International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 16, Number 2

April – 2015

Perceived Usability Evaluation of Learning Management Systems: Empirical Evaluation of the System Usability Scale

Konstantina Orfanou[1], Nikolaos Tselios[1], and Christos Katsanos[2]
[1]University of Patras, Greece, [2]Hellenic Open University, Greece and Technological Educational Institute of Western Greece

Abstract

Perceived usability affects greatly student's learning effectiveness and overall learning experience, and thus is an important requirement of educational software. The System Usability Scale (SUS) is a well-researched and widely used questionnaire for perceived usability evaluation. However, surprisingly few studies have used SUS to evaluate the perceived usability of learning management systems (LMSs). This paper presents an empirical evaluation of the SUS questionnaire in the context of LMSs' perceived usability evaluation. Eleven studies involving 769 students were conducted, in which participants evaluated the usability of two LMSs (eClass and Moodle) used within courses of their curriculum. It was found that the perceived usability of the evaluated LMSs is at a satisfactory level (mean SUS score 76.27). Analysis of the results also demonstrated the validity and reliability of SUS for LMSs' evaluation, and that it remains robust even for small sample sizes. Moreover, the following SUS attributes were investigated in the context of LMSs evaluation: gender, age, prior experience with the LMS, Internet self-efficacy, attitude towards the Internet and usage frequency of the LMS.

Keywords: Usability evaluation; educational software; system usability scale; SUS; learning management system; LMS; e-learning

Introduction

Numerous usability evaluation questionnaires designed to measure users' subjective assessments of a system's usability (i.e., its perceived usability) are available at the moment (Brooke, 1996; Lewis, 1991, 1992, 1995; Tullis & Albert, 2008; Tullis & Stetson, 2004). Possibly, the most widely adopted is the System Usability Scale (SUS) (Brooke, 1996). This is mainly because it is characterized by a number of advantages.

SUS is very short; it comprises only 10 items to be rated on a five point scale ranging from strongly disagree to strongly agree, among which five are positive statements and the rest are negative. The small number of questions arises from Brooke's (1996) goal to construct a 'quick and dirty' usability evaluation questionnaire. In addition, the SUS score is single and ranges from 0 to 100. Therefore, its results can be comprehensible even to non-experts. Moreover, it has been found to be a remarkably robust measure of system usability (Bangor, Kortum, & Miller, 2008; Lewis & Sauro, 2009; Tullis & Stetson, 2004), even with a small sample size (Tullis & Stetson, 2004). In addition, SUS can be used to assess the usability of any software system, device or service. In other words, it is "technology agnostic" (Bangor et al., 2008). Another great advantage is that in contrast to other questionnaires, such as SUMMI, SUS is provided free of charge under the only condition of the reference of the source.

More and more teachers use LMSs to enhance their teaching methods (Torkzadeh & van Dyke, 2001). Asynchronous education or distance learning helps teachers and learners to overcome limitations of typical teaching such as space and time, and it also allows a personalized learning on the part of the learner. Thus, in recent years the interest is focused on web-based LMSs and usage of reliable usability evaluation questionnaires within LMSs should be considered as an effective option. However, an initial literature review reported in the following contrasted the aforementioned suggestion; few studies report effective usage of instruments such as SUS. It becomes apparent that the study of LMSs' usability is still at an early stage and the investigation of the SUS questionnaire within the context of LMSs' usability evaluation is important.

The goal of this paper is twofold: (a) to examine the applicability of the SUS method in the context of the perceived usability evaluation of LMSs, (b) to investigate associations (if any) between SUS score and students' characteristics, and in specific gender, age, previous experience with the LMS, Internet self-efficacy, Internet user attitude, and LMS usage frequency.

The paper is organized as follows: Initially, the research methodology, the profile of the participants and the LMSs (Moodle and Eclass, an open platform based on Claroline) which were the subject of evaluation are presented. Subsequently, the research results are presented focusing on the investigation of SUS within LMSs. The implications of these results are also discussed.

Related Work

SUS and Perceived Usability Evaluation of Products

SUS benchmark data.

Bangor et al. (2008) conducted an extensive research studying the usability evaluation of various products and services using the SUS questionnaire. Within a decade they gathered more than 2,300 individual surveys from more than 200 studies. It was found that the mean SUS score was 70.14 with a median of 75 (on a per study basis 69.69 and 70.91, respectively) and Cronbach's alpha was 0.91. Moreover, it was observed that no evaluated product received a score below 30, and fewer than 6% of study scores fell below 50.

Bangor et al. (2008; 2009) also wanted to add an interpretation of the SUS score. To this end, they compared the quartiles of the SUS score and adjective ratings of perceived usability. They found that if the SUS score is over 85 the system/product is highly usable, over 70 to 85 it is characterized from good to excellent, a value from about 50 to about 70 shows that the system is acceptable but it has some usability problems and needs improvement, and finally a system with SUS score below 50 is considered unusable and unacceptable.

In a recent survey, Kortum and Bangor (2013) assessed the perceived usability of 14 common everyday products using the SUS questionnaire. Their aim was to provide practitioners and researchers enough data (benchmarks) so that they would be able to characterize and present results of their own usability investigations. Surprisingly, e-learning platforms are neither included in the categories of web products chosen nor in the nine additional product categories that the authors mention that they would also study if they had the opportunity for a larger product list.

SUS validity and reliability.

