The Turkish Online Journal of Design, Art and Communication - TOJDAC
ISSN: 2146-5193, October 2021 Volume 11 Issue 4, p.1328-1347

ADAPTING SUPR-Q INTO TURKISH FOR ASSESSING USER EXPERIENCE IN WEB AND MOBILE SERVICES

Mehmet İlker BERKMAN
Bahçeşehir University, Turkey
ilker.berkman@comm.bau.edu.tr
https://orcid.org/0000-0002-2340-9373

Şafak ŞAHİN
Bahçeşehir University, Turkey
safak.sahin@comm.bau.edu.tr
https://orcid.org/0000-0003-2459-8476

Atıf	Berkman, M. İ. ve Şahin, Ş. (2021). ADAPTING SUPR-Q INTO TURKISH FOR ASSESSING USER EXPERIENCE IN WEB AND MOBILE SERVICES. The Turkish Online Journal of Design Art and Communication, 11 (4), 1328-1347.

ABSTRACT

SUPR-Q (- Standardized User Experience Percentile Rank Questionnaire) is a usability scale that is highly suitable for assessing websites and mobile applications which have transactional capabilities, such as shopping and e-commerce services. The Turkish adaptation of 8-item SUPR-Q is evaluated through 120 responses collected from users of Turkish shopping websites. The evidence for the 4-factor model of Appearance, Loyalty, Usability and Trust was partly verified based on the dataset, using Partial Least Squares based Confirmatory Factor Analysis. For this reason, we proposed a two-factor measurement model of Ergonomics and Credibility dimensions, considering the ongoing discussions on the close relationship between hedonic and pragmatic quality of digital artefacts, and the strong causal relationship between trust and loyalty given in the literature. Both four-factor and two-factor models, showed a good level of internal consistency and they are sensitive to the differences between web sites as well as mobile apps. Both scales' scores have correