journal of Systems and Software*, vol. 75, no. 1–2, 2005, pp. 3-16.

[8] E. O. Navarro and A. Van der Hoek. "Design and evaluation of an educational software process simulation environment and associated model", Proc. of the 8th Conference on Software Engineering Education and Training, 2005, pp. 25-32. Ottawa, Canada.

[9] C. Gresse von Wangenheim, R. Savi, and A. F. Borgatto. "SCRUMIA - An educational game for teaching SCRUM in computing courses", Journal of Systems and Software, vol. 86, no. 10, 2013, pp. 2675-2687.

[10] E. A. Boyle, T. M. Connolly, and T. Hainey, T. "The role of psychology in understanding the impact of computer games", Entertainment Computing, vol. 2, no. 2, 2011, pp. 69–74.

[11] G. Petri & C. Gresse von Wangenheim. "How to evaluate educational games: a systematic literature review", Journal of Universal Computers Science, vol. 22, no. 7, 2016, pp 992-1021.

[12] R. Savi, C. Gresse von Wangenheim, and A. F. Borgatto. "A model for the evaluation of educational games for teaching software engineering", Proc. of the 25th Brazilian Symposium on Software Engineering, 2011, pp. 194-203. São Paulo, Brazil (in Portuguese).

[13] V. R. Basili, G. Caldiera, and H. D. Rombach. Goal, Question Metric Paradigm. In J. J. Marciniak, Encyclopedia of Software Engineering, Wiley-Interscience, 1994, pp. 528-532. New York, NY, USA.

[14] R. Savi, C. Gresse von Wangenheim, A. F. Borgatto, L. Buglione, and V. R. Ulbricht. MEEGA – A model for the evaluation of games for teaching software engineering. Technical Report INCoD/GQS.01.2012.E, 2012. Available at: http://goo.gl/vJiQ8I. Accessed: 30 jul. 2016.

[15] E. G. Carmines and R. A. Zeller. Reliability and validity assessment, 5th ed. Beverly Hills: Sage Publications Inc, 1982.

[16] R. F. DeVellis. Scale development: theory and applications. SAGE Publications, 2003.

[17] W. M. Trochim and J. P. Donnelly. Research methods knowledge base, 3rd. ed., 2008. Mason, OH: Atomic Dog Publishing.

[18] L. J. Cronbach. "Coefficient alpha and the internal structure of tests", Psychometrika, vol. 16, no. 3, 1951, pp. 297–334.

[19] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. Experimentation in Software Engineering, 2012. Springer-Verlag Berlin Heidelberg.

[20] J. Keller. "Development and use of the ARCS model of motivational design", Journal of Instructional Development, vol. 10, no. 3, 1987, pp. 2-10.

[21] P. Sweetser and P. Wyeth. "GameFlow: a model for evaluating player enjoyment in games", Computers in Entertainment, vol 3, no. 3, 2005, pp. 1-24.

[22] K. Poels, Y. D. Kort, and W. Ijsselsteijn, W. "It is always a lot of fun!: exploring imensions of digital game experience using focus group methodology", Proc. of Conf. on Future Play, 2007, pp. 83-89. Toronto, Canada.

[23] J. Takatalo, J. Häkkinen, J. Kaistinen, and G. Nyman. Presence, Involvement, and Flow in Digital Games. In: Bernhaupt, R. (Ed.). Evaluating User Experience in Games: Concepts and Methods, 2010, pp. 23-46. Springer.

[24] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Longman, 2001.

[25] G. Sindre and D. Moody. "Evaluating the effectiveness of learning interventions: an information systems case study", Proc. of the 11th European Conf. on Information Systems, 2003, Paper 80. Naples, Italy.

[26] T. Tullis and W. Albert, W. Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Morgan Kaufmann, 2008.

[27] E. H. Gámez. On the Core Elements of the Experience of Playing Video Games (Dissertation). UCL Interaction Centre, Department of Computer Science, 2009. London, UK.

[28] R. K. Yin. Case study research: design and methods, 4th. ed., 2009. Sage Publications, Beverly Hills.

[29] T. A. Brown. Confirmatory factor analysis for applied research. New York: The Guilford Press, 2006.

[30] L. Kish. "Multipopulation survey designs: five types with seven shared aspects", International Statistical Review, vol. 62, no. 2, 1994, pp.167–186.

[31] J. Cohen. Statistical Power Analysis for the Behavioral Sciences. Routledge Academic, 1998.

[32] J. Keller. Motivational Design for Learning and Performance: The ARCS Model Approach. Springer, 2009.

[33] W. Huang, W. Huang, H. Diefes-Dux, and P. K. Imbrie. "A Preliminary Validation of Attention, Relevance, Confidence and Satisfaction Model-Based Instructional Material Motivational Survey in a Computer-Based Tutorial Setting", British Journal of Educational Technology, vol. 37, no. 2, 2006, pp. 243-259.

[34] M. Johnson, M. "A pilot study examining the motivational effect of instructional materials on EFL learning motivation", Journal of Language and Culture of Hokkaido, vol. 10, 2012, pp. 39-47.

[35] G. Petri, C. Gresse von Wangenheim, and A. F. Borgatto. MEEGA+: an evolution of a model for the evaluation of educational games. Technical Report, INCoD/GQS.03.2016.E., 2016. Available at: http://goo.gl/lPz1OY. Accessed: 12 aug. 2016.

[36] P. Bourque and R. E. Fairley. Swebok v3.0 Guide to the software enginnering body of knowledge., IEEE Computer Society, 2014.

[37] G. Petri, C. Gresse von Wangenheim, and A. F. Borgatto. Data analysis of a large-scale evaluation of a model for the evaluation of games for teaching software engineering. Working paper, WP_GQS_01.2016_v1., 2016. Available at: http://goo.gl/RYic6q