Hayashi, Y., et al. (Eds.) (2017). Workshop Proceedings of the 25th International Conference on Computers in Education. New Zealand: Asia-Pacific Society for Computers in Education

Pre-Testing the Chinese Version of the System Usability Scale (C-SUS)

Feng-Ru SHEU[a], Hui-Jung FU[b*], & Meilun SHIH[c]

[a]University Libraries, Kent State University, USA

[b]Physical Education Center, Southern Taiwan University of Science and Technology, Taiwan

[c]Center for Teaching & Learning Development, National Taiwan University, Taiwan

*hufu@stust.edu.tw

Abstract: Background: Given many advancement in technology, information & communication technology (ICT) in education for enhancing effectiveness of teaching and learning has become a widely applied and discussed area. Usability is central for the success of any instructional design product or learning materials, including any educational websites, learning management system, mobile devices, and wearable technology. The System Usability Scale (SUS) is one of the commonly used questionnaires for usability rating. Objectives: With the increasing interest in usability studies and user experience research, there is a need to officially translate it into Chinese and also to validate the translation. The aim of this paper is to describe the process of translating the original System Usability Scale (SUS) from English into Chinese (C-SUS), and to evaluate its reliability and validity in the college students. Methods: This study consisted of two phases. In phase one, the SUS was translated into Chinese by a group of translators and experts in education using Brislin's (1970, 1986) translation and back-translation method. Both semantic equivalence and content validity were assessed. In the second phase, the psychometric properties of the C-SUS were tested with two studies and with convenience samples of 125 (study 1) and 104 (study 2) college students recruited from a private university in southern Taiwan. Reliability was assessed by internal consistency and construct validity was tested using exploratory factor analysis. Data analyses was performed using SPSS 23.0 to assess reliability and validity. Results: The sematic equivalence and content validity index of the Chinese version of SUS were satisfactory. Results also indicated that the Chinese version had a high level of equivalence with the original English version and demonstrated a high internal consistency. Exploratory factor analysis revealed the presence of two factors supporting the conceptual dimension of the original instrument. Conclusion: The study provides initial psychometric properties of the Chinese version of the SUS and supports it as a reliable and valid instrument to measure usability for design products and services for Chinese speaking individuals.

Keywords: System Usability Scale (SUS), translation, validation, usability testing

1. Introduction

Information & communication technology (ICT) has been important and used in education for decades to promote effective teaching and learning. It is even more so with recent technology advancements, such as interactive educational websites, mobile applications (APP) in both Android and OS, learning management system (LMS), virtual reality (VR), and wearable technology, to name a few. Among all elements for design and development, usability is the most important element for instructional design products and services to be successful. There are several questionnaires available for professionals to assess the usability of given products or services with target users, such as After Scenario Questionnaire (ASQ), Computer System Usability Questionnaire (CSUQ), Software Usability Measurement Inventory (SUMI), Usefulness, Satisfaction and Ease of Use (USE), Website Analysis and Measurement Inventory (WAMMI), and System Usability Scale (SUS). In that list of tools, the System Usability Scale (SUS), first developed by Brooke in 1986 as a quick and easy-to-use scale, is the most commonly used questionnaire for rating usability (Brooke, 1996; Sauro, 2011). In

Tullis and Stetson's study (2004), SUS, the shortest survey in the study, was among those providing the most reliable results across sample sizes for user satisfaction on web assessment.

As mentioned, the SUS was initially developed by John Brooke for a quick measurement on usability. The standard SUS consists of ten items with 5 Likert scale from 1= strongly disagree to 5= strongly agree and odd-numbered items worded positively and even-numbered items worded negatively:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

There are many positive attributes that lead to the wide use of the SUS. The SUS is short, containing only 10 items and is easy to use, allowing professionals in the usability field to quickly and easily assess the usability rating of a given product from users' perspective. The SUS has been shown to have good reliability and validity (Bangor, Kortum, & Miller, 2008; Sauro, 2011). A key factor is that the scale is technology-agnostic so that it can be used for a wide range of products, such as websites, cell phones, software, applications, and TV programs etc. It also can be understood by a wide range of people. In other words, it can be used for assessing usability of educational materials and devices, including learning management system, educational websites, and any other information communication technology for teaching and learning. In addition, the scale is free of charge and open access, which makes it a good, cost-effective tool (Sauro, 2011).

The SUS has been unofficially translated into several languages, including Spanish, French, and Dutch (Brooke, 2013; Sauro, 2011) and has been used on projects in various development stages. With the increasing interest in usability studies and user experience research, there is a need to officially translate it into Chinese and to validate the translation. This study addresses those needs. This present study reports the process of translating the original SUS from English into Chinese and assessing its reliability and validity among Chinese speaking individuals.