Tullis and Stetson (2004) measured the usability of two websites using five different surveys, including the Questionnaire for User Interaction Satisfaction (QUIS), the SUS, the Computer System Usability Questionnaire (CSUQ), and two vendor specific surveys, and found that the SUS provided the most reliable results across a range of sample sizes. They reported that two of the questionnaires, SUS and CSUQ, achieved this goal faster, extracting the correct conclusion (which website had superior usability) to over 90% of cases, when the sample size was between 12 and 14. This result implies that usability studies applying the SUS questionnaire could have reliable results even with a small sample, as low as 12 participants.

In addition, studies (Bangor et al., 2008, 2009) have found significant correlation (from r=0.806 to r=0.822) between the SUS score and a 7-point adjective rating scale with different wordings that assessed the "user-friendliness" of the system; the worst rating was 'worst-imaginable',

whereas the best rating was 'best-imaginable'. This association strengthens the validity of the SUS questionnaire.

SUS attributes.

Bangor et al. (2008) examined the possible associations between gender or age and the final SUS score in a subset of their dataset (N = 213). They found a significant negative correlation between SUS score and age (r=-0.203, p=0.03), but no significant difference between the mean SUS scores obtained from women (mean = 71.6) and men (mean = 70.2).

Sauro (2011) examined the influence of prior experience with a website on the users' SUS scores. Using a large dataset from about 2,000 users who evaluated the usability of 62 websites and 16 software products, he found that users with previous experience rate the websites and software products from 6% to 15% more usable than first-time users.

In addition, Bangor et al. (2008) report that a one-way analysis of variance on the data referring to different type of devices (i.e., cell phones, customer equipment such as modems, GUIs, interactive voice response systems, and web pages/applications) showed that SUS scores do vary significantly ($p<0.001$) by the type of interface being tested.

SUS and Educational Software Evaluation

In the context of an e-learning platform evaluation, Renaut, Batier, Flory, and Heyde (2006) conducted a study in order to suggest how to use assessment methods based on observation and test. Their aim was to detect usability problems and to cultivate knowledge of web developers on their end-users and user centered design. The platform used was SPIRAL, realized in Lyon University. The researchers used the SUS scale as a post-test assessment of the usability of the platform. They found that 72% of the participating teachers described the platform as easy to use, but they do not cite specific SUS ratings. The overall conclusion of the study was that while teachers seemed to positively assess the usability of the platform, they formed their own way of using the platform which was different from the one that was envisaged from the platform's developers.

Ayad and Rigas (2010) evaluated three edutainment platforms in terms of user performance, learning effectiveness and satisfaction in order to explore usability aspects of educational entertainment in e-Learning. The three platforms were Virtual Classroom, Game-based and Storytelling. The SUS questionnaire was employed to measure users' satisfaction. The average SUS scores for the three platforms were 75.3, 73.4 and 64.5 respectively. The Game-based platform was found to be the best in terms of user performance and overall learning experience.

The SUS questionnaire was also used to assess the usability of the first version of the Topolor system (Shi, Awan, & Cristea, 2013). It is a Social Personalized Adaptive E-Learning Environment (SPAEE) and constitutes an attempt to combine social e-learning with adaptive e-learning. The SUS score was 75.75 and Cronbach's alpha was found to be 0.85.

Marco, Penichet, and Gallud (2013) proposed a way of remote collaboration in real time via the platform Moodle. This method is based on the use of the tool Drag & Share, a collaborative tool which enables sharing and synchronization of files in real time. To evaluate the Drag & Share tool, the researchers asked a group of users to perform the same collaborative work in Moodle with and without the use of the tool. Subsequently, they compared their efficiency and satisfaction across the two conditions. Efficiency was operationalized using the time taken for completion of work and user satisfaction was evaluated by the SUS questionnaire. SUS scores for the Moodle system with and without the tool Drag & Share were 89.5 and 46.75, respectively. These SUS scores confirmed the perceived usability of the tool proposed by the researchers.

A recent study (Luo, Liu, Kuo, & Yuan, 2014) proposed a simulation-based learning system for international trade and used the SUS questionnaire to assess its perceived usability. The researchers studied the usability of the system and the opinions of the students that used it during two semesters. They collected 49 and 24 valid completed surveys for each semester respectively. After the first semester, the SUS score was 58.1, which showed that the system needed improvement. The researchers made some modifications to the system based on the collected user data and found that after the second semester the SUS score increased to 65.93. The usability of the system was improved, but there was still room for further improvement. In addition, at the end of the second semester nine teachers were interviewed and they were also asked to complete the SUS questionnaire. The SUS score was 74.45, indicating that the part of the system which is addressed to the teachers was sufficiently usable. This finding also shows that students' and teachers' perceptions of LMS usage may vary (Emelyanova & Voronina, 2014).

Simoes and de Moraes (2012) used SUS to evaluate the usability of the virtual learning environment adopted by the Distance Education Center of the Federal Institute of Espírito Santo – Brazil, which has the Moodle platform as a basis. They examined the usability of the environment from three different perspectives using three different methods of evaluation. Initially, they used the SUS questionnaire to assess users' satisfaction, then they evaluated the design of the interface with the method of heuristic evaluation, and finally they evaluated the usability of the system by applying the method of cooperative evaluation. The sample size was 59 users: students who attended the first semester's course "Information Technology" and people who were already using the environment for four months. All three methods unveiled that the virtual learning environment had serious usability problems. The authors characterized the SUS questionnaire as an effective tool for evaluating usability and users' satisfaction. However, they did not report the obtained SUS scores.

