Journal of Audiovisual Translation

Usability as a Premise of Quality: First Steps Towards the Validation of the System Usability Scale (SUS) into Spanish

Irene Tor-Carroggio ✉
Universitat Autònoma de Barcelona

Daniel Segura ✉
Universitat Autònoma de Barcelona

Olga Soler-Vilageliu ✉
Universitat Autònoma de Barcelona

Abstract

Usability is a key factor when talking about the quality of a product. The System Usability Scale (SUS) is one of the most popular tools to measure usability due to its numerous advantages and, therefore, a very useful quality assessment tool. Originally designed in English, it is available in some other languages, such as Persian and Greek but no validated version in Spanish can be found yet. This paper bridges this gap by describing the process of statistically validating the SUS into Spanish. The results show that our translation of the SUS is reliable, although our modest sample of informants (*n* = 50) leaves room for improvement and future research. The validation of the SUS is framed within a European project that will use it for its testing phase.

Key words: usability evaluation, system usability scale, SUS, questionnaire translation, questionnaire validation, Spanish.

Citation: Tor-Carroggio, I., Segura, D. & Soler-Vilageliu, O. (2019). Usability as a premise of quality: First steps towards the validation of the System Usability Scale (SUS) into Spanish. *Journal of Audiovisual Translation, 2*(2), 57–71.

Editor(s): G.M. Greco & A. Jankowska

Received: April 24, 2019

Accepted: November 21, 2019

Published: December 31, 2019

Funding: This article has been funded by the EasyTV Project (GA761999). The authors are all members of the TransMedia Catalonia research group (2017SGR113).

✉ Irene.Tor@uab.cat, https://orcid.org/0000-0003-2924-014X
✉ Daniel.Segura@uab.cat, https://orcid.org/0000-0001-7669-0755
✉ Olga.Soler@uab.cat, https://orcid.org/0000-0001-9219-1913

1. Introduction

Although often vaguely glossed as “ease of use”, usability is a difficult construct to define, probably because it is a non-functional requirement that “does not specify parts of the system functionality” (Lauesen & Younessi, 1988, p. 1). Yet, researchers such as Lauesen and Younessi (1988, p. 1–2) compiled the five factors that have traditionally been associated with any product tagged as “usable”, namely ease of learning, task efficiency, ease of remembering, understandability and subjective satisfaction. More recently, it has been formally defined under the ISO 9241-11: 2018(E) standard as “the effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments”. User satisfaction with a product or service that successfully meets their requirements is one of the angles from which quality may be looked at (Elshaer, 2012). In fact, usability has directly been conceptualised as quality of use, although this seems to be changing into quality of experience (McNamara & Kirakowski, 2005). Be that as it may, usability is a key quality factor of many products, such as successful software systems (Winter & Deissenboeck, 2008). Therefore, it seems that the following conclusion can be drawn: any product that aims at ensuring quality, must be usable.

Given both its importance and implications, usability needs to be measurable. This means that usability requirements must be tangible enough to verify and trace them during the development of any product (Lauesen & Younessi, 1988, p. 1). Usability evaluation is often carried out through questionnaires. There are many usability evaluation questionnaires available nowadays, such as the Software Usability Measurement Inventory (SUMI) (Kirakowski & Corbett, 1993), the Post-Study System Usability Questionnaire (PSSUQ) (Lewis, 1992), the Single Usability Metric (SUM) (Sauro & Kindlund, 2005), the Website Analysis and Measurement Inventory (WAMMI) (Claridge & Kirakowski, 2011) and the System Usability Scale (SUS) (Brooke, 1996). Baumgartner, Sonderegger and Sauer (2019) even developed a pictorial single-item scale for measuring perceived usability. The SUS is amongst the most popular ones and Brooke (2013, p. 29) reported that some publications have even referred to it as an “industry standard”, although it has never formally been through a standardisation process. It has a plethora of advantages, such as brevity and robustness, as well as being free of charge (Katsanos, Tselios, & Xenos, 2012, p. 302; Bangor, Kortum, & Miller, 2008). Despite its simplicity, Tullis and Stetson (2004) noted that the SUS, when compared to other usability questionnaires, yielded among the most reliable results across sample sizes. Klug (2017) underlines that it has also been successfully applied to a wide range of devices and systems, such as scholarly repositories, websites, medical technologies, decision aids and print materials. It has also been used to test landline telephones, non-web graphical user interfaces and automated telephone interfaces (Bangor et al., 2008), which proves its flexibility and lack of dependence on the system under study. Designed in the 80s by Brooke (1996), it consists of ten standardised questions, originally written in English, half of which are positive statements, while the rest are negative. Informants need to express how much they agree with the proposed statements selecting one of five options available, ranging from *strongly disagree*