2. Research Design

The purpose of this study is to formally evaluate the Chinese translation of System Usability Scale from English to Chinese. The common procedure of psychological scale/test adaptation usually consists of two phases: translation and validation. In the phase one, the SUS was translated into Chinese using Brislin's (1970, 1986) translation and back-translation method by a group of 3 translators and 3 experts in education. Both semantic equivalence and content validity were assessed. In the second phase, the psychometric properties of the C-SUS were tested with two studies. In study one, it was tested with a convenience sample of 125 college students on an educational website. In study two, the C-SUS was tested with 87 students with three different type of educational systems.

3. Phase One: Translation and Back-Translation

In the first phase, the SUS was translated into Chinese using Brislin's (1970, 1986) translation and back-translation method. The semantic equivalence and content validity were assessed. The translation process was conducted by applying Brislin's methods of translation and back- translation (1970) as well as translation guidelines by Guillemin, Bombardier, and Beaton (1993). Translators were fluent in both Chinese and English and were familiar with the cultures. The quality of translation was tested by considering semantic equivalence and cross-culture relevance of the scale. The translations were compared and analyzed by three experts in educational technology and education administration. Both the original English scale and the translations were compared by the researcher and the translators. Where there was a disagreement on translation, discussion took place until consensus was reached. The translated version was pre-tested by five people before being used in the usability studies in phase two. Below shows the Chinese translation of SUS:

1. 我覺得我會常使用這個系統。
2. 我覺得這個系統太過不必要的複雜。
3. 我覺得這個系統是容易操作的。
4. 我覺得我需要透過專人的協助才能操作這個系統。
5. 我覺得這個系統許有多不同功能，整合的很好。
6. 我覺得這個系統很多地方不一致令人困惑。
7. 我覺得大部分的人都能很快知道怎麼使用這個系統。
8. 我覺得這個系統使用起來有點麻煩。
9. 我非常有信心下次能自己順利操作這個系統。
10. 我在能操作這個系統前，要學很多東西。

4. Phase Two: Psychometric Testing

4.1. Study 1

According to Fang and Liu (2002), the sample size should be 5-10 times larger than the number of items in the instrument used and expanded by at least 10% to ensure a sufficient sample size. As a result, a reasonable/legitimate/effective sample size of ranging 60 to 110 was calculated, as the number of items of C-SUS is 10. In study 1, a convenience sample of 125 freshmen who speak Chinese from a private university in southern Taiwan was recruited through an announcement post on the learning management system at the school. The translated Chinese version of SUS was used. The test system was an educational website about volleyball. Students were instructed to complete 3 information-searching tasks using the website. After completion of the tasks, they were asked to fill out Chinese version of SUS.

4.2. Study 2

In Study 2, a convenience sample of 104 Chinese-speaking freshmen was recruited to participate using the same recruitment method as study 1. Three selected systems were introduced to the participants separately in random order and two weeks apart in order to prevent order bias. The three systems were mobile application NIKE+, a website of a fitness association, and an educational website about basketball. All systems were in Chinese and they were all new to the participants. For each usability test, all participants were asked to perform the same tasks, including creating a profile, taking a screenshot of a particular screen, and searching for specific information to answer to the questions given by the researcher. At the end of each usability test, the students were asked to fill out Chinese version of SUS (C-SUS). A total of 81 students completed all tasks, including C-SUS survey.

5. Data Analysis

The SUS consists of 10 5-point Likert items ("1" representing "Strongly disagree" and "5" representing "Strongly agree"). Scoring of each item alternates between positive and negative. Overall SUS scores are scaled from 0 (lowest usability) to 100 (highest usability), and reflect a general measure of user-perceived usability. Brooke (1996) and Sauro (2011) are sources for detailed scoring. SUSCalc was used to obtain overall SUS scores. The SUS score equals adding each raw scores and multiplied by 2.5.

The reliability of the C-SUS was determined in terms of homogeneity - that is, Cronbach's alpha (α) coefficients - by examining the internal consistency of the questionnaire. Cronbach's alpha indicates how well the items are measuring the same dimension. Alpha values range from 0 to 1, with $\alpha > 0.80$ are considered "good reliability" and $\alpha > 0.90$ is considered "excellent reliability" (Kirakowski, 1994). A recommended interval for Cronbach's alpha value is .70 – .90 (Terwee et al., 2007).

Semantic equivalence is rated on a 4-point Likert scale ("not appropriate" to "most appropriate"). The translations were compared and analyzed by three experts in educational technology and education administration. Both the original English scale and the translations were compared by the researcher and the translators. The content validity of the C-SUS was established on a 4-point rating scale (1 = not relevant, 2 = somewhat relevant, 3 = quite relevant, and 4 = very relevant) content validity index (CVI). The CVI is the percentage of total items rated by experts as 3 or 4 and with a value of > .8 indicated good content validity.

