PRE-PRINT COPY

To cite this article:

Diana Castilla, Irene Jaen, Carlos Suso-Ribera, Gemma Garcia-Soriano, Irene Zaragoza, Juani Breton-Lopez, Adriana Mira, Amanda Diaz-Garcia & Azucena Garcia-Palacios (2023): Psychometric Properties of the Spanish Full and Short Forms of the System Usability Scale (SUS): Detecting the Effect of Negatively Worded Items, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2023.2209840

To link to this article: https://doi.org/10.1080/10447318.2023.2209840

FULL AND SHORT FORM OF SPANISH SYSTEM USABILITY SCALE IS AVAILABLE AT:

https://docs.google.com/document/d/1B6K3BpoUaEZ2b-kCGQ3JKkESKL_Moqmh/edit?usp=drive_link&ouid=117372221466962332753&rtpof=true&sd=true

PSYCHOMETRIC PROPERTIES OF THE SPANISH FULL AND SHORT FORMS OF THE SYSTEM USABILITY SCALE (SUS): DETECTING THE EFFECT OF NEGATIVELY WORDED ITEMS

Diana Castilla[1,], Irene Jaen*[1,2], Carlos Suso-Ribera[2], Gemma Garcia-Soriano[1], Irene Zaragoza[3], Juani Breton-Lopez [2], Adriana Mira[1], Amanda Diaz-Garcia[4], Azucena Garcia-Palacios[2]

[1]Departamento de Personalidad, Evaluación y Tratamiento Psicológico, Universidad de Valencia, Valencia, Spain
[2] Departamento de Psicología Básica, Clínica y Psicobiología, Universitat Jaume I, Castellón de la Plana, Spain.
[3] Departamento de Informática, Universidad de Valencia, Valencia, Spain
[4] Departamento de Psicologia y Sociologia. Universidad de Zaragoza
* Corresponding author

Address for correspondence (current adress):

Universitat Jaume I. Facultad de Ciencias de la Salud. Departamento de Psicología Básica, Clínica y Psicobiología. Avda. Sos Baynat s/n. Castellón de la Plana E-12071 España, +34 964 38 76 46, ijaen@uji.es

Funding sources: This study was funded in part by Consellería de Innovación, Universidades, Ciencia y Sociedad Digital under the grant "Ayudas urgentes para proyectos de investigación, desarrollo tecnológico e innovación (I+D+i) por la covid19" [GVA-COVID19/2021/074]; and by MCIN/AEI/10.13039/501100011033 and ERDF under the grant [RTI2018-098349-B-I00].

Psychometric properties of the Spanish full and short forms of the System Usability Scale (SUS): detecting the effect of negatively worded items

This study examined the SUS's psychometric properties with the Spanish population considering the plausible method effects associated with negatively worded items. A short form consisting of either direct or reversed items was also examined. Participants were 1321 Spaniards who completed the SUS. Confirmatory analyses showed that the SUS was a valid measure with a one-factor structure when method errors associated with negatively worded items were considered (CFI=.932, TLI=.898; RMSEA=.055, CI 90%= 0.047, 0.062), and shown evidence of reliability (Cronbach's alpha=.76). The short version with only positively worded items also showed to be a valid (CFI=.973, TLI=.946; RMSEA=.057, CI 90%= .041, .075) and reliable measure (Cronbach alpha=.77) to assess usability. This is the first study to clarify the effect of negatively worded items in the structure of the SUS and propose a short version of the SUS to be used with Spaniards when a brief version is preferred to test usability.

Keywords: System Usability Scale, validity, reliability, method effects, Spanish, psychometric properties

Introduction

The International Standards Organization (ISO 9241-11:2018) defines usability as *"the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*". In the last years, there has been an acceleration in the use of health technologies (Wind et al., 2020). Usability is a desirable feature of any technology, and it has recently been considered a prerequisite for the success of health and wellness applications as its lack turns ICTs into a barrier to the delivery of psychological treatments (Liew et al., 2019). A guide by the World Health Organization (2019) highlighted the importance of identifying the potential barriers and facilitators associated with adopting technology in health settings. For this purpose, one of the first stages when developing a technological device is testing usability since it provides information about the strengths and weaknesses of development, which prevent applications from being difficult or tedious to use (Baharuddin et al., 2013).

Currently, several self-report measures of usability exist. Some of the most popular examples are the Post-Study System Usability Questionnaire (PSSUQ; Lewis, 1995), the Computer System Usability Questionnaire (CSUQ; Lewis, 1995), the Usability Metris for User Experience (UMUX; Finstad, 2010), or the System Usability Scale (SUS; Brooke, 1996) among others. Of these, the latter has become the most popular (Maramba et al., 2019; Sauro & Lewis, 2009), accounting for the 43% of post-test questionnaire usage (Lewis & Sauro, 2017). The use of SUS has grown considerably in recent years, as evidenced by its growing number of citations. The paper that introduces the SUS (Brooke, 1996) had 5,664 citations on Google Scholar five years ago (Lewis, 2018), and this same paper currently has 15,667 citations (accessed on 2/13/2023), in other words, in just five years it has had 10,000 citations.

The SUS is a 10-item scale developed by Brooke (1996) and is considered a quick method that allows low-cost assessments of usability. Brooke (1996) described the SUS as a valuable, robust, and reliable tool that provides a global, subjective view of usability, although he did not report reliability or validity estimates. The main advantages of this questionnaire are: 1) studies using large sample sizes have shown evidence of the validity of the SUS (Bangor et al., 2008); 2) it is an agile questionnaire with only 10 items; and 3) no fees are required to use the SUS, only acknowledgement to its original source (Brooke, 1996).

Evidence of the psychometric properties of the SUS has been reported in several languages, including Arabic (AlGhannam et al., 2017), Italian (Borsci et al., 2009; Borsci, Federici, Bacci, et al., 2015; Borsci, Federici, Mele, et al., 2015), Slovene (Blažica & Lewis, 2015), Persian (Dianat et al., 2014), Polish (Borkowska & Jach, 2016), Danish (Hvidt et al., 2020), Finnish (Raita & Oulasvirta, 2011), Portuguese (Rosa et al., 2015), and Spanish (Sevilla-Gonzalez et al., 2020). These studies have reported evidence of reliability (Cronbach's alphas from 0.79 to 0.82) and construct validity. However, some inconsistent results have been obtained regarding its internal structure. Whereas some investigations replicated the one-factor structure of the original study (Bangor et al., 2008; Kortum & Sorber, 2015; Lewis, Brown, et al., 2015; Lewis, Utesch et al., 2015; Sauro & Lewis, 2011), other studies revealed a two-factor structure (Lewis & Sauro, 2009). In this sense, research showing a two-factor structure (Lewis & Sauro, 2017) supports the idea that the internal structure might be obscured because method effects associated with positively- and negatively-toned items, which might lead to different factors, as shown in previous research with scales using a similar approach (Wolgast, 2014). Thus, although evidence powerfully indicates that SUS is two-dimensional as a function of item tone, Lewis & Sauro (2017) suggested that these

dimensions are not of theoretical or practical relevance, and the best research practice is to interpret SUS as a unidimensional measure. In addition, other studies also supported that the structure of the SUS is built by two correlated factors: usability and learnability (item 4 and item 10). (Borsci et al., 2009; Lewis & Sauro, 2009). Also, the study conducted by Borsci et al. (2009) showed that the higher the perceived usability, the higher the perceived learnability, concluding that the two factors were probably indissociable in practice.

Based on the previous research, it seems necessary to increase the robustness of the SUS, exploring the effect of negatively worded items, which may introduce systematic measurement errors leading to interpretation errors regarding its internal structure. In addition, although a Spanish- Language version of the SUS has been previously validated (in a study conducted in Mexico with 88 participants Sevilla-González, Moreno-Loaeza, Lazaro-Carrera et al., 2020), its internal structure was not tested. Also, there might be cross-cultural differences between Mexicans and other Spanish speaker populations, such as Spaniards. Following this, the present study's first aim is to translate and validate a Spanish adaption of the SUS to be used by Spanish researchers and engineers. For this purpose, sources of construct validity and reliability will be tested. Specifically, this study will test the internal structure of the SUS by performing different Confirmatory Factor Analyses (CFA) with one and two-factor models that have been proposed in previous research. In addition, this study aims to investigate the effect of the use of negatively-worded items on the internal structure of the SUS. Finally, we will test the psychometric properties of two 5-item forms of the SUS comprised of either positive or negative items only, which can be useful in research when shorter measures are needed.

Method

Participants

Participants were 1321 Spaniards (50.5% males, 48.8% females, 0.2% other, and 0.2% preferred not to specify their gender). Their age ranged from 18 to 81 years (mean age=36.86 years, SD=13.93). This sample size is clearly above the minimum established to obtain adequate results for validation studies (Guadagnoli & Velicer, 1988).

The sample for this study was collected from 8 different research studies conducted from 2012 to 2022 at the [blinded for peer review], the psychological assistance service from the [blinded for peer review], and different hospital pain units in Spain. All the research studies were performed with applications used in real user contexts.

Six web apps were intended to provide psychoeducational information (35% of users), two web apps provided general information (13% of users), one smartphone app was intended for daily pain assessment in people with chronic pain and cancer pain (3% of users) and one smartphone app in serious game format was intended to provide psychoeducational information to the general population (47%).

SUS was added to the evaluation protocol as a summative evaluation of the applications after the first or second session of use. In addition, 52,2% of the whole sample belonged to the general population, while 47,8% were persons from a clinical population. Moreover, 50,3% of users used web platforms while 49,7% used APPs for mobile devices. The study has been performed in accordance with the Declaration of Helsinki. All the participants provided written consent to participate in the study.

Instruments

Spanish adaptation of the System Usability Scale (SUS; Brooke, 1996). The SUS is a 10-item self-report measure designed for usability assessment. Items are rated on a 5-

point Likert scale (1 = strongly disagree, 5 = strongly agree), ranging their contribution from 0 to 4. Total SUS scores have a range of 0 to 100, being higher scores in the SUS are indicative of greater usability. The psychometric properties of the scale have shown evidence of validity and reliability (Bangor et al., 2008). The translation into Spanish was conducted following the back-translation method proposed by Brislin (1976). Two Spanish-speaking authors of the present study, who were familiar with the SUS, translated the English version into Spanish. Next, a professional translator unfamiliar with the questionnaire translated the SUS from Spanish to English. Finally, the study authors discussed and resolved the discrepancies between the English and the Spanish versions. Table 1 shows the Spanish-translated version of the SUS with the correspondence items in English.

As suggested by Bangor et al. (2008), item 8 of the System Usability Scale (SUS) presented some difficulties due to the use of the term "cumbersome," which was not easily understood by some users. To address this issue, we followed their recommendation and replaced the Spanish equivalent of "cumbersome" (i.e., "farragoso") with "awkward" (i.e., "engorroso"). However, we found that the latter term was also difficult to understand for some Spanish-speaking users. As a result, we ultimately opted to replace it with the more commonly recognised synonym: "difficult" ("difícil" in Spanish). In addition, to enhance the clarity of the statement, we substituted the general term "system" with the specific name of the product in each study.

Data analysis

Because there is evidence of different internal structures of the SUS in past research, we performed a series of Confirmatory Factor Analyses (CFA) using R Studio. Model 1 was a one-factor model including all the items of the SUS. Model 2 was comprised of two correlated factors, one with positively worded items and the other

with negatively worded items. Model 3 was a two-correlated factor model with one factor for usability items (1, 2, 3, 5, 7, 8 and 9) and one factor with two items (4 and 10) that has been associated with learnability (Borsci et al. 2009). Model 4 is based on Correlated Traits Correlated Methods (CTCM) models, which allow the separation of substantive content from the method used to collect the responses (DiStefano & Motl, 2006). Following this CTCM framework, Model 4 has two factors, one including all the items of the SUS and a negative method factor comprised of the negatively worded items (2, 4, 6, 8 and 10). Correlation between the two factors was fixed to zero. Finally, two additional 5-item models were tested only with positive (Model 5) or negative items (Model 6) in order to test the internal consistency of a short form of the SUS.

To investigate the fit of the data to the models, the Comparative Fit Indexes (CFI), the Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA) indices were used. For interpretation purposes, CFI and TLI >.90 and >.95 indicated good and excellent adequacy of the model, respectively (Marsh et al., 2004). RMSEA values ≤.10 indicate an acceptable fit, but values ≤.05 are preferred (Weston & Gore, 2006). Confidence intervals (90%) for RMSEA were also calculated and reported. For CFA analyses, the maximum likelihood with robust standard errors (MLR) estimator was used due to the ordinal nature of the data and the lack of normality in the data distribution. The reliability of the System Usability Scale was examined using SPSS version 26 with Cronbach's alpha. After, Student's t-test analyses for independent samples were calculated to compare the means obtained in the SUS by gender.

Results

Descriptive analysis of the items

The distributional properties of the 10 items of the Spanish version of the System Usability Scale (i.e., means, standard deviations, skewness, and kurtosis) are presented in Table 2. Skewness ranged from -0.68 to -2.94. Kurtosis ranged from -0.08 to 9.44, obtaining a leptokurtic distribution (>3) for items 2, 4, 6, 7, 8 and 10. Negatively worded items (even items) were reversed for the descriptive analysis.

Structural validity evidence

The goodness of fit indices for each confirmatory analysis is presented in Table 3. Model 1 (one factor) and Model 3 (two factors: usability and learnability) showed poor fit indices. Model 2 and Model 4 showed good fit indexes for CFI and TLI, being superior to the CFI and TLI of Model 4, which has a method factor representing the negatively worded items. Also, the RMSEA index was superior for Model 2 compared with Model 4. Regarding Model 5 and Model 6, although both showed adequate indexes, the model comprised of positive items (Model 5) showed the best goodness fit with excellent CFI and good TLI and RMSEA. Factor loadings for the models tested with confirmatory analyses are reported in Table 4 (all p's <.001).

Reliability and analysis of gender differences

Given that Model 4 showed the best goodness of fit for the full version of the SUS (10 items), reliability analyses are reported for the SUS as a one-factor scale. Likewise, reliability was also analyzed for the short version with the best goodness fit, which only includes positively worded items (Model 5). Cronbach's alpha indices are presented in Table 5. These indices are reported for the total sample and also for females and males. Reliability coefficients for the categories "other" and "prefer do not specify" were not reported due to the low number of participants. Evidence of reliability was obtained for the total sample and across gender (Cronbach alphas ≥ 0.73) Gender differences in the total scores of the SUS were not significant for the full (t (1309) = .56, p= .58) or for the

246 short version with 5 items (t (1309) = .76, p= .45). The correlation between the total
247 scores of the SUS and age was not significant (r=-.03, p=0.26).
248
249 **Discussion**
250 The purpose of the present study was to adapt the System Usability Scale into Spanish
251 and examine its psychometric properties in a sample of Spaniards for the full 10-item
252 version testing different models provided by previous research, as well as investigating
253 the effect of negative items on the internal structure of the SUS. In addition, two short
254 forms consisting of either positive or negative items were analyzed.
255 Based on the CFA analyses performed in this study for the full version of the SUS,
256 Model 4 showed the best fit. This model included a general usability factor and a
257 negative method factor following the CTCM framework. These results suggested that
258 there is a source of method error associated with using negatively worded items in the
259 SUS, which reduce the SUS's construct validity when tested as a one-factor model (i.e.,
260 Model 1). The results obtained in this study are in line with previous studies that
261 suggest that the unidimensionality of a scale can be jeopardized by secondary sources of
262 variance and question the use of direct and reverse items for reducing the acquiescence
263 response bias (Salazar, 2015; Schriesheim & Hill, 1981; Suárez-Alvarez et al., 2018).
264 As Lewis & Sauro (2017) proposed, the two-factor solution based on item tone has little
265 practical or theoretical interest in assessing perceived usability. This study supports
266 previous studies on the SUS with the idea that the construct validity of the SUS was
267 obscured by the tone of the item (Borsci et al., 2009; Lewis, 2018; Lewis & Sauro,
268 2009), providing an empirical basis for explaining the effect of negatively worded
269 items. Therefore, treating the SUS as a unidimensional scale is recommended instead of
270 bidimensional when interpreting the participant’s scores.

Regarding reliability, the full 10-item version of the SUS showed acceptable Cronbach's alpha and Omega values in the overall sample and also across gender groups (>.73), so the full version of the SUS showed to be a reliable measure to be used with the Spanish population. It is noteworthy that our study did not find any gender or age differences in the SUS scores. These findings are in contrast to Bangor et al. (2008), who observed a slightly significant correlation with age. However, it should be considered that this study was conducted 15 years ago. Therefore, the differences in our findings may be attributed to changes in the profile of elderly users during this time. Additionally, two confirmatory analyses were conducted in this study, either with the 5 positively worded items or with the 5 negatively worded items of the scale, in order to obtain a short and valid usability measure without tone item effects. Although both models showed acceptable fit indices, the first showed better goodness fit indexes. In addition, the short and positively worded version of the SUS showed evidence of reliability (Cronbach's alpha = .77), so using this short version of the SUS is recommended for future studies.

The findings obtained in this study have important implications. First of all, this work studied the psychometric properties of a Spanish version of the SUS that were not examined in previous literature, such as the internal structure. In addition, this study includes an examination of method effects associated with using negatively worded items, which supports previous explanations regarding the effect of the item's tone on the SUS structure. Lastly, this study offers a short version with acceptable validity and reliability indexes and eliminates the problems arising from using positive and negative items. Having a short version of the SUS may be an advantage for those studies in which the SUS is part of a larger battery of questionnaires, and the use of brief questionnaires is preferred. However, caution should be exercised when utilizing the

short version in large batteries that evaluate usability electronically. In such cases, employing a mixed-tone version of the SUS or using control questions in the battery can help reduce response bias by compelling participants to carefully read the questions before providing responses.

In addition to these strengths, this study also has limitations that should be considered. First, this study did not include an analysis of concurrent validity with other usability measures. In this regard, the literature supports the fact that if a construct is defined and measured consistently using the same method across different cultures (using item and scalar equivalence by the nomological network method and by examination of internal structure congruence), the resulting measurement tool will have equivalent construct operationalization across those cultures (Hui & Triandis, 1985). However, Hilton et al. (2002) highlighted that the relevance of a specific concept could not be assumed to be universal across cultures. A word-for-word translation from English to another language is insufficient to address the linguistic and cultural disparities that exist. This statement underscores the importance of acknowledging cultural and linguistic differences when translating concepts from one language to another. One should not assume that concepts and terminology from one language hold the same significance in another language and culture. To achieve a suitable and effective translation, it is imperative to consider the diverse ways of thinking, feeling, and communicating that exist within each culture. For this reason, although our study confirms the SUS structure, it would be advisable to include other usability measures in future studies to analyze concurrent validity.

Secondly, this study did not perform a test-retest feasibility analysis, which is usually essential to examine the stability of the measurement. However, from the theoretical point of view of the usability construct, we believe it may not be appropriate to conduct a usability retest because experience moderates the perceived ease of use (Castañeda et

al., 2007). Thus, a retest measurement of the SUS will always obtain a better usability result due to the learning effect. In addition, it would be recommendable that future studies test the validity and reliability of the Spanish version of the SUS using the 10-item version with negative items converted to positive, which has provided evidence of its validity and reliability in previous studies (Kortum et al., 2021; Sauro & Lewis, 2011). Also, it is recommended to compare the short form of the SUS with other short usability scales, such as the UMUX-LITE or UX-Lite, which have shown correspondence with the SUS scores (Lewis, Utesch et al., 2015; Lewis & Sauro (2021). In conclusion, the present study demonstrated that the Spanish version of the SUS showed evidence of reliability and validity and evidenced that the internal structure of this scale is obscured by method errors. In addition, this study proposes to use a short version of the SUS comprised of the positively worded items, recommending its use as a brief measure of usability. This short version of the SUS could overcome the problems of the full version of the SUS associated with the tone of the items and has demonstrated to be a valid and reliable measure for assessing perceived usability with the Spanish population.

Declaration of interest

No potential conflict of interest was reported by the authors.

References

AlGhannam, B. A., Albustan, S. A., Al-Hassan, A. A., & Albustan, L. A. (2017). Towards a Standard Arabic System Usability Scale: Psychometric Evaluation using Communication Disorder App. *Https://Doi.Org/10.1080/10447318.2017.1388099*, *34*(9), 799–804. https://doi.org/10.1080/10447318.2017.1388099

Baharuddin, R., Singh, D., & Razali, R. (2013). Usability dimensions for mobile applications-a review. *Research Journal of Applied Sciences, Engineering and Technology*, *5*(6), 2225–2231. https://doi.org/10.19026/rjaset.5.4776

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *24*(June 2015), 574–594. https://doi.org/10.1080/10447310802205776

Blažica, B., & Lewis, J. R. (2015). A Slovene Translation of the System Usability Scale: The SUS-SI. *Http://Dx.Doi.Org/10.1080/10447318.2014.986634*, *31*(2), 112–117. https://doi.org/10.1080/10447318.2014.986634

Borkowska, A., & Jach, K. (2017). Pre-testing of Polish Translation of System Usability Scale (SUS). In Z. Borzenski, L., Grzech, A., Swiatek, J., Wilimowska (Ed.), *Advances in Intelligent Systems and Computing* (Vol. 521, pp. 143–153). Springer Verlag. https://doi.org/10.1007/978-3-319-46583-8_12/TABLES/7

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience. *Https://Doi.Org/10.1080/10447318.2015.1064648*, *31*(8), 484–495. https://doi.org/10.1080/10447318.2015.1064648

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale: a test of alternative measurement models. *Cognitive Processing*, *10*(3), 193–197. https://doi.org/10.1007/s10339-009-0268-9

Borsci, S., Federici, S., Mele, M. L., & Conti, M. (2015). Short scales of satisfaction assessment: A proxy to involve disabled users in the usability testing of websites. In Masaaki Kurosu (Ed.), *17th International Conference on Human-Computer Interaction* (Vol. 9171, pp. 35–42). Springer Verlag. https://doi.org/10.1007/978-3-319-21006-3_4/TABLES/4

Brislin, R. W. (1976). Comparative research methodology: Cross-cultural studies. *International Journal of Psychology*, *11*(3), 215–229. https://doi.org/https://doi.org/10.1080/00207597608247359

Brooke, J. (1996). SUS: A quick and dirty usability scale. *Usability Evaluation in Industry*, 189–194.

Castañeda, J. A., Muñoz-Leiva, F., & Luque, T. (2007). Web Acceptance Model (WAM): Moderating effects of user experience. *Information & Management*, *44*(4), 384–396. https://doi.org/10.1016/J.IM.2007.02.003

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. https://doi.org/10.5681/hpp.2014.011

DiStefano, C., & Motl, R. W. (2006). Further Investigating Method Effects Associated With Negatively Worded Items on Self-Report Surveys. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004

Guadagnoli, E., & Velicer, W. F. (1988). Relation of Sample Size to the Stability of Component Patterns. *Psychological Bulletin*, *103*(2), 265–275. https://doi.org/10.1037/0033-2909.103.2.265

Hilton, A., & Skrutkowski, M. (2002). Translating Instruments Into Other Languages: Development and Testing Processes. *Cancer Nursing*, *25*(1), 1-7.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A Review and Comparison of Strategies. *Journal of Cross-Cultural Psychology*, *16*(2), 131–152. https://doi.org/10.1177/0022002185016002001

Hvidt, J. C. S., Christensen, L. F., Sibbersen, C., Helweg-Jørgensen, S., Hansen, J. P., & Lichtenstein, M. B. (2020). Translation and Validation of the System Usability Scale in a Danish Mental Health Setting Using Digital Technologies in Treatment Interventions. *International Journal of Human–Computer Interaction*, *36*(8), 709–716. https://doi.org/10.1080/10447318.2019.1680922

International Organization for Standardization. (2018). *ISO 9241-11:2018(en), Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts*. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

Kortum, P., Acemyan, C. Z., & Oswald, F. L. (2021). Is It Time to Go Positive? Assessing the Positively Worded System Usability Scale (SUS). *Human Factors*, *63*(6), 987–998. https://doi.org/10.1177/0018720819881556

Kortum, P., & Sorber, M. (2015). Measuring the Usability of Mobile Applications for Phones and Tablets. *International Journal of Human–Computer Interaction*, *31*(8), 518–529. https://doi.org/10.1080/10447318.2015.1064658

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. https://doi.org/10.1080/10447319509526110

Lewis, J. R. (2018). The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction*, *34*(7), 577–590. https://doi.org/10.1080/10447318.2018.1455307

Lewis, J. R., Brown, J., & Mayes, D. K. (2015). Psychometric Evaluation of the EMO and the SUS in the Context of a Large-Sample Unmoderated Usability Study. *International Journal of Human–Computer Interaction*, *31*(8), 545–553. https://doi.org/10.1080/10447318.2015.1064665

Lewis, J. R., & Sauro, J. (2017). Revisiting the Factor Structure of the System Usability Scale. *Journal of Usability Studies*, *12*(4), 183–192.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In M. Kurosu (Ed.) (Ed.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 5619 LNCS* (pp. 94–103). Springer Verlag. https://doi.org/10.1007/978-3-642-02806-9_12

Lewis, J. R., & Sauro, J. (October 19, 2012). *How to Estimate SUS Using the UX-Lite.* Measuring U. https://measuringu.com/how-to-estimate-sus-with-ux-lite/

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human–Computer Interaction*, *31*(8), 496–505. https://doi.org/10.1080/10447318.2015.1064654

Liew, M. S., Zhang, J., See, J., & Ong, Y. L. (2019). Usability Challenges for Health and Wellness Mobile Apps: Mixed-Methods Study Among mHealth Experts and Consumers. *JMIR MHealth and UHealth*, *7*(1), e12160. https://doi.org/10.2196/12160

Maramba, I., Chatterjee, A., & Newman, C. (2019). Methods of usability testing in the development of eHealth applications: A scoping review. *International Journal of Medical Informatics*, *126*(November 2018), 95–104. https://doi.org/10.1016/j.ijmedinf.2019.03.018

Raita, E., & Oulasvirta, A. (2011). Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with Computers*, *23*(4), 363–371. https://doi.org/10.1016/j.intcom.2011.04.002