to *strongly agree* on a Likert-type scale. Final scores for the SUS can range from 0 to 100 and 68 is considered the average score (Sauro, 2011b). Yet, the SUS has also some limitations that need to be taken into consideration, such as it not being a "diagnostic instrument", meaning that "it cannot contribute to understanding the underlying reasons that explain participants' perceptions about the quality of the user experience" (Katsanos et al., 2012, p. 302). Therefore, the SUS is no replacement for identifying usability problems (Sauro, 2011b). Grading the SUS is also a difficult process that can result in scoring errors if not carried out properly. Calculated grades should be normalised to obtain a percentage, which will be easier to understand especially by those not familiar with this scale and, therefore, prone to believe that a raw SUS score equals a percentage simply because it falls somewhere between 0 and 100 (Usability.gov, n.d.).

1.1. The EasyTV project

The SUS was selected as the usability testing tool for the EU-funded project EasyTV (www.easytvproject.eu). This project aims at making audio-visual content and media accessible to persons with sight and hearing loss and to the growing ageing population of Europe by developing new access services. Some examples of these new services are customised subtitles, and a crowdsourcing platform by means of which videos in sign language can be uploaded and shared. These access services are expected to grant equal and better access to audio-visual content in terms of both choice and quality.

The EasyTV project follows a user-centric approach, which means that the services developed need to be tested and approved by users at different stages of development (Matamala et al., 2017). This approach had previously been defended by, for instance, Lauesen and Younessi (1988, p. 2) in the case of usability testing. These researchers pointed out that "nobody can foresee the usability problems for a given user interface – not even usability experts (…) Only some kind of testing with real users can reveal the usability problems". They also underlined that these kind of problems need to be identified during development. For the tests it was decided that we would be focusing on the service's usability rather than their accessibility, which was taken for granted. It was thought that, since accessibility precedes usability, testing the latter would shed more light on the quality of the user experience. Also, since the "SUS data can help provide a more complete picture of the attitudes toward a website or system being tested when used in conjunction with usability test measures such as timed tasks and the number of tasks successfully completed" (Klug, 2017, n.p.), the EasyTV's live tests will consist of users executing some specific tasks first and then answering the SUS.

The SUS was chosen as the testing tool, for several reasons. First, it was considered to be a simple and quick way to find out whether it was worth continuing the development of the access services. Second, as Manchón and Orero (2018) explained, its questions apply to the whole system under scrutiny, ignoring opinions about the content presented or shown by the system

(i.e., it gives a usability score about a subtitling platform, but not about the quality of the subtitles themselves). Third, it has been used in other similar European projects, such as HBB4ALL (RBB, 2016) and ImAc (Agulló, forthcoming). Fourth, since it is technology-independent, scores can be compared regardless of the technology used. Fifth, results are easy to share and to understand (Klug, 2017), and this was deemed important, since the project is interdisciplinary. Lastly, even though a large sample of participants will be drawn for the final tests, the number of users recruited for the intermediate tests was expected to be quite modest (five to ten users per service), hence the need to find a reliable tool that can deal with such modest samples.

When carrying out any sort of tests it is vital to make participants feel comfortable. This comfort can and should be provided by, for instance, administering the questionnaire in the participants' mother tongue, even if they can speak more than one language. This is necessary not only because it can ease the complexity of the activity, but also because it has been found that participants answering questionnaires in a language different than their native one are more likely to give higher non-respondent rates (Groves & Couper, 1998). Also, Choi and Pak (2005) noted that the language and culture of the participants in a questionnaire can affect their perception of questions and, therefore, have an impact on their answers. Thus, were the SUS to be used in the testing phases of the project, it would be necessary to have its translation statistically validated into Spanish. In fact, statistical validation is imperative to validate this kind of tests after translating them because the reliability of the original version does not necessarily transfer to its translations. The change of one term or the new way a question is phrased could result in an unpredicted alteration of the final score.