Venturi and Bessis (2006) used SUS to evaluate the perceived usability of DELTA, a distributed learning resources repository. The sample was 14 users who performed six tasks before completing the SUS questionnaire. The authors state that usability tests were extremely effective in discovering usability issues. However, no SUS scores are reported.

In a case study (Granic & Cukusic, 2011), SUS was used among other usability evaluation techniques to evaluate UNITE, an e-learning platform that supports education in 14 European

secondary schools with a total of 512 students and 46 teachers. The survey was conducted in nine schools and 47 students and 23 teachers participated. The average SUS score was 59.36 for students and 53.15 for teachers. They also found that there was a significant negative correlation (r=-0.467, p=0.001) between the SUS score and students' age, as it was also found in Bangor et al. (2008).

Method

Research Design

All in all, 11 studies were carried out. In all studies, university students were asked to evaluate the perceived usability of their course's Moodle-based LMS (Fig. 1) or eClass-based LMS (Fig. 2) by completing SUS. Moodle was selected because it constitutes a widely acceptable and globally popular LMS solution. Eclass, an open platform based on Claroline, is the LMS solution used in the University of Patras (and in most Greek higher education institutes) and it was selected because: a) we had access to a wide variety of courses, and b) we wanted to have benchmark data for its perceived usability. In all studies, participation was voluntary.

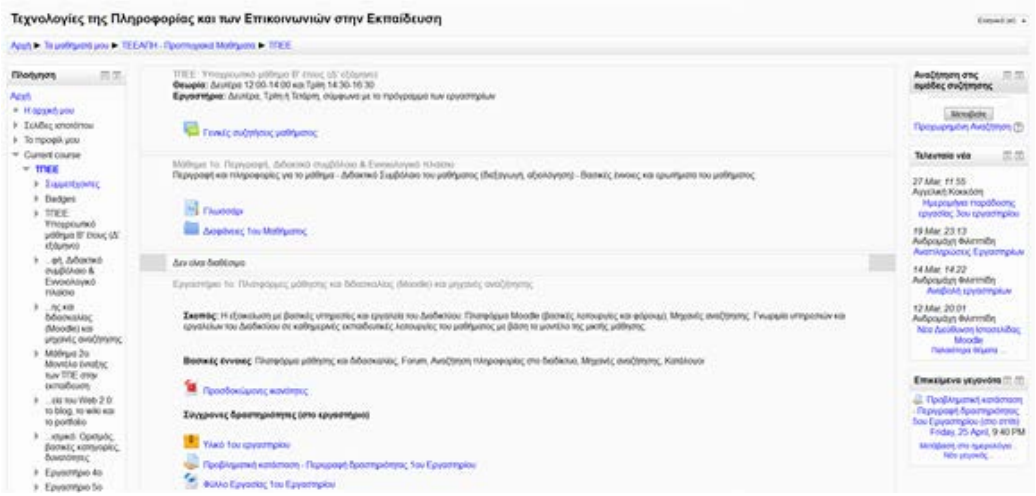

Figure 1. The Moodle-based system for the course "Technologies of Information and Communication in Education" (tenth study).

Figure 2. The eClass-based system for the course "Digital Telecommunications" (second study).

The provided questionnaire comprised SUS items as well as general demographics questions and questions related to the participants' internet self-efficacy (ISE) and internet user attitude (IUA). Already validated constructs were used to measure ISE and IUA (Papastergiou, 2010). Furthermore, students were asked to state whether they had prior experience with the LMS system used in their course (yes/no), and also rate on a 1 (not at all frequently) to 5 (very frequently) scale the usage frequency of their course's LMS. Moreover, an adjective rating was used to collect students' ratings of the LMS perceived usability according to a 7-point adjective rating scale (Bangor et al., 2009).

Participants were presented with either the Greek (Katsanos, Tselios, & Xenos, 2012) or the English (original) version of the SUS items based on their self-rated knowledge of English. We provided this option in an attempt to minimize potential threats to the validity and reliability of SUS data obtained from non-native English speakers (Finstad, 2006). In our previous research (Katsanos et al., 2012), three studies involving 280 university students in both blended and distance learning education were conducted. Analysis of the results demonstrated the validity and reliability of the Greek version of SUS in the context of Moodle evaluation. The questionnaire was pilot-tested in a preliminary study involving 63 students.

Participants and Procedures

Overall, the 11 studies reported in this paper involved 769 university students of eight University departments, 487 female and 282 male, aged 18 to 52 (mean=21.78, sd=4.31). They reported using the Internet for 8.02 years (sd=2.94) on average. A total of 436 students completed the Greek version of the SUS questionnaire, whereas 333 students completed the English version of SUS. In addition, 355 students had prior experience with the system used at their course. On a 1–5 scale, their mean self-reported LMS usage frequency was 3.81 (sd=0.99), their mean internet self-efficacy was 4.10 (sd=0.61) and their mean internet user attitude was 4.27 (sd=0.52). The

data collection process lasted thirty-four days, from 22 May to 9 June 2013 and from 30 January to 13 February 2014, at the end of each semester and took place after the last lecture of the course. The characteristics of the sample per study are elaborated in Table 1.