Rosa, A. F., Silva, A., Rocha, N. P., Queirós, A., & Martins, A. I. (2015). European Portuguese Validation of the System Usability Scale (SUS). *Procedia Computer Science*, *67*(Dsai), 293–300. https://doi.org/10.1016/j.procs.2015.09.273

Salazar, M. S. (2015). El dilema de combinar ítems positivos y negativos en escalas. *Psicothema*, *27*(2), 192–199. https://doi.org/10.7334/psicothema2014.266

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? *Conference on Human Factors in Computing Systems - Proceedings*, 2215–2223. https://doi.org/10.1145/1978942.1979266

Schriesheim, C. A., & Hill, K. D. (1981). Controlling Acquiescence Response Bias by Item Reversals: The Effect on Questionnaire Validity. *Educational and Psychological Measurement*, *41*(4), 1101–1114. https://doi.org/10.1177/001316448104100420

Sevilla-Gonzalez, M. D. R., Moreno Loaeza, L., Lazaro-Carrera, L. S., Bourguet Ramirez, B., Vázquez Rodríguez, A., Peralta-Pedrero, M. L., & Almeda-Valdes, P. (2020). Spanish Version of the System Usability Scale for the Assessment of Electronic Tools: Development and Validation. *JMIR Human Factors*, *7*(4), e21161. https://doi.org/10.2196/21161

Suárez-Alvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in likert scales: A questionable practice. *Psicothema*, *30*(2), 149–158. https://doi.org/10.7334/psicothema2018.33

Weston, R., & Gore Jr., P. A. (2006). A Brief Guide to Structural Equation Modeling. In *The Counseling Psychologist* (Vol. 34, Issue 5, pp. 719–751). Sage Publications. https://doi.org/10.1177/0011000006286345

Wind, T. R., Rijkeboer, M., Andersson, G., & Riper, H. (2020). The COVID-19 pandemic: The 'black swan' for mental health care and a turning point for e-health. *Internet Interventions*, *20*(March), 100317. https://doi.org/10.1016/j.invent.2020.100317

Wolgast, M. (2014). What does the acceptance and action questionnaire (AAQ-II) really measure? *Behavior Therapy*, *45*(6), 831–839. https://doi.org/10.1016/j.beth.2014.07.002

World Health Organization. (2019). The WHO Special Initiative for Mental Health (2019-2023): Universal Health Coverage for Mental Health. In *The WHO special initiative for mental health (2019-2023)*. https://apps.who.int/iris/bitstream/handle/10665/310981/WHO-MSD-19.1-eng.pdf?ua=1

Table 1. Original and Spanish items of the System Usability Scale

English version (Brooke, 1996)	Spanish adaptation
1. I think that I would like to use this system frequently	1. Creo que me gustaría usar este sistema frecuentemente
2. I found the system unnecessarily complex	2. Encontré el sistema innecesariamente complejo.
3. I thought the system was easy to use	3. Pensé que el sistema era 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 personal técnico para poder usar este sistema
5. I found the various functions in this system were well integrated	5. Encontré que las diversas funciones de este sistema estaban bien integradas
6. I thought there was too much inconsistency in this system	6. Pensé que había demasiada inconsistencia en este sistema
7. I would imagine that most people would learn to use this system very quickly	7. Imagino que la mayoría de las personas podrían aprender a usar este sistema muy rápidamente
8. I found the system very cumbersome to use	8. Encontré el sistema muy difícil de usar
9. I felt very confident using the system	9. Me sentí muy seguro usando el sistema
10. I needed to learn a lot of things before I could get going with this system	10. Necesité aprender muchas cosas antes de poder empezar a usar este sistema

Table 2. Descriptive analysis of items in the Spanish adaptation of the System Usability Scale

	Mean	SD	Skewness	Kurtosis
Item 1	2.81	1.07	-0.68	-0.08
Item 2	3.47	0.90	-1.94	3.59
Item 3	3.28	1.11	-1.67	2.03
Item 4	3.66	0.80	-2.91	8.84
Item 5	3.31	0.94	-1.62	2.69
Item 6	3.46	0.88	-1.84	3.24
Item 7	3.47	0.91	-2.08	4.34
Item 8	3.55	1.01	-2.57	5.73
Item 9	3.39	0.96	-1.78	2.98
Item 10	3.69	0.74	-2.94	9.44

