Usability and User Experience, Vol. 39, 2022, 224–232
https://doi.org/10.54941/ahfe1001711

A Global Questionnaire? An International Comparison of the System Usability Scale in the Context of an Infotainment System

Alexandra Löw, Denise Sogemeier, Sarah Kulessa, Yannick Forster, Frederik Naujoks, and Andreas Keinath

BMW AG, Petuelring 130, 80809 Munich, Germany

ABSTRACT

The System Usability Scale (SUS) is a widely used questionnaire to assess the subjective usability of interactive products or services. Past research has already demonstrated psychometric properties of the SUS in different languages. However, there are no international psychometric studies that empirically proof that the SUS can be applied for the *same* product in *different* international markets. Therefore, the aim of this study was to investigate if the SUS provides comparable results. Participants from China, Germany and the USA were asked to perform different use cases using the infotainment system of a series production vehicle followed by an evaluation using SUS in their corresponding language. We assessed various psychometric quality measures to evaluate the SUS. Based on the results, the translations received validation support to a certain extent, but further research or adjustments are necessary to validate the SUS as a global questionnaire in the context of an infotainment system.

Keywords: Usability, Method development, Human machine interface, System usability scale, Intercultural comparison

INTRODUCTION

Due to technological advancement, the interaction between systems and users is playing an increasingly important role worldwide. Globalization emphasizes the cultural factor of such an interaction. In-Vehicle Information Systems (IVIS) are a standard technology in many road vehicles but new features can lead to distraction and safety issues (Svenson & Patten, 2005). Therefore, the usability design of IVIS is crucial. The evaluation of usability in turn is important to assess the extent an IVIS meets the characteristics of usability. The greater aim of those evaluations is to increase the usability by identifying areas of improvement in the interactions (Gray & Salzman, 1998). Results can be used to provide feedback on a product, to indicate the likely success of a product in the intended market or to compare two or more similar products (Butler, 1996; Rennie, 1981). Usability of a product or a system can be evaluated using questionnaires which are a cost-effective and time-efficient method to collect self-reported data from users. The quality of those questionnaires

can be classified by their psychometrics. Psychological questionnaires should generally include reliability, construct validity, and sensitivity considerations (Nunnally, 1978). The most widely used questionnaire to assess the usability of a system or product is the System Usability Scale (SUS) by Brooke (1986). The survey consists of ten questions, each to be answered on a five-point Likert scale from "strongly agree" to "strongly disagree". The scale is now being used in various surveys to assess the usability of user interfaces, such as software interfaces, websites or In-Vehicle interfaces (Bangor et al., 2008) and has been established as an industry standard (Baumgartner et al., 2019; Lewis & Sauro, 2017). Measuring usability using the SUS only takes mere minutes per system and participant which makes it a convenient method in the industry context (Orfanou et al., 2015). Since the SUS was developed, numerous studies have confirmed its excellent psychometric properties (Lewis, 2018). It can be used for small sample sizes while still yielding reliable test results (Tullis & Stetson, 2004). Another advantage of the SUS is that it is technology agnostic and can therefore be used to assess the usability of a wide range of products, such as phones or IVIS (Li et al., 2017) and automated driving (Forster et al., 2019). Originally, the SUS was developed having a one factor structure with the one factor of perceived usability. Multiple studies reanalyzed SUS indicating a two factors structure (Borsci et al., 2009), with items 4 and 10 aligning on a separate factor from the other items. Lewis and Sauro (2009) defined the two factors as *Learnability* (items 4 and 10) and *Usability* (all other items). An increasing number of studies are focusing on usability across cultures (Clemmensen, 2011; Clemmensen & Roese, 2010). Therefore, the SUS has already been translated into different languages. Vatrapu and Perez-Quinones (2006) claim that usability testing cannot provide accurate information of a local product when it is tested using techniques that do not consider cultural influences. Numerous cross-culture studies have been conducted on a wide range of products and cultures using the SUS in order to investigate usability. Those studies show significant differences in performance among users from different cultures (Gao et al., 2020). However, it is difficult to identify the reasons for these differences. Finstad (2006) argued that the original SUS may not be suitable in multicultural environments as non-native English speakers may interpret it differently. This assumption aligns with the results of Marzuki et al. (2018) showing that different cultures can interpret similar words or phrases in a different manner. Cultural and linguistic differences are often insufficiently considered in the literal translation of questionnaires (Hilton & Skrutkowski, 2002). This negligence can also be found in the usage of the SUS: In the Polish translation the word "cumbersome" in item 8 of the SUS is replaced by "inconvenient" (Borkowska & Jach, 2017); in the Arabic translation it is replaced by "strange" (AlGhannam et al., 2018). Wang et al. (2020) stressed that a certain concept does not have the same relevance in different cultures. Although there are multiple studies that examined the reliability and validity of the questionnaires' translations, a literature research did not reveal the extent to which an international comparison of the *same* product in *different* markets using the SUS is reliable, valid and sensitive. Therefore, the goal of this study is to investigate if the SUS leads to comparable results or if cultural background may influence the