Table 1

Overview of Dataset

Study	N	Subject	LMS	Gender (female, male)	Age (mean ±sd)	GR, EN version of SUS	Mean LMS usage frequency (sd)	Mean Internet self-efficacy (sd)	Mean Internet attitude (sd)	Prior experience with LMS (yes, no)	Mean Internet usage in years (sd)
1	65	Electrical Machines I	eClass	19 female, 46 male	20 to 30 (21.85 ±2.60)	31 GR, 34 EN	3.58 (1.07)	4.42 (0.56)	4.45 (0.45)	57 yes, 8 no	9.78 (2.99)
2	56	Digital Telecommunications	eClass	17 female, 39 male	19 to 27 (23.27 ±1.69)	19 GR, 37 EN	3.18 (0.74)	4.42 (0.57)	4.34 (0.53)	48 yes, 8 no	9.77 (3.00)
3	54	Zoology I	eClass	37 female, 17 male	18 to 25 (18.65 ±1.18)	22 GR, 32 EN	4.24 (0.95)	3.88 (0.68)	4.01 (0.51)	18 yes, 36 no	7.28 (1.97)
4	87	Differential Equations	eClass	36 female, 51 male	19 to 27 (21.43 ±1.72)	40 GR, 47 EN	3.46 (0.97)	4.19 (0.54)	4.19 (0.53)	59 yes, 28 no	8.11 (2.44)
5	63	Pharmaceutical Immunology	eClass	31 female, 32 male	19 to 29 (21.14 ±1.93)	30 GR, 33 EN	4.02 (0.73)	3.96 (0.62)	4.23 (0.42)	43 yes, 20 no	8.38 (2.84)
6	52	Mathematics for Chemists	eClass	39 female, 13 male	18 to 22 (18.88 ±1.06)	27 GR, 25 EN	3.42 (0.98)	3.92 (0.58)	4.19 (0.43)	9 yes, 43 no	6.48 (1.85)
7	67	Mathematics I	eClass	58 female, 9 male	18 to 32 (18.99 ±2.14)	32 GR, 35 EN	3.51 (0.84)	3.97 (0.59)	4.17 (0.47)	21 yes, 46 no	7.19 (2.37)
8	42	Electronic Elements of Power	eClass	7 female, 35 male	21 to 26 (22.62 ±1.43)	15 GR, 27 EN	3.90 (0.82)	4.28 (0.59)	4.01 (0.73)	29 yes, 13 no	7.83 (2.31)

9	52	Decision Theory	eClass	28 female, 24 male	22 to 45 (24.29 ±s3.31)	23 GR, 29 EN	3.54 (0.92)	4.31 (0.56)	4.25 (0.50)	41 yes, 11 no	8.88 (2.09)
10	178	ICT in Education	Moodle	174 female, 4 male	19 to 48 (21.04 ±4.28)	all (178) GR	4.37 (0.86)	3.95 (0.53)	4.34 (0.50)	5 yes, 173 no	6.56 (2.57)
11	53	Methodology of Educational Research	Moodle	41 female, 12 male	22 to 52 (30.38 ±7.43)	19 GR, 34 EN	3.68 (1.12)	4.11 (0.66)	4.60 (0.44)	25 yes, 28 no	10.89 (3.80)
All	769	11 subjects, 8 Univ. Depart.	9 eClass, 2 Moodle	487 female, 282 male	18 to 52 (21.78 ±4.31)	436 GR, 333 EN	3.81 (0.99)	4.10 (0.61)	4.27 (0.52)	355 yes, 414 no	8.02 (2.94)

Research Materials

The Google Forms service was used to create and distribute the questionnaire for each study. The collected data were organized and preprocessed using Microsoft Excel 2010 and were analyzed using IBM SPSS Statistics v19.0. The materials provided to the students through the eClass and the Moodle system were organized according to each subject and were available to the students until the end of the semester (or the academic year).

Results

SUS within LMSs

SUS benchmark data for LMSs.

The SUS questionnaire is investigated within the context of assessing usability of LMSs. Table 2 presents a summary of the data collected in the eleven studies evaluating perceived usability of the eClass and the Moodle LMSs. The number of students who completed each SUS version, the mean and standard deviation of their SUS scores and perceived usability ratings are provided per study and across all studies. All SUS scores for the 11 studies are beyond the value of 72/100, therefore according to the criteria of Bangor et al. (2009) the usability of the evaluated LMSs is at a satisfactory level.

Table 2

Sample Size and Descriptive Statistics (SUS score, Usability Rating) in our Dataset

Study	LMS	Students (N)			SUS Score (0-100)		Usability Rating (1-7)	
		All	*Greek SUS*	*English SUS*	*Mean*	*SD*	*Mean*	*SD*
1	eClass	65	31	34	79.73	13.31	4.75	0.87
2	eClass	56	19	37	73.35	11.43	4.71	0.76
3	eClass	54	22	32	81.57	11.93	5.43	0.60
4	eClass	87	40	47	73.76	14.06	4.86	0.99
5	eClass	63	30	33	81.59	13.90	5.21	0.83
6	eClass	52	27	25	75.58	12.49	5.10	0.63
7	eClass	67	32	35	76.23	12.43	5.25	0.80
8	eClass	42	15	27	74.11	13.10	5.17	0.76
9	eClass	52	23	29	75.43	10.48	5.12	0.68
10	Moodle	178	178	0	75.84	12.14	5.23	0.76
11	Moodle	53	19	34	72.17	17.73	4.87	0.83
All	eClass	538	239	299	76.81	13.02	5.05	0.82
	Moodle	231	197	34	75.00	13.67	5.15	0.79
	All	769	436	333	76.27	13.24	5.08	0.81

SUS validity and reliability within LMSs' evaluation.

For the needs of the analyses, the negative statements of SUS (i.e. Q2, Q4, Q6, Q8, and Q10) were recoded so that positive responses are associated with a larger number, like the rest five positive statements. In all subsequent statistical analyses, we use the correlation coefficient *r* as an effect size, which is calculated according to the formulas reported in Field (2009).