Note: Negatively worded items (2,4,6,8 &10) were reversed

Table 3. Fit indices obtained from confirmatory analyses

	Chi	*p*	RMSEA	RMSEA 90% CI	CFI	TLI
Model 1	959.91	<.001	*.119*	.112, .125	.625	.517
Model 2	266.995	<.001	*.062*	.055, .069	.901	.870
Model 3	775.094	<.001	*.109*	.102, .115	.695	.597
Model 4	198.931	<.001	*.055*	.047, .062	.932	.898
Model 5	44.325	<.001	*.057*	.041, .075	.973	.946
Model 6	459.227	<.001	*.068*	.052, .086	.931	.863

Table 4. Standardized factor loadings for Confirmatory Factor Analyses

	Model 1 One-Factor	Model 2 Two-Factor		Model 3 Two-Factor		Model 4 Two-Factor		Model 5 One-Factor	Model 6 One-Factor
	General Usability	Positive usability	Negative usability	Usability	Learnability	General Usability	Negative Method	Positive short-form	Negative Short-form
Item 1	.514	.534		.525		.544		.541	
Item 3	.516	.528		.517		.518		.528	
Item 5	.756	.782		.772		.798		.783	
Item 7	.637	.650		.636		.631		.646	
Item 9	.694	.709		.424		.702		.707	
Item 2	.393		.621	.374		.309	.511		.597
Item 4	.302		.580		.692	.194	.565		.596
Item 6	.442		.633	.424		.367	.489		.593
Item 8	.163		.388	.144		.074	.422		.412
Item 10	.266		.600		.588	.144	.644		.637
		r= .409		*r*= *.389*		*r*= 0			

Note: The correlation between the General Usability Factor and the Negative method factor was fixed at 0.

Table 5. Reliability indexes for the full and the short positive form of the System Usability Scale

	Full form of the System Usability Scale		Short form of the System Usability Scale	
	Mean total score (SD)	**Cronbach's Alpha (95% CI)**	**Mean total score (SD)**	**Cronbach's Alpha (95% CI)**
Total	85.20 (13.12)	.76 (.74, .78)	81.26 (20.00)	.77 (.75, .79)
Females	85.49 (12.94)	.73 (.70, .76)	81.73 (18.19)	.75 (.72, .78)
Males	85.09 (13.07)	.77 (.75, .80)	80.98 (17.68)	.79 (.76, .81)

Note: Mean total scores for the Short Form of the System Usability Scale were calculated by multiplying the sum of the scores by 5, instead of 2.5 as Brooke (1996) proposes for the original System Usability Scale. Thus, the scores for the proposed short version have the same range as the original (i.e., from 0 to 100).

Biographies

Dr Diana Castilla is a psychologist specialized in the development of eHealth applications. Her main field is Cognitive Ergonomics, with a special interest in Usability and User Experience with ICTs. She leads the research group "SAEM-TIC. Mental Health Technologies" at University of Valencia (Spain).

Dr Irene Jaén is researcher at the Universitat Jaume I (Spain). Her main line of research is emotion regulation, with a special interest in the application of Information and Communication Technologies to the field of health and clinical psychology.

Dr Carlos Suso (susor@uji.es) is professor in the Universitat Jaume I (Spain). His main line of research is chronic pain and new technologies, with special interest in ecological momentary interventions and machine learning.

Dr Gemma Garcia-Soriano is Professor at University of Valencia (Spain), and senior researcher of I'TOC research group at University of Valencia. Her main field of work is focused on mental health applications and Obsessive-compulsive disorder.

Dr Irene Zaragoza is engineer in Computer Science with interest in the development of different eHealth applications. She is a senior researcher of "SAEM-TIC. Mental Health Technologies" group at University of Valencia (Spain).

Dr Juani Breton-Lopez is professor at Universitat Jaume I (Spain). Her research interests include health and clinical psychology, and the application of Information and Communication Technologies to Psychology. She leads the Emotional Disorders Clinic at UJI, which offers psychological treatments to the community and training for clinical psychologists.

Dr Adriana Mira is professor at University of Valencia (Spain), and researcher in Labpsitec. Her research interests include health and clinical psychology and the application of Information and Communication Technologies to Psychology, especially in emotional disorders.

Dr Amanda Díaz-García is professor at University of Zaragoza (Spain), and researcher of IPES research group. Her research interests include health and clinical psychology, and intervention of emotional disorders through the Unified Protocol and the application of Information and Communication Technologies to Psychology

Dr Azucena Garcia-Palacios is a clinical psychologist with an interest in health and clinical applications based on Information and Communication Technologies; she is professor in Psychopathology and director of Labpsitec, Jaume I University (Spain).