results. The study aims to answer the research question if the SUS is indeed a global questionnaire.

METHOD

Participants

In total, $N = 102$ participants took part in the on-road study with $n = 36$ from China, $n = 36$ in the USA and $n = 30$ in Germany. In Germany, three females and 27 males participated in the study with 15 participants owning a BMW. The mean age was 54.0 years ($SD = 11.0$) ranging from 26 to 74 years. None of the participants experienced BMW's ID 7.0 previously. In the USA 24 male and 12 female participants took part in the study. Mean age in the USA was 39.5 years ($SD = 8.3$) with a range from 27 to 68 years. Sixteen participants owned a BMW. In total, 36 participants took part in the study in China with 28 being male. The mean age was 35.5 years ($SD = 7.1$) with a range from 25 to 54 years. The majority of the participants ($N = 18$) owned a BMW.

Human Machine Interface

The participants were able to interact with the infotainment system of the BMW X5 using the vehicles Human-Machine-Interface (HMI). The interaction with the system is possible using multiple interaction modalities. Users can operate with a central control element which consists of control buttons (e.g., menu, back, options) and a controller which is operated by pressing, shifting and rotating. Also, the users can use the touch screen or the voice control which can be activated through a button on the steering wheel.

Study Design and Procedure

The study was conducted as a between-subjects design with country being the independent variable and the score of the SUS being the dependent variable. The study was conducted in three different countries and was carried out in Shanghai, China in 2019, in Sherman Oaks, California, USA in 2019 and in Würzburg, Germany in 2018. In each country, a corresponding native speaking investigator conducted the study. After welcoming the participants and proving a consent form, a preliminary survey was conducted. The participants filled out a demographic questionnaire and answered additional questions about vehicle usage and experience with the interaction modalities. After an introduction into the vehicle operations of the BMW X5, participants were asked to perform different use cases in the vehicle. The investigator sat in the passenger seat and read the tasks out loud. Each participant completed a total of nine use cases with five use cases being conducted while driving. After finishing each use case in all modalities, the next use case was carried out. The tasks were always started from the home screen of the IVIS. The driving use cases were conducted in a traffic restrictive area. The participants drove a given route with a maximum speed of 30 km/h. It was pointed out that the driving safety had the highest priority and that drivers should refuse to handle use cases if the use cases seem to be too distracting. After

completing the nine use cases, participants evaluated the infotainment system of the BMW X5 regarding usability using the SUS in their corresponding language.

Statistical Procedure and Data Analysis

Statistical tests were conducted using the Software IBM SPSS Statistics. Internal consistency was calculated as a measure of scale reliability. Construct validity was examined by conducting a factor analysis. The total sample of $N = 102$ was included in the analysis.

RESULTS

Descriptive Statistics

The total SUS score was calculated for each country. The average score across all three countries was $M = 74.95$ ($SD = 19.69$). In Germany the average SUS score was $M = 72.08$ ($SD = 18.37$), in the USA $M = 78.89$ ($SD = 21.22$) and in China $M = 73.89$ ($SD = 19.49$).

Sensitivity

A Kolmogorov-Smirnov test revealed that the assumption of a normal distribution was not fulfilled ($p = .002$). Because the ANOVA seems to be robust to violations of non-normality when sample sizes are equal, an ANOVA was applied. Homogeneity of variances asserted using Levene's test showed equal variances could be assumed ($p = .890$). The F-test showed no significant difference between the SUS scores, $F(2,99) = 1.07$, $p = .346$, $\eta_p^2 = .02$.