This paper aims to make the tests to be carried out in the EasyTV project easier by validating the SUS into Spanish so that it can be used as a solid and reliable usability testing tool. This article is divided into four sections: the first one reviews previous research regarding the SUS and questionnaire validation; the second one describes the methodology of the study; the third one presents the results obtained; and the last one draws some conclusions, as well as discusses some of the challenges we faced in our study. It also outlines new research paths worth exploring in the future.

2. Research Background

The SUS is a widely researched usability tool which many academic and industrial papers revolve around. One of the most comprehensive studies is that of Bangor et al. (2008), who analysed 206 studies in which the SUS had been applied. Several findings are presented in their paper: to begin with, out of the 2,324 individual surveys that those studies comprised, the average SUS score was 70.14 and the median 75. Moreover, fewer than 6% of the mean scores fell below 50, and there were no group scores below 30. Also, after having used the SUS with many different devices, graphical user interface for OS-based computer interfaces scored the highest with an average score

of 75.24. Fourth, unlike with sex, a significant relationship was found between age and SUS scores. Finally, the authors also explored what an acceptable SUS score was:

> […] This means that products which are at least passable have SUS scores above 70, with better products scoring in the high 70s to upper 80s. Truly superior products score better than 90. Products with scores of less than 70 should be considered candidates for increased scrutiny and continued improvement and should be judged to be marginal at best. (Bangor et al., 2008, n.p.)

Bangor, et al. (2009) conducted a study with almost 1,000 participants using a version of the SUS with an extra question, which asked them to rate the user-friendliness of a product as *worst imaginable*, *awful*, *poor*, *OK*, *good*, *excellent*, or *best imaginable*. Their results pointed out that the scale ratings were very similar to SUS scores and that, therefore, the inclusion of an additional scale may be of help. Finstad (2010) underlined the frustration respondents can feel with the 5-point Likert scale and proposed a 7-point Likert scale with the options *entirely disagree*, *mostly disagree*, *somewhat disagree*, *neither agree nor disagree*, *somewhat agree*, *mostly agree*, and *entirely agree*. Sauro (2011b) came up with very interesting conclusions too, such as SUS scores being able to predict customer loyalty, five seconds with a system being enough to generate a stable SUS score and positively and negatively worded items being more harmful than helpful. Klug (2017) collected some tips of advice that according to researchers could facilitate successful administration of the SUS, such as reminding users of the alternative nature of the SUS statements, alternating the order in which tools are tested, and participants filling in the questionnaire as soon as they finish testing the product so that they provide an accurate summation. This last point had also been recommended by Brooke (1996).

As for the SUS statistical properties, Lewis and Sauro (2018) report that, initially, 10 SUS items were selected from an initial pool of 50, based on the responses of 20 people who used the full set of items to rate two software systems, one of which was known to be easier to use that the other. The items selected for the SUS were those that provided the strongest discrimination between the systems. Since the original SUS just reported strong correlations between the selected values, some researchers have investigated its reliability (Bangor et al., 2008), validity (Bangor et al., 2008) and sensitivity (Lewis & Sauro, 2009).

2.1. The Concept of Validity and the Validation Process

Gandek and Ware (1998, p. 953) briefly defined "validity" as "the extent to which a score [of a questionnaire] means what is supposed to mean". Arribas (2004, p. 27) differentiates three kinds of validity in a questionnaire: content validity (the extent to which the items in that questionnaire are indicative of what we want to measure), construct validity (the degree to which the questionnaire reflects the theory on the measured concept) and criteria validity (the comparison of each subject's score on that questionnaire with a gold standard).

Thus, we can conclude that validity, as a broad concept used in the context of questionnaire design, refers to the ability of a questionnaire to accurately measure a variable representing an object of study. Under that premise, we define "validation" as the process followed to corroborate a questionnaire's validity on an empirical basis. There can be many procedures to achieve that, hence the word validation referring to a specific methodology. It must be noted, however, that some translation scholars have already used the terms validity and validation before with very different meanings in mind. For example, Martín Casado and Sánchez-Reyes (1999, p. 149) defined validation as "the adaptation of the questionnaire to the receiving culture's concepts".

Regarding the topic of questionnaire translation and validation, extensive literature can be found. As an example, we can take the paper by Aguilar, de la Garza González, Miranda, and Villegas (2016) explaining their effort to validate the Computer System Usability Scale questionnaire (Lewis, 1995) in Spanish. Later, Aguilar and Villegas (2016) continued their research on questionnaires and validated an adaptation of the SUS that consisted of a positive version of the original questions. Also, some papers have been written about several translations of the SUS questionnaire and their respective validations. Greek (Katsanos et al., 2012), Persian (Dianat, Ghanbari, & AsghariJafarabadi, 2014), European Portuguese (Martins, Rosa, Queirós, Silva, & Rocha, 2015) Slovene (Blažica & Lewis, 2015) and American Sign Language (Huenerfauth, Patel, & Berke, 2017) are some of the languages the SUS has already been validated in. It has also been translated into Swedish and Finnish, but these were ad hoc translations that lacked psychometric evaluation (Blažica & Lewis, 2015, p. 112). Similarly, a German translation was produced by Rummel, Ruegenhagen, & Reinhardt (2013) and validated using backwards translation, i.e., translating the SUS in German back to English, but doing this alone does not completely validate the questionnaire (Blažica & Lewis, 2015, p. 112). Even though Sauro (2011a) reports the existence of unofficial Spanish translations, none of them seems to have been validated so far.

3. Methodology

The first step to validate the Spanish translation of the SUS was, obviously, to translate it. For that purpose, we decided to follow the instructions regarding questionnaire translation described by Tsang, Royse and Terkawi (2017). Their method was chosen because it is simple to follow, comprehensive and has proven to be effective in the past. Also, very similar procedures were followed in other articles focusing on questionnaire translation and validation, such as Domínguez, Balkrishnan, Ellzey, and Pandya's (2006). The first step was to establish an expert committee, which Terkawi et al. (2017, n.p.) suggest as a preliminary step to "produce a prefinal version of the translation". According to these researchers (2017, n.p.), the committee "should include experts who are familiar with the construct of interest, a methodologist, both the forward and backward translators, and if possible, developers of the original questionnaires". Representatives of all the suggested categories were on the committee except for the developers of the original questionnaire. The committee agreed on, for example, the profile

of users that would participate in the validation. It was decided that the users in the study would be English speakers (both native and non-native) and native Spanish speakers, as respectively the control and the experimental group, from all age groups and of various educational backgrounds. The committee also chose the system to be tested in terms of usability (the EasyTV project website), as the results could also shed some light on what improvements could be made in it and therefore help the project, although that was not the main objective of our study. The main translation challenges were also discussed within this group and, with that feedback, the SUS questionnaire was translated into Spanish by two of the present study's researchers, who are Spanish professional translators. This team-based approach is said to be preferred among translation researchers, since it generates more translation options and provides more solid and less idiosyncratic translation evaluation (Forsyth, Stapleton, Lawrence, Levin, & Lewis, 2006, p. 4114). They worked separately and reached an agreement on the final draft, which was then sent to two different translators. These were briefed on the task and carried out a backwards translation into English. The researchers were available to answer questions and provide guidance, in case it was necessary. The objective of this exercise was to check that the translation did not convey any other meaning that was not present in the original questionnaire. Although the persons carrying out the backwards translation were professional translators, they were Spanish too, which obviously constituted a limitation in our research.

The following version of the SUS questionnaire was elaborated based on the input received from the two backwards translations. It was then posted on Google Forms, preceded by five tasks that the informants had to carry out in order to evaluate the usability of the EasyTV's website. The questionnaire also included some demographic questions about the informants' sex, age and mother tongue. The informants were also asked if they had already visited the website, since it was pointed out by Sauro (2011a) that previous experience with the website under study results in higher SUS scores. An important remark to be made is that instead of using the word *sistema* ("system"), which is the one used in the official version, we opted for substituting it with *página web* ("website"). Bangor et al. (2008) also reported that using "product" instead of "system" based on user feedback did not affect the results. It must also be highlighted that an English version of the questionnaire was also drafted so that it could be used in the control group.

For the pilot, four informants were recruited. After answering the questionnaire in Spanish, they also participated in a cognitive interview, similar to the one reported in Forsyth et al. (2006). This interview aimed at finding out how they understood each of the SUS questions and their feelings on the proposed tasks. The results of the pilot brought about several modifications. First, it was discovered that each informant had a different understanding of what the word "usability" means, so the researchers decided to eliminate this word in the title of the questionnaire to avoid biasing the subjects beforehand. The questions that contained this word were also rephrased. For example, the question "Did you find the website usable?" was substituted for "Did you feel you could use the website in an effective, efficient and satisfactory way?". It was also found out that the informants took less time to complete the tasks than what the researchers had anticipated,

so the information on how much time was needed was modified accordingly. Also, some of the tasks were changed or rewritten, as they were not clear enough to the informants or useful for the researchers. This is the case of, for example, a task that originally asked the respondents if they would be able to send an email to the project leader if needed. It was discovered that some of them did not even try to do it and just answered "yes". Thus, it was decided that all tasks should look for specific information rather than just offering a yes/no type of answer. The following questions, extracted from the final questionnaire, indicate what kind of tasks informants had to perform in order to be able to answer them:

1. How many publications does Pilar Orero have (with other co-authors)?
- 3
- 4
- 5
- Don't know

2. How many members does the EasyTV consortium have?
- 7
- 8
- 9
- Don't know

3. What sex is the sign language interpreter in the "general overview"?
- Male
- Female
- Don't know

4. How many languages are available in the EasyTV website?
- 3
- 4
- 5
- Don't know

5. How many objectives does EasyTV have?
- 4
- 5
- 6
- Don't know

6. How many categories can the news section be divided into?
- 2
- 3
- 4
- Don't know

Regarding the translation itself, it was revealed that some words like *inconsistencia* "inconsistency" led to confusion and they had to be changed (in this case to *inconsistente* "inconsistent"). This was not the first time that a word was found to be problematic, as the word "cumbersome" had already been found uncertain in previous studies in English and had, therefore, been changed for "awkward" (Bangor et al., 2008). Lastly, some commentaries about the questionnaire design were also made (for example, informants complained that the questions were too repetitive), but the researchers could do nothing about that, as the goal of the study was to validate the translation, and not to modify or improve the original SUS. The final version of the SUS that was approved was the following:

Table 1.

Items of the SUS and their Translation into Spanish

English version of SUS	Spanish version of SUS
1. I think that I would like to use this system frequently.	1. Creo que me gustaría usar esta página web frecuentemente.
2. I found the system unnecessarily complex.	2. Me ha parecido innecesariamente compleja esta página web.
3. I thought the system was easy to use.	3. La página web me ha parecido fácil de usar.
4. I think that I would need the support of a technical person to be able to use this system.	4. Creo que necesitaría la ayuda de una persona con conocimientos técnicos para usar esta página web.
5. I found the various functions in this system were well integrated.	5. Me ha parecido que las distintas funciones de esta página web están bien integradas.
6. I thought there was too much inconsistency in this system.	6. Creo que la página web es demasiado inconsistente.
7. I would imagine that most people would learn to use this system very quickly.	7. Imagino que la mayoría de la gente aprendería a usar esta página web de forma muy rápida.
8. I found the system very cumbersome to use.	8. La página web me ha parecido engorrosa.
9. I felt very confident using the system.	9. Tenía muy claro cómo usar esta página web todo el rato.
10. I needed to learn a lot of things before I could get going with this system.	10. Tuve que adquirir muchos conocimientos antes de poder usar esta página web.

Once the final version was agreed on, the approval of our university's ethics committee was requested and obtained in March 2019. The questionnaires, both in Spanish and in English were distributed between March and early April 2019. Although the users recruited for the EasyTV tests will be functionally diverse and elderly persons, there was no specific demographic characteristic requested for this study, since the SUS can be applied regardless of the user profile. The test results are described next.

4. Results

In this section, the results obtained during the tests will be analysed. We will try to find out if the questionnaire is reliable by using statistical methods and if the translation is valid by comparing the SUS scores of the English and Spanish versions. Also, we will analyse the usability of the EasyTV webpage on its own.

4.1 Spanish SUS Validation

A total of 50 informants answered the questionnaire in Spanish. Among those, 70% chose Spanish as their mother tongue, while 25% chose Catalan (2 users, or 5% of the total, chose "other"). The control group consisted of 19 informants, who performed the same exact tasks and answered the questionnaire in English. Because of the difficulty of finding native English speakers (only 10% of the respondents), non-native speakers with high language proficiency were allowed to answer. As the questionnaire's content was not difficult and the tasks were simple, we do not think that the control group results were affected by this choice. What could affect the results, however, were each participant's technological skills, as the system being tested was a website and the tasks to be done in the tests required some basic computer knowledge. Because of that, informants were asked how often they surfed the internet. All participants in the control group answered "Every day or almost every day", while 98% of the ones in the experimental group would give the same answer (1 user answered "3 times per week or more"). The participants in the control group were slightly younger on average than the ones in the experimental group (the average being 29.58 and 32.52, respectively). The questionnaires were distributed through Google Forms and the data for the statistical analysis was processes using the software IBM SPSS (version 23). Before performing any analyses, we corrected the SUS scale using the transformations to obtain positive scoring between 0 and 4 for all 10 items. This is important because questions 2, 4, 6, 8 and 10 are reversed questions.