The students' SUS scores were compared to their overall adjective ratings of perceived usability, following the same approach used in a study (Bangor et al., 2009) within a different context. A non-parametric measure of association was used (Spearman's coefficient) since the assumption of normality was violated by both dependents; $W(769)=0.979$, $p<0.001$ and $W(769)=0.849$, $p<0.001$ respectively. A significant correlation between the SUS score and the overall adjective rating was found; $rs=0.525$, $p<0.01$. This value is lower than the ones found by previous studies in examining the usability of other systems; $r=0.806$ (Bangor et al., 2008), $r=0.822$ (Bangor et al., 2009), and *r* between 0.498 and 0.787 (Kortum & Bangor, 2013). However, moderate correlations with absolute values as small as 0.30 to 0.40 are considered (Nunnally & Bernstein, 1994) large enough to justify the validity of psychometric instruments, such as questionnaires.

The values of the correlation between the overall adjective rating and the SUS score of the Greek and English language version of the questionnaire were rs=0.491, p<0.01 and rs=0.577, p<0.01, respectively. Furthermore, both language versions of the SUS had good internal consistency; Cronbach's alpha=0.820 for the English version (N=333 surveys) and Cronbach's alpha=0.808 for the Greek version (N=436 surveys).

In addition, it was examined whether the SUS is reliable within LMSs even with a small-sized sample. A similar process to that of Tullis and Stetson (2004), who studied usability of websites, was conducted. The context of websites can be considered a similar framework to e-learning platforms, and can provide a basis for comparison. First, twenty sub-samples were randomly chosen for a sample size of 6, 8, 10, 12, and 14 participants. Next, for each sample size, the percentage of 20 t-tests that yielded the same conclusion as the analysis of the full dataset (i.e., the mean SUS score of the sub-sample did not significantly differ from the overall mean SUS score) was calculated. Tullis and Stetson (2004) report that a sample of 12 to 14 users provides the correct finding 100% of the time. In the present study, it was found that the SUS score of a sample of 6 to 14 users does not differ from the total score at least 90% of the time.

Investigation of SUS Attributes within LMSs

Additional analyses were conducted to investigate the following SUS attributes in the context of LMSs' evaluation: a) gender, b) age, c) prior experience with the LMS, d) Internet self-efficacy, e) attitude towards the Internet, and f) usage frequency of the LMS. The first three attributes have been previously investigated in the context of other products' usability evaluation (Bangor et al., 2008; Granic & Cukusic, 2011; Kortum & Bangor, 2013; Sauro, 2011; Tullis & Stetson, 2004) and we were interested to explore whether the reported findings are confirmed within LMSs' evaluation. To the best of our knowledge, the latter three SUS attributes have not been previously explored in any evaluation context.

Students' gender and SUS score.

A two-tailed Mann-Whitney U test investigated the effect of students' gender on their SUS score. A non-parametric test was selected because the SUS score distributions of both men and women deviated significantly from the normal distribution; W(282)=0.976, p<0.001 and W(487)=0.980, p<0.001 respectively. Results showed no significant difference in SUS score between women (mean=76.30, sd=13.16) and men (mean=76.21, sd=13.40); U=68549.000, z=0.040, p=0.968. Comparing this result with the findings of previous research in a different context, it is consistent with the research of Bangor et al. (2008).

Students' age and SUS score.

A small, non-significant negative correlation was found between the SUS score and age of students who evaluate the usability of LMSs; rs=-0.061, p=0.09. Therefore, age does not seem to be significantly associated with the perceived usability of a LMS, which is in conflict with previous studies (Bangor et al., 2008; Granic & Cukusic, 2011). However, it is important to mention that

the range of students' age in our dataset may not be varying enough; our participants had a mean age of 21.78 years and the 95% confidence interval for the mean was from 21.47 to 22.08 years old. Enriching our dataset with SUS scores from older adult students is required before providing a definitive answer on the association between students' age and SUS score.

Students' prior experience with the LMS and SUS score.

A two-tailed Man-Whitney U test investigated the effect of students' prior experience with the LMS on their SUS score. The assumption of normality was violated for both conditions.Thus a non-parametric test was applied; $W(414)=0.979$, $p<0.001$ and $W(355)=0.974$, $p<0.001$. A significant difference was observed ($U=66467.000$, $z=2.289$, $p=0.022$, $r=0.08$) in the evaluation of perceived usability of LMSs between users with prior experience with the system (mean=77.36, sd=13.37) and first-time users (mean=75.33, sd=13.06). This result is consistent with findings of previous studies in different evaluation contexts (Kortum & Bangor, 2013; Sauro, 2011).

However, in the context of LMSs, students with prior experience score the perceived usability of the system averaged 2.12 percentage points higher (2.12%) than students with no prior experience, while Sauro (2011), who examined usability in websites, found that users with prior experience rate usability averaged 11% higher than first-time users.

Students' Internet self-efficacy and SUS score.

A significant positive correlation was observed between the SUS score and Internet self-efficacy (Papastergiou, 2010) of students who evaluate the usability of LMSs; $rs=0.326$, $p<0.01$. Therefore, the more efficient on the Internet students feel, the higher perceived usability ratings for the used LMS they provide.