Item Analyses

As a measure of reliability, Cronbach's alpha was calculated to assess the internal consistency. The reliability analysis revealed high internal consistency in all three countries. In Germany, the SUS revealed an average Cronbach's alpha of $\alpha = 0.88$, in the USA of $\alpha = 0.92$ and in China of $\alpha = 0.87$. In multiple item analyses the correlations of each item with the total scale were calculated. The correlations between each item and the overall questionnaire score for each country can be found in Table 1. A correlation less than $r = 0.30$ would indicate that the item may not belong to the scale. This wasn't the case for any item in the presented study. The corrected item total correlations ranged between $0.30 < r < 0.86$. If Cronbach's alpha becomes much larger after a certain item has been removed, the regarding question may not fit the higher-level construct of usability. Cronbach's alpha when the item of interest is omitted can be found in Table 1. Deleting item 6 would increase Cronbach's alpha to $\alpha = 0.88$, which could lead to the consideration of removing item 6 whereas in Germany a removal of item 2 could be considered.

Table 1. Results of the item analyses in Germany, USA and China.

Item	Germany		USA		China	
	Corrected Item-Total Correlation	Cronbach's α if item deleted	Corrected Item-Total Correlation	Cronbach's α if item deleted	Corrected Item-Total Correlation	Cronbach's α if item deleted
1	.649	.869	.475	.922	.419	.870
2	.416	**.885**	.801	.905	.707	.849
3	.693	.867	.855	.902	.749	.845
4	.512	.881	.712	.910	.649	.852
5	.788	.861	.656	.913	.509	.858
6	.632	.870	.862	.903	.306	**.881**
7	.550	.876	.837	.902	.680	.853
8	.854	.854	.740	.908	.575	.861
9	.481	.880	.418	**.929**	.302	.867
10	.642	.870	.719	.910	.572	.856

Note. The bold values show the increased Cronbach's alpha after the respective item removal.

Factor Analyses

Construct validity was investigated by means of a factor analysis for each country individually. Because there are controversial opinions in the literature whether the SUS is unidimensional or bidimensional, two factor analyses were carried out. Factor loads below ± .20 were not taken into account. Only factors with eigenvalues ≥ 1 were considered (Kaiser, 1960). The values can be found in Table 2. Principal Component Analyses (PCA) were performed for each country. Kaiser-Meyer-Olkin (KMO) measures of sampling adequacy were above the minimum of .50 (Field, 2013) in each country. Bartlett's tests of sphericity were significant in each country ($p < .001$).

Germany. The KMO measure for the PCA of sampling adequacy was KMO = 0.81. The examination of the scree-plot yielded empirical justification for retaining one factor which accounted for 51.06 % of the total variance. Furthermore, because all ten items have a result of at least ± .40, the factor can be interpreted. The second factor analysis with two factors revealed, that two factors accounted for 65.43 % of the total variance. Among the factor solutions, the varimax-rotated two-factor solution yielded the most interpretable solution, and most items loaded highly on only one of the two factors.

USA. The KMO measure for the PCA of sampling adequacy was KMO = 0.83. The examination of the scree-plot yielded empirical justification for retaining one factor which accounted for 60.53 % of the total variance. Furthermore, because all ten items have a result of at least ± .40, the factor can be interpreted. Although the scree-plot indicates the presence of one factor with eigenvalues greater than 1, a two-factor solution was also performed revealing that two factors accounted for 71.94 % of the total variance.

China. The KMO measure for the PCA of sampling adequacy was KMO = 0.74. The examination of the scree-plot yielded empirical justification for retaining one factor which accounted for 48.14 % of the total variance. Furthermore, because all ten items have a result of at least ± .40,

Table 2. Results of the one-factor and two-factor solution in Germany, USA and China.

	Germany			USA			China		
	Component Matrix	Rotated Component Matrix		Component Matrix	Rotated Component Matrix		Component Matrix	Rotated Component Matrix	
Factor	1	1	2	1	1	2	1	1	2
Item 1	.765	**.833**		.569		.876	.533	.303	**.494**
Item 2	.453		**.886**	.861	**.744**	.442	.804	**.670**	.445
Item 3	.799	**.911**		.899	.629	**.660**	.855	**.641**	.574
Item 4	.584	.338	**.508**	.767	**.802**	.207	.768	**.744**	.278
Item 5	.868	**.740**	.466	.738	.306	**.824**	.696	.394	**.649**
Item 6	.724	.478	**.557**	.901	**.671**	.607	.400		**.827**
Item 7	.651	.358	**.589**	.878	**.702**	.527	.775	**.796**	.216
Item 8	.909	**.708**	.570	.788	**.865**		.684	**.762**	
Item 9	.586	**.706**		.489	**.389**	.297	.566	.296	**.562**
Item 10	.684	.237	**.786**	.778	**.875**		.728	**.784**	

Note. Bold values show on which factor the items load. Factor loads ± .20 are not shown.

the factor can be interpreted. Although the scree-plot indicates the presence of one factor with eigenvalues greater than 1, a two-factor solution was also performed showing that two factors accounted for 59.36 % of the total variance.