We first tested the reliability of the Spanish version of SUS with two different reliability tests: Cronbach's alpha, a widely used measure of internal consistency, and Guttman lambda-2. Both Cronbach's alpha coefficient (.904) and lambda2 values (.911) are very high for the ten items, indicating that the questionnaire is highly consistent.

We also carried a factor analysis to identify possible subscales. Lewis and Sauro (2009) detected that English SUS can be subdivided into two subscales that measure Usability (questions 1, 2, 3, 5, 6, 7, 8, 9) and Learnability (questions 4 and 10). A factor analysis using common varimax rotation with Kaiser normalisation extracted two principal components, clearly identifying the suggested. Table 2 shows the components identified in the SUS scale. Questions 4 and 10 are clearly related to a different dimension than the rest.

Table 2.

Rotated Component Matrix (rotation converged in 3 iterations)

	Component 1	Component 2
Q7N	.795	
Q2N	.778	
Q1N	.775	
Q3N	.774	
Q9N	.769	
Q8N	.754	
Q5N	.745	
Q6N	.513	
Q10N		.894
Q4N		.814

Furthermore, we tested the reliability within both subscales. Cronbach's alpha for usability scale (8 items) is .907, and for learnability scale (2 items) equals .786.

According to these results, the Spanish version of SUS is a highly reliable scale. Within the scale we could identify two subscales, as defined by Lewis and Sauro (2009). Both subscales, usability and learnability, also show good consistency.

Next analyses consisted of comparing Spanish SUS scores with the control group scores, that is, the group of 19 people that answered the questionnaire in English. We ran a non-parametric test, the Independent-Samples Mann-Whitney U Test, which did not detect any differences in the scoring of all the questions across groups (all $p > .05$), meaning that the scores are not statistically different from one to another.

4.2 Evaluation of EasyTV website

The results obtained in our analyses indicate that our translation of the SUS questionnaire in Spanish is consistent and comparable to English SUS. Therefore, we can make use of the evaluation that our informants did on the EasyTV website' usability. The English SUS group qualified the website

with a mean score of 74.07, whereas the Spanish SUS group gave a rating of 69.6. The difference between these two scores cannot be interpreted as significant taking into account the number of respondents for each questionnaire, especially the one in English (<20). Both ratings are above the average score that Sauro (2011b) specified (68). Also, if we combine both scores we obtain a mean score of 71.83, which, since it is above 70, it is "at least passable" (Bangor et al., 2008). Nonetheless, it stands to reason that this score leaves room for further improvement so that the website's usability can increase. In fact, future improvements could and should start by looking at the issues encountered by some informants. For example:

1. "The web is not user-friendly".
2. "What I found the hardest to find were the publications".
3. "Although there are several languages to choose from, not all the functions have the corresponding translation".
4. "I think the website is easy to use and clear. I found all the information easily. My only comment perhaps is the images chosen for the front page – I think they are not very clear in terms of the message that the website wants to send. They are very cluttered when compared to the rest of the website."

Finally, it should be borne in mind that the scores could change depending on the informants' profile (for example, it could be less usable for users with functional diversity or of certain age), but, as it was mentioned in the methodology section, the demographic characteristics of the informants were not taken into account in this test. In general, though, we can conclude that the EasyTV project website is usable enough and the users were satisfied with it.

5. Conclusions

This study has its limitations. First, the backwards translation carried out during the process of coming up with the final Spanish version of the SUS was performed by two non-English speakers. Second, the samples obtained in the questionnaires (in both the English and the Spanish versions) were not randomly selected and their sizes also leave room for improvement. Third, some of the persons who filled in the questionnaire in English were not native speakers, which obviously can compromise the results. In fact, Finstad (2006) had found that non-native speakers had difficulty in understanding the word "cumbersome" as opposed to English native speakers.

Due to the limited sample of users, this article can be read as a first step towards a definitive validation of the SUS questionnaire in the Spanish language. The next logical step towards successful validation would be to conduct a study involving a higher number of informants. The vast majority of validation studies cited in the current article used samples ranging from 150 to 300 informants.

Consequently, we think that a similar number of informants for both the control group and the test group would yield statistically more powerful results.