In addition, we recoded our dataset to create two between-subject groups based on students' Internet self-efficacy: a) low Internet self-efficacy (N=376), which included students with a value below the mean value of all students, and b) high Internet self-efficacy (N=393), in which the rest students were assigned. A two-tailed Man-Whitney U test investigated the effect of students' Internet self-efficacy on their SUS score. A non-parametric test was applied since the assumption of normality was violated for both conditions; $W(376)=0.981$, $p<0.001$ and $W(393)=0.966$, $p<0.001$. A significant difference was observed ($U=47558.000$, $z=8.565$, $p<0.0001$, $r=0.31$) in the evaluation of perceived usability of LMSs between students with low Internet self-efficacy (mean=72.23, sd=12.73) and students with high Internet self-efficacy (mean=80.13, sd=12.56).

Students' attitude towards the Internet and SUS score.

Results showed that the SUS score and students' attitude towards the Internet as a learning tool (Papastergiou, 2010) were significantly correlated ($rs=0.223$, $p<0.01$). Therefore, students who have a more positive attitude towards the Internet for learning purposes rate higher the usability of the LMS by completing the SUS questionnaire.

Furthermore, we recoded our dataset to create two between-subject groups based on students' attitude towards the Internet: a) negative attitude (N=398), which included students with a value below the mean value of all students, and b) positive attitude (N=371), in which the rest students were assigned. A two-tailed Man-Whitney U test investigated the effect of students' attitude towards the Internet as a learning tool on their SUS score. The assumption of normality was violated for both conditions and therefore a non-parametric test was applied; $W(398)=0.986$, $p<0.001$ and $W(371)=0.961$, $p<0.001$. Results showed a significant difference ($U=57391.000$, $z=5.350$, $p<0.0001$, $r=0.19$) in the evaluation of perceived usability of LMSs between students with negative attitude towards the Internet (mean=74.04, sd=12.28) and students with positive attitude towards the Internet (mean=78.66, sd=13.81).

Students' usage frequency of the LMS and SUS score.

A significant positive correlation was observed between the SUS score and usage frequency of the LMS; $rs=0.267$, $p<0.01$. Thus, the more often students use the LMS that supports their course, the higher SUS score for the LMS they provide.

In addition, we recoded our dataset to create two between-subject groups based on students' usage frequency of the LMS: a) low usage frequency (N=255), which included students with a value below the mean value of all students, and b) high usage frequency (N=514), in which the rest students were assigned. A two-tailed Man-Whitney U test investigated the effect of students' usage frequency of the LMS on their SUS score. A non-parametric test was applied, since the assumption of normality was violated for both conditions; $W(255)=0.986$, $p<0.05$ and $W(514)=0.971$, $p<0.001$. A significant difference was observed ($U=48700.500$, $z=5.815$, $p<0.0001$, $r=0.21$) in the LMS's SUS score between students that used the LMS less frequently (mean=72.40, sd=12.85) and students that used the LMS more frequently (mean=78.19, sd=13.02).

Investigation of the Greek SUS Scale

Validity analysis.

Validity refers to the extent to which an instrument, such as a questionnaire, measures what it is intended to measure (Nunnally & Bernstein, 1994). Similarly to the English version of SUS, the Greek one is intended to measure perceived usability of software. Its validity was investigated following the same two approaches reported in our previous research (Katsanos et al., 2012).

First, the students' SUS scores were compared to their overall adjective ratings of perceived usability. A significant correlation between the SUS score provided by students who completed the Greek version of the questionnaire and the overall adjective rating was found; $rs=0.491$, $p<0.01$. This correlation value is very close to the one found in our previous research; $rs=0.474$, $p<0.01$.

Table 3

Validity Analysis of the Greek SUS Scale

Study	Students (N)			Mean Greek SUS score	Mean English SUS score	Comparison	Sig. (2-tailed)
	All	*Greek SUS*	*English SUS*				
1	65	31	34	80.81	78.75	t(63)=0.619	p=0.583
2	56	19	37	69.34	75.41	t(54)= -1.924	p=0.060
3	54	22	32	79.32	83.13	t(52)= -1.156	p=0.253
4	87	40	47	72.37	74.95	t(85)= -0.849	p=0.398
5	63	30	33	81.67	81.52	u=470.000	p=0.730
6	52	27	25	73.52	77.80	t(50)= -1.241	p=0.220
7	67	32	35	75.39	77.00	t(65)= -0.527	p=0.600
8	42	15	27	72.00	75.28	t(40)= -0.773	p=0.444
9	52	23	29	76.30	74.74	t(50)=0.531	p=0.598
10	178	178	0	75.84	-	-	-
11	53	19	34	73.95	71.18	t(51)=0.542	p=0.590
All	769	436	333	75.80	76.88	u=66886.000	p=0.314

Second, differences between the Greek and the English version of the SUS questionnaire were investigated (Table 3). An independent samples non-parametric test (Mann-Whitney U Test) was conducted to compare the means of the SUS scores obtained by the Greek and the English version of the questionnaire, since both dependents were significantly non normal; $W(436)=0.981$, $p<0.05$ and $W(333)=0.972$, $p<0.05$ respectively. On average, students filling the Greek version of the questionnaire provided slightly lower SUS scores (mean=75.80, sd=13.17) than those that completed the English version of the SUS questionnaire (mean=76.88, sd=13.31). However, this difference was not significant; U=69052.500, z=-1.162, p=0.245. All in all, the Greek version of the SUS questionnaire was found to be a valid instrument for measuring perceived usability in the context of LMSs usability evaluation both on a per-study and cross-study analysis.

Reliability analysis.