DISCUSSION

Assessing usability of a product or a system is crucial for its evaluation. To get access to the customer's feedback and to enable comparisons of products, psychometrically suitable methods are needed. A widely used questionnaire to assess usability is the SUS. It is applied for different technologies all around the globe in several languages and its psychometrics have been investigated in multiple studies. However, a literature research did not reveal the extent to which an international comparison of the *same* product in *different* markets is reliable, valid, and sensitive. Therefore, the aim of this study was to investigate whether the SUS leads to comparable results for the same product in different markets. For this purpose, the infotainment system of the BMW X5 was evaluated in three different countries: Germany, USA and China. The infotainment system of the BMW X5 can be considered having acceptable usability in all three countries. Participants rated the same system. Assuming there are no cultural factors affecting the evaluation, the SUS should show similar results across countries. This was the case in the present study: The total mean scores did not differ between Germany, USA and China. To further evaluate sensitivity, supplementary questionnaires or methods to assess usability should be included in further studies in order to ensure that there are no differences in the usability rating. Additionally, to proof the ability of the SUS and its translations to differentiate between different systems, more infotainment systems should be included in further studies. Analyzing Cronbach's alpha as an indicator for internal consistency revealed a high internal

consistency for each country. According to George and Mallery (2003), the results for China and Germany can be considered good and even excellent for the USA. In an item analysis for each country was found that all items of the SUS seem to belong to the scale in its respective language. Removing certain items which may not fit the higher-level construct of usability would increase indicators of internal consistency. The items that may be taken into account for deletion differ between the three countries. Despite the consistent results on reliability and validity, differences were found when comparing which item deletion would increase internal consistency in each language. This might question the assumption of the SUS being a global questionnaire. For the USA, the item analysis showed conspicuous findings for item 8. Item 8 was already discussed in previous work due to the adjective *cumbersome* (AlGhannam et al., 2018; Borkowska & Jach, 2017). The concept of something being cumbersome may not have the same relevance in different cultures. Inconsistencies in single item analyses potentially arise due to different interpretations of the adjective in different languages. Hilton and Skrutkowski (2002) already stressed that cultural and linguistic differences are often insufficiently considered in the translation of questionnaires. The results of the factor analyses are controversial. As with Brooke (1996), a one-factor analysis was used to confirm the unidimensionality of the SUS. To test the bidimensionality, a two-factor analysis has been performed. Both can be interpreted across all countries with the serious difference that the two factor analysis explains a higher proportion of the total variance. This suggests that the SUS is bidimensional which questions the globality of the SUS, since the individual items of each country load differently on the two factors. Also, the two-factor structure, in which items 4 and 10 load on the same factor (Lewis & Sauro, 2009) could not be replicated. So, construct validity could be proven, but a globality of the SUS with the factor analysis is difficult to assess. Based on the results, the English, German, and Chinese translations of the SUS received validation support to a certain extent. The mean scores ranked similar across the three languages and correlations between single items and the total score were high across languages. However, there are inconsistencies in single item reliability analyses. Which might question the assumption of the SUS being a global questionnaire with regard to its item structure. Furthermore, different factor loadings across countries hinder to assess the globality of the SUS. Overall, the three SUS translations seem to be capable of measuring the subjective usability of an infotainment system, but further research and adjustments of the translations are necessary to validate the SUS as a global questionnaire in the context of an infotainment system.

REFERENCES

AlGhannam, B.A., Albustan, S.A., Al-Hassan, A.A., Albustan, L.A. (2018). Towards a Standard Arabic System Usability Scale: Psychometric Evaluation using Communication Disorder App. International Journal of Human-Computer Interaction, Volume 34, No. 9, pp. 799–804.

Bangor, A., Kortum, P.T., Miller, J T. (2008). An Empirical Evaluation of the System Usability Scale. International Journal of Human-Computer Interaction, Volume 24, No. 6, pp. 574–594.

Baumgartner, J., Frei, N., Kleinke, M., Sauer, J., Sonderegger, A. (2019). Pictorial System Usability Scale (P-SUS). In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Volume 69.

Borkowska, A., Jach, K. (2017). Pre-testing of Polish Translation of System Usability Scale (SUS). In: Proceedings of 37th International Conference on Information Systems Architecture and Technology, Volume 521.

Borsci, S., Federici, S., Lauriola, M. (2009). On the dimensionality of the System Usability Scale: A test of alternative measurement models. Cognitive Processing, Volume 10, No. 3, pp. 193–197.

Brooke, J. (1996). SUS – A "quick and dirty" usability scale. Usability Evaluation Industry, Volume 194, No. 189, pp. 4–7.

Butler, K.A. (1996). Usability engineering turns 10. Interactions, Volume 3, No. 1, pp. 58–75.

Clemmensen, T. (2011). Templates for Cross-Cultural and Culturally Specific Usability Testing: Results From Field Studies and Ethnographic Interviewing in Three Countries. International Journal of Human-Computer Interaction, Volume 27, No. 7, pp. 634–669.

Clemmensen, T., Roese, K. (2010). An Overview of a Decade of Journal Publications about Culture and Human-Computer Interaction (HCI). In: Human Work Interaction Design: Usability in Social, Cultural and Organizational Contexts, Volume 316, pp. 98–112.

Field, A. (2013). Discovering statistics using IBM SPSS statistics, SAGE, London.

Finstad, K. (2006). The System Usability Scale and non-native English speakers. Journal of Usability Studies, Volume 1, No. 4, pp. 185–188.

Forster, Y., Hergeth, S., Naujoks, F., Beggiato M., Krems, J.F., Keinath, A. (2019). Learning to use automation: Behavioral changes in interaction with automated driving systems. Transportation research part F: traffic psychology and behaviour, Volume 62, pp. 599–614.

Gao, M., Kortum, P., Oswald, F.L. (2020). Multi-Language Toolkit for the System Usability Scale. International Journal of Human-Computer Interaction, Volume 36, No. 20, pp. 1883–1901.

George, D., Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Allyn & Bacon, Boston.

Gray, W.D., Salzman, M.C. (1998). Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. Human–Computer Interaction, Volume 13, No. 3, pp. 203–261.

Hilton, A., Skrutkowski, M. (2002). Translating instruments into other languages: Development and testing processes. Cancer Nursing, Volume 25, No. 1.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. Educational and psychological measurement, Volume 20, No. 1, pp. 141–151.

Lewis, J.R. (2018). The System Usability Scale: Past, Present, and Future. International Journal of Human-Computer Interaction, Volume 34, No. 7, pp. 577–590.

Lewis, J.R., Sauro, J. (2009). The Factor Structure of the System Usability Scale. In: Human Centered Design, Volume 5619, pp. 94–103.

Lewis, J.R., Sauro, J. (2017). Revisiting the Factor Structure of the System Usability Scale. Journal of Usability Studies, Volume 13, No. 1, pp. 17–37.

Li, R., Chen, Y.V., Sha, C., Lu, Z. (2017). Effects of interface layout on the usability of In-Vehicle Information Systems and driving safety. Displays, Volume 49, pp. 124–132.

Marzuki, M.F., Yaacob, N.A., Yaacob, N.M. (2018). Translation, Cross-Cultural Adaptation, and Validation of the Malay Version of the System Usability Scale Questionnaire for the Assessment of Mobile Apps. JMIR Human Factors, Volume 5, No. 2.

Nunnally, J.C. (1978). Psychometric Theory. McGraw-Hill, New York.

Orfanou, K., Tselios, N., Katsanos, C. (2015). Perceived usability evaluation of learning management systems: Empirical evaluation of the System Usability Scale. The International Review of Research in Open and Distributed Learning, Volume 16, No. 2, pp. 227–246.

Rennie, A.M. (1981). The application of ergonomics to consumer product evaluation. Applied Ergonomics, Volume 12, No. 3, pp. 163–168.

Svenson, O., Patten, C.J.D. (2005). Mobile phones and driving: a review of contemporary research. Cognition, Technology & Work, Volume 7, No. 3, pp. 182–197.

Tullis, T.S., Stetson, J.N. (2004). A Comparison of Questionnaires for Assessing Website Usability. In: Usability professional association conference, Volume 1.

Wang, Y., Lei, T., Liu, X. (2020). Chinese System Usability Scale: Translation, Revision, Psychological Measurement. International Journal of Human-Computer Interaction, Volume 36, No. 10, pp. 953–963.