However, even considering these limitations, the tests carried out show that the Spanish translation of the SUS questionnaire has been an (interim) success. The results of Cronbach's Alpha and Guttman Lambda tests demonstrate that the questionnaire is very reliable. This happens both if we consider the 10 questions as a whole and if we measure usability and learnability as two separate variables. Also, no differences were detected in the SUS scores in the English and Spanish versions. This demonstrates that the phrasing of the questions in the Spanish version did not lead to a different interpretation of their meaning or change the original questions' goal. We can conclude that, at least with our limited number of informants, the translation of the SUS questionnaire is good and precise enough as to be used in subsequent usability tests and, therefore, to guarantee their quality.

References

Aguilar, M. I. H., de la Garza González, A., Miranda, M. P. S., & Villegas, A. A. G. (2016). Adaptación al español del Cuestionario de Usabilidad de Sistemas Informáticos CSUQ/Spanish language adaptation of the Computer Systems Usability Questionnaire CSUQ. *Revista Iberoamericana de las Ciencias Computacionales e Informática*, *4*(8), 84–99.

Aguilar, M. I. H., & Villegas, A. A. G. (2016). Análisis comparativo de la Escala de Usabilidad del Sistema (EUS) en dos versiones/Comparative analysis of the System Usability Scale (SUS) in two versions. *Revista Iberoamericana de las Ciencias Computacionales e Informática*, *5*(10), 44–58.

Agulló, B. (forthcoming). Technology for subtitling: a 360-degree turn. To appear in *Hermeneus*.

Arribas, M. (2004). Diseño y validación de cuestionarios [Design and validation of questionnaires]. *Matronas Profesión*, *5*(17), 23–29.

Bangor, A., Kortum, P., & Miller, J. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594.

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.

Baumgartner, J., Sonderegger, A., & Sauer, J. (2019). No need to read: Developing a pictorial single-item scale for measuring perceived usability. *International Journal of Human-Computer Interaction*, *122*, 78–89.

Blažica, B., & Lewis, J. (2015). The SUS-SI. Intl. *Journal of Human–Computer Interaction*, *31*, 112–117.

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* (pp. 189–194). London: Taylor and Francis.

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

Choi, B. C., & Pak, A. W. (2005). Peer reviewed: A catalog of biases in questionnaires. *Preventing chronic disease*, *2*(1).

Claridge, N., & Kirakowski, J. (2011). WAMMI: website analysis and measurement inventory questionnaire. Retrieved from http://www.wammi.com/

Dianat, I., Ghanbari, Z., & AsghariJafarabadi, M. (2014). Psychometric properties of the Persian language version of the System Usability Scale. *Health Promotion Perspectives*, *4*(1), 82–89.

Dominguez, A., Balkrishnan, R., Ellzey, A., & Pandya, A. (2006). Melasma in Latina patients: Cross-cultural adaptation and validation of a quality-of-life questionnaire in Spanish language. *Journal of the American Academy of Dermatology*, *55*(1), 59–66.

Elshaer, I. (2012). What is the meaning of quality? Retrieved from https://mpra.ub.uni-muenchen.de/57345

Finstad, K. (2006). The system usability scale and non-native English speakers. *Journal of Usability Studies*, *1*(4), 185–188.

Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies*, *5*(3), 104–110. Retrieved from http://uxpajournal.org/response-interpolation-and-scale-sensitivity-evidence-against-5-point-scales/

Forsyth, B., Stapleton, M., Lawrence, D., Levin, K., & Lewis, G. (2006). Methods for translating survey questionnaires. AAPOR – ASA Section on Survey Research Methods, 4114–4119. Retrieved from https://www.researchgate.net/publication/239856416_Methods_for_Translating_Survey_Questionnaires1

Gandek, B., & Ware Jr, J. E. (1998). Methods for validating and norming translations of health status questionnaires: the IQOLA project approach. *Journal of Clinical Epidemiology*, *51*(11), 953–959.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238.

Huenerfauth, M., Patel, K., & Berke, L. (2017). Design and psychometric evaluation of an American Sign Language translation of the System Usability Scale. *Proceedings of the 19th International ACM SIGACCESS Conference*.

Groves, R. M., & Couper, M. P. (1998). *Non-response in household interview surveys*. New York: Wiley & Sons.

ISO (International Organization for Standardization) (2018). *Ergonomics of human – system interaction. Part 11: Usability: Definitions and concepts* (ISO 9241-11:2018(E)).

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. *Proceedings of the 2012 16th Panhellenic Conference on Informatics*.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, 24(3), 210–212.

Klug, B. (2017). An overview of the System Usability Scale in library website and system usability testing. *Journal of Library User Experience*, *1*(6). Retrieved from https://quod.lib.umich.edu/w/weave/12535642.0001.602?view=text;rgn=main

Lausen, S., & Younessi, H. (1988). Six styles for usability requirements. In *Proceedings of REFSQ'98.* Namur: Presses Universitaires de Namur.

Lewis, J. (1992). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *36*(16), 1259-1260. Los Angeles, CA: SAGE Publications.

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

Lewis, J., & Sauro, J. (2009). The factor structure of the System Usability Scale. In M. Kurosu (Ed.), *Human Centered Design, HCII 2009* (pp. 94–103). Heidelberg, Germany: Springer-Verlag.

Lewis, J., & Sauro, J. (2018). Item benchmarks for the System Usability Scale. *Journal of Usability Studies*. *13*(3), 158–167

Manchón, L. M., & Orero, P. (2018). Usability tests for personalised subtitles. *Translation Spaces*, *7*(2), 263–284.

Martins, A. I., Rosa, A. F., Queirós, A., Silva, A., & Rocha, N. P. (2015). European Portuguese validation of the system usability scale (SUS). *Procedia Computer Science*, *67*, 293–300.

Matamala, A., Orero, P., Rovira-Esteva, S., Casas-Tost, H., Morales-Morante, F., Soler-Vilageliu, O., Agulló, B., Fidyka, A., Segura, A., & Tor-Carroggio, I. (2018). User-centric approaches in access services evaluation: Profiling the end user. *Proceedings of the Eleventh International Conference on Language Resources Evaluation (LREC 2018),* 1–7.

McNamara, N., & Kirakowski, J. (2005). Defining usability: quality of use or quality of experience? In *Proceedings of the International Professional Communication Conference* (pp. 200–204). Piscataway, NJ: IEEE.

Rummel, B., Ruegenhagen, E., & Reinhardt, W. (2013). System Usability Scale [German trans.]. Retrieved from http://www.sapdesignguild.org/resources/sus.asp

RBB. (2016). D3.4 – Pilot-A evaluations and recommendations. Retrieved from http://pagines.uab.cat/hbb4all/sites/pagines.uab.cat.hbb4all/files/d3.4-rbb_pilot-a-evaluation-and-recommendations_v1.00.pdf

Sauro, J., & Kindlund, E. (2005). A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 401–409. New York, NY: ACM Press.

Sauro, J. (2011a). *A practical guide to the System Usability Scale: Background, benchmarks, & best practices*. Denver, CO: Measuring Usability LLC.

Sauro, J. (2011b) SUStisfied? Little-known system usability scale facts. *User Experience: The Magazine of the User Experience Professionals Association*, *10*(3). Retrieved from http://uxpamagazine.org/sustified/

Martín Casado, M., & Sánchez-Reyes, M.S. (1999). Más allá de la traducción: la validación de cuestionarios científico-técnicos inglés-español [Beyond translation: the validation of scientific-technical questionnaires English-Spanish]. *Livius (Revista de Estudios de Traducción)*, *14*, 149–155.Terkawi, A. S., Tsang, S., Abolkhair, A., Alsharif, M., Alswiti, M., Alsadoun, A., & Altirkawi, K. A. (2017). Development and validation of Arabic version of the short-form Mcgill pain questionnaire. *Saudi Journal of Anaesthesia*, *11*(Suppl 1), S2.

Tsang, S., Royse, C., & Sulieman, A. (2017). Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi Journal of Anaesthesia*, *11*(1): S80–S89. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5463570/

Tullis, T., & Stetson, J. (2004). A comparison of questionnaires for assessing website usability. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.396.3677&rep=rep1&type=pdf

Usability.gov. (n.d.). *System Usability Scale (SUS)*. Retrieved from http://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

Winter, S., & Deissenboeck, F. (2008). A comprehensive model of usability. In J. Gulliksen, M. B. Harning, P. Palanque, G. C. van der Veer, & J. Wesson (Eds.), *Engineering Interactive Systems. EIS 2007 Joint Working Conferences EHCI 2007, DSV-IS 2007, HCSE 2007, Salamanca, Spain, March 22–24, 2007. Selected Papers* (pp. 106–122). Berlin, Heidelberg: Springer.