The construct validity of the C-SUS was estimated by exploring its factor model with an exploratory factor analysis (EFA - Principal component analysis - Varimax with Kaiser normalization) to determine the factor loading of the items and their dimensions. The factor-loading criterion of the items was set to 0.40 in this study.

6. Results and Discussion

The internal consistency of C-SUS was illustrated by Cronbach's alpha coefficient of 0.93, indicating good reliability. The CVI was calculated to estimate the content validity at the item level (I-CVI) and scale level (S-CVI). The I-CVIs of each item were assessed by the four experts, and the values ranged from 0.75-1.00. The S-CVI of equivalence was 0.9. The CVI was 0.95, indicating the content validity of the items of the C-SUS. The study sample of 125 Taiwanese college students consisted of 69 (55.2%) males and 56 (44.8%) females. The C-SUS scores ranged from 40.81 - 48.03, with a mean score of 44.42 (SD = 20.38).

An EFA was conducted. A Kaiser-Meyer-Olkin (KMO) value was 0.925 and the Bartlett spherical test value was 925.297 ($p < .000$), which meant that the factor analysis was feasible. The scree plot suggested generating a two-factor model (Figure 1). Two common factors, where the Eigenvalues were > 1, were extracted after varimax orthogonal rotation, and 74.93% of the variance was explained by a two-factor solution. Each item had an acceptable factor loading on one of the two common factors and the communalities were from 0.656 - 0.815 (Table 1). The two factors were labeled as "usability" and "learnability" (Lewis & Sauro, 2009).

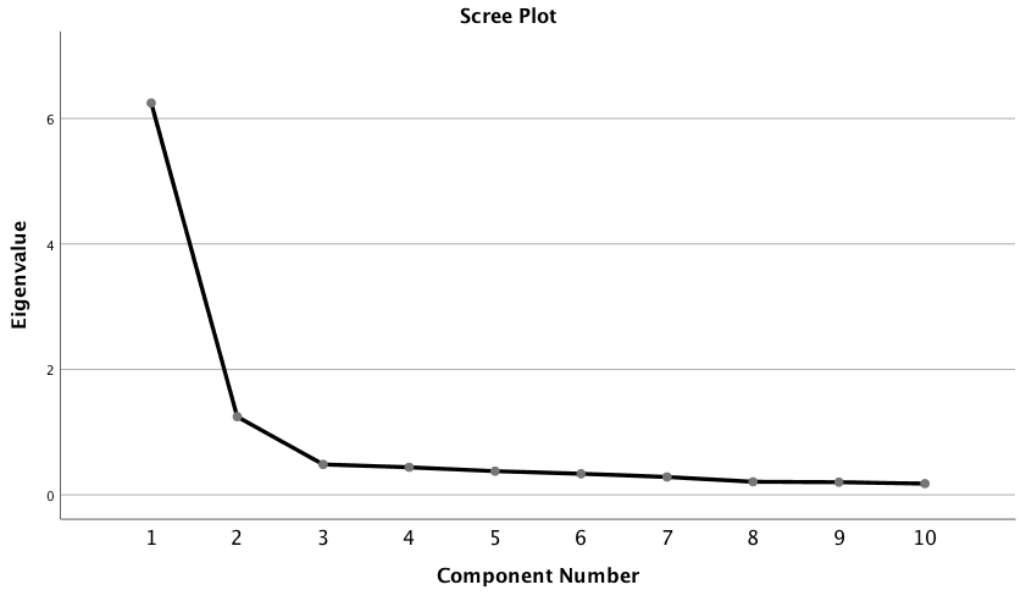

Figure 1. Scree plot illustrating the factor loading of the Chinese version of the SUS.

Cronbach's alpha (α) is a measure of the internal consistency of a questionnaire, which indicates how well the items are measuring the same dimension. As mentioned earlier, Alpha values range from 0 to 1, with $\alpha > 0.80$ considered "good reliability" and $\alpha > 0.90$ considered "excellent reliability" (Kirakowski, 1994). In this sample of 81 Taiwanese college students, the Chinese version of the SUS for testing mobile app NIKE +, Fitness website, and Basketball website had alpha coefficients of 0.90, 0.85, and 0.90, indicating good to excellent reliability. A comparison of Cronbach's alpha coefficients of the Chinese SUS scores in three usability testing is provided in Table 2.

The findings of both studies provided initial support that Chinese version of the SUS is a reliable tool for assessing usability rating with intended target users. Translation and back-translation method was applied and the versions of both Chinese and English were compared and assessed by the researcher, experts, and translators. Translation of "I feel…", "I think…", "I found…" were semantically the same or exchangeable in Chinese in the context of describing opinions and thinking. Translations between two languages were found to be satisfactory by researchers, translators, and the experts involved in the process.