Reliability refers to the extent to which a questionnaire yields the same results under consistent conditions (Nunnally & Bernstein, 1994). It is most commonly measured using Cronbach's alpha,

which is a measure of internal consistency. For the dataset of 436 completed surveys, the 10-item Greek SUS questionnaire had a good internal consistency; Cronbach's alpha=0.808. In our previous research (Katsanos et al., 2012) the value of Cronbach's alpha for the Greek version of SUS was slightly lower; α=0.777.

Factor analysis.

Given that a study (Lewis & Sauro, 2009) has found that the English SUS scale consists of two reliable subscales that measure Learnability (items Q4 and Q10) and Usability (the rest items), additional analyses were conducted for the Greek SUS. First, reliability analysis indicated that the aforementioned Usability subscale had good internal consistency (Cronbach's alpha=0.810, N=436). However, the aforementioned Learnability subscale did not have sufficient reliability (Cronbach's alpha=0.508<0.70, N=436). In addition, a principal component analysis (PCA) was conducted on the 436 completed Greek SUS surveys with orthogonal rotation (varimax) and two-, three- and four-factors solutions. The results showed that the Greek SUS questionnaire does not include any reliable subscales. Thus, all statements form a single reliable scale that measures perceived usability, a finding consistent with our previous research (Katsanos et al., 2012).

Conclusions and Discussion

In this paper, an empirical evaluation of the SUS questionnaire in the context of LMSs' perceived usability evaluation is presented. Our research is motivated by the increasingly important need to have reliable and robust psychometric instruments for evaluating the perceived usability of LMSs. To this end, we present findings related to various SUS attributes in the context of LMS usability evaluation, using a dataset produced by 769 university students. Students were asked to evaluate the perceived usability of their course's Moodle-based or eClass-based LMS by completing SUS and providing a 7-point scale adjective rating. They were presented with either a Greek or an English version of the questionnaire based on their self-rated knowledge of English.

Analysis of the results showed that the SUS questionnaire is a valid tool for the assessment of LMSs' usability. The usability of the evaluated LMSs seems to be at a satisfactory level (mean SUS score equals 76.27/100, range from 72.17 to 81.59). Also, a sufficiently strong correlation (r_s=0.525, $p<0.01$) was observed between the score of the LMS usability as derived from the questionnaire SUS, and a concurrent 7-point scale which measures overall perceived usability (overall adjective rating) (Bangor et al., 2009). In addition, both the English and the Greek language versions of the SUS had a good internal consistency; Cronbach's alpha=0.820 for the English version (N=333 surveys) and Cronbach's alpha=0.808 for the Greek version (N=436 surveys). Furthermore, it was found that SUS ratings for LMSs can provide a reliable evaluation conclusion even with a sample size of 6 to 14 participants.

Moreover, it was found that students' gender does not significantly affect their SUS score. This finding is in alignment with previous research evaluating the perceived usability of other products (Bangor et al., 2008). In the context of LMSs' perceived usability evaluation, Ituma (2011) also found no effect of students' gender on usability ratings of the Blackboard WebCT. Moreover, students' age was not found to be significantly associated with SUS score. The latter contradicts previous research (Bangor et al., 2008; Granic & Cukusic, 2011) in the usability evaluation of other products that found a significant negative correlation between SUS and users' age. In addition, a significant effect of students' prior experience with the LMS on SUS score was found. Previous research in website usability (Sauro, 2011) reports a similar finding, but with a larger effect size compared to ours. Furthermore, students' self-efficacy on the Internet (ISE) (Papastergiou, 2010), attitude towards the Internet as a learning tool (IUA) (Papastergiou, 2010) and usage frequency of the LMS were found to significantly affect their SUS scores.

Finally, the Greek version of the questionnaire SUS (Katsanos et al., 2012) was confirmed as suitable for evaluating the usability of LMSs by students who speak Greek. No significant difference between the Greek (mean=75.80, sd=13.17) and the English (mean=76.88, sd=13.31) version of SUS was observed. In addition, reliability analysis demonstrated the internal consistency (Cronbach's alpha=0.808) of the SUS in Greek. It was also found that the Greek version of SUS forms a single scale without any subscales, contrary to the English version of SUS that was found (Lewis & Sauro, 2009) to form two reliable subscales that measure Learnability and Usability.

Regarding the contribution of this study, it is argued that the collected data can help the designers of LMSs since they can be used as benchmarks for the SUS score in the context of LMSs' perceived usability evaluation. This can provide significant feedback for their work, indicating the extent of the possible need for improvement. In addition, the SUS score might provide a common code of communication between manufacturers and their customers, who may not have the appropriate expertise in software usability. The SUS score provides good insight into the perceived usability, an important factor for e-learning adoption (Tselios, Daskalakis & Papadopoulou, 2011) and can be understood by all the LMSs' stakeholders. Furthermore, the findings of this study can help students, professors and educational organizations, who will be able to compare different systems in order to choose the right one and reject a non-usable system for the course they wish to follow or create.

The research presented in this paper has limitations. The sample includes students only from a single University with specific characteristics. Therefore, further investigation in different educational institutions and levels of education, and with more LMSs is required. The present study explores only students' views. In future work additional user groups could be studied, such as educators who may have quite diverging perceptions compared to students (Emelyanova & Voronina, 2014). Future work, also, includes investigating relationships (if any) between SUS score and a) the attitudes of participants towards the professor of the course who uses the LMS, and b) the education delivery method used, such as blended learning or distance education.

Moreover, the influence of students' behavior while seeking for information in a LMS on their SUS rating will be investigated, using eye-tracking equipment and the theoretical foundations of information foraging theory (Katsanos, Tselios, & Avouris, 2010; Tselios Katsanos, Kahrimanis, & Avouris, 2007). Finally, the reliability of the SUS questionnaire compared within LMSs with other questionnaires measuring usability, such as SUMI, CSUQ, and QUIS, requires further examination.

References

Ayad, K., & Rigas, D. (2010). Comparing virtual classroom, game-based learning and storytelling teachings in e-learning. *International Journal of Education and Information Technologies*, *4*(1), 15–23.

Bangor, A., Kortum, P., & Miller, J. (2008). An empirical evaluation of the sSystem uUsability sScale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594. doi:10.1080/10447310802205776

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS Scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Brooke, J. (1996). SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability evaluation in industry*. London: Taylor and Francis.

Emelyanova, N., & Voronina, E. (2014). Introducing a learning management system at a Russian university: Students' and teachers' perceptions. *The International Review of Research in Open and Distance Learning*, *15*(1).

Field, A. P. (2009). *Discovering statistics using SPSS*. Los Angeles, [Calif.]: London: SAGE.

Finstad, K. (2006). The System Usability Scale and non-native english speakers. *Journal of Usability Studies*, *1*(4), 185–188.

Granic, A., & Cukusic, M. (2011). Usability testing and expert inspections complemented by educational evaluation: A case study of an e-Learning platform. *Educational Technology & Society*, *14*(2), 107–123.

Ituma, A. (2011). An evaluation of students' perceptions and engagement with e-learning components in a campus based university. *Active Learning in Higher Education*, *12*(1), 57–68. doi:10.1177/1469787410387722

Katsanos, C., Tselios, N., & Xenos, M. (2012). Perceived usability evaluation of learning management systems: a first step towards standardization of the System Usability Scale in Greek. In *2012 16th Panhellenic Conference on Informatics (PCI)* (pp. 302–307). doi:10.1109/PCi.2012.38

Katsanos, C., Tselios, N., & Avouris, N. (2010). Evaluating web site navigability: Validation of a tool-based approach through two eye-tracking studies. *New Review of Hypermedia and Multimedia*, vol. 16(1-2), pp. 195-214.

Kortum, P., & Bangor, A. (2013). Usability ratings for everyday products measured with the system usability scale. *International Journal of Human-Computer Interaction*, *29*(2), 67–76. doi:10.1080/10447318.2012.681221

Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGCHI Bull.*, *23*(1), 78–81. doi:10.1145/122672.122692

Lewis, J. R. (1992). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proc. of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 36, pp. 1259–1263). Santa Monica, CA: HFES.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78. doi:10.1080/10447319509526110

Lewis, J. R., & Sauro, J. (2009). The factor structure of the sSystem uUsability sScale. In *Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International 2009* (pp. 94–103). Berlin, Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-02806-9_12

Luo, G.-H., Liu, E. Z.-F., Kuo, H.-W., & Yuan, S.-M. (2014). Design and implementation of a simulation-based learning system for international trade. *The International Review of Research in Open and Distance Learning*, *15*(1). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/1666

Marco, F. A., Penichet, V. M. R., & Gallud, J. A. (2013). Collaborative e-Learning through Drag & Share in synchronous shared workspaces. *J. UCS*, *19*(7), 894–911.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill Humanities/Social Sciences/Languages.

Papastergiou, M. (2010). Enhancing physical education and sport science students' self-efficacy and attitudes regarding information and communication technologies through a computer literacy course. *Computers & Education*, *54*(1), 298–308. doi:10.1016/j.compedu.2009.08.015

Renaut, C., Batier, C., Flory, L., & Heyde, M. (2006). Improving web site usability for a better e-learning experience. *Current Developments in Technology-Assisted Education*, 891–895.

Sauro, J. (2011). Does prior experience affect perceptions of usability? Retrieved from http://www.measuringusability.com/blog/prior-exposure.php

Shi, L., Awan, M. S. K., & Cristea, A. I. (2013). Evaluating system functionality in social personalized adaptive e-Learning systems. In D. Hernández-Leo, T. Ley, R. Klamma, & A. Harrer (Eds.), *Scaling up learning for sustained impact* (pp. 633–634). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-40814-4_87

Simões, A. P., & de Moraes, A. (2012). The ergonomic evaluation of a virtual learning environment usability. *Work: A Journal of Prevention, Assessment and Rehabilitation*, *41*, 1140–1144. doi:10.3233/WOR-2012-0293-1140

Torkzadeh, G., & van Dyke, T. P. (2001). Development and validation of an Internet self-efficacy scale. *Behaviour & Information Technology*, *20*(4), 275–280. doi:10.1080/01449290110050293

Tselios, N., Daskalakis, S., & Papadopoulou, M. (2011). Assessing the acceptance of a blended learning university course. *Educational Technology & Society*, *14*(2), 224–235.

Tselios, N., Katsanos, C., Kahrimanis, G., & Avouris, N. (2007). *Design and evaluation of web-based learning environments using information foraging models*. In Pahl, C. (ed.), Architecture Solutions for e-learning systems, pp. 320-339, Hershey, PA, USA: Information Science Reference.

Tullis, T., & Albert, W. (2008). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Morgan Kaufmann.

Tullis, T., & Stetson, J. (2004). A comparison of questionnaires for assessing website usability. In *Usability Professionals Association (UPA) 2004 Conference* (pp. 7–11).

Venturi, G., & Bessis, N. (2006). User-centred evaluation of an e-Learning repository. In *Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles* (pp. 203–211). New York, NY, USA: ACM. doi:10.1145/1182475.1182497