In terms of reliability, Cronbach's alphas was 0.93 (Volleyball site) in study 1 and were 0.90 (NIKE+ mobile APP), 0.85 (Fitness Association), and 0.90 (Basketball site) in study 2, which indicated good internal consistency and acceptable reliability (DeVellis, 2003; Kline, 2005). According to Kirakowski (1994), the typical minimum reliability goal for questionnaires used in research and evaluation is 0.70. In a study conducted on the original SUS in English by Bangor et al. (2008), the coefficient alpha of the SUS of 2324 cases to be 0.91. The exploratory factor analysis yielded an interpretable two-factor solution, which accounted for 74.93% of the variance. The two-factor structure found in this study is consistent with the dimensional nature of the finding of Lewis and Sauro (2009); however, although it contrasts with the finding of Brook (1996).

Table 1: Summary of principal component analysis with varimax rotation.

Item	Factor 1 Usability	Factor 2 Learnability	Communalities
Q4 I think that I would need the support of a technical person to be able to use this system.	.879		.794
Q2 I found the system unnecessarily complex.	.832		.812
Q8 I found the system very cumbersome to use.	.834		.813
Q6 I thought there was too much inconsistency in this system	.817		.800
Q10 I needed to learn a lot of things before I could get going with this system	.776		.763
Q1 I think that I would like to use this system f requently		.853	.743
Q5 I found the various functions in this system were well integrated.		.802	.734
Q3 I thought the system was easy to use		.742	.702
Q7 I would imagine that most people would l earn to use this system very quickly		.737	.656
Q9 I felt very confident using the system		.708	.677
Eigenvalues % of Variance (total)	62.467	12.462	(74.929)

Table 2: Means, standard deviations, and Cronbach's alpha coefficients of C-SUS on NIKE+, Fitness website, and Basketball website.

Tested System	*M*	*SD*	*α*
NIKE+	60.7	18.0	0.90
Fitness	53.5	16.1	0.85
Basketball	58.2	15.1	0.90

7. Conclusion

This paper describes the process of the translation and validation of the System Usability Scale (SUS) from English into Chinese. To the best of our knowledge, the present study is the first effort to investigate and report the psychometric properties and the equivalences of the Chinese version SUS. To make the translation more suitable to Chinese language, a decision was made not to translate them literally to better reflect the intended purpose of scale in English. Words like "thought, felt, and found" were interchangeable in this case. Overall the Chinese version of the SUS is appropriate for use when conducting usability test with Chinese speakers. The translated version was well accepted

and understood by the participants. Therefore, the finding of this study may be valuable for providing usability professionals an easy-to-use tool for assessing usability for products or services in education, especially as these products or services involve information communication technology, such as educational websites, mobile APP, or wearable technology. It would also provide a tool for cross-culture research. Long-term development of this research should include a follow-up study with both English and Chinese speakers on the same system(s) that are free of culture bias.

Acknowledgements

We would like to thank Kristin Yeager for her statistical expertise and assistance, and all the participants, the experts, and the translators involved in this study.

References

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. Intl. Journal of Human–Computer Interaction, 24(6), 574-594.

Brislin, R. W. (1970). Back-translation for cross-cultural research. Journal of Cross-Culture Psychology, 1, 185-216.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), Field methods in cross-cultural research. (pp. 137-164). Thousand Oaks, CA: Sage Publications, Inc.

Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In P.W. Jordan, B. Thmoas, B. A. Weerdmeester, & I. L. McClelland (Eds.), Usability evaluation in industry (pp. 189-194). London: Taylor & Francis.

Brooke, J. (2013). SUS: A retrospective. Journal of Usability Studies, 8(2), pp. 29-40.

DeVellis, R. (2003). Scale development: Theory and applications. Thousand Oaks, CA: Sage Publications, Inc.

Fang, J. Q., & Lu, Y. (2002). Advanced medical statistics. Beijing, China: People's Medical Publishing House.

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

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. Journal of Clinical Epidemiology, 46(12), 1417-1432.

Kirakowski, J. (1994). The Use of Questionnaire Methods for Usability Assessment. In T. Bösser (Ed.), Measures and methods for quality of use. Retrieved from http://sumi.ucc.ie/sumipapp.html

Kline, T. (2005). Psychological testing: A Practical approach to design and evaluation. Thousand Oaks, CA: Sage Publications, Inc.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In Human Centered Design (pp. 94-103). Springer Berlin Heidelberg.

Sauro, J. (2011). A practical guide to the System Usability Scale (SUS): Background, Benchmarks & Best Practice. Denver, USA: A measuring usability LLC Publication.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology, 60, 34–42.

Tullis, T. S., & Stetson, J.N. (2004). A Comparison of questionnaires for assessing website usability. Proceedings of Usability of Professionals’ Association (pp. 1-12), Minneapolis, MN, June 7-11. Retrieved from http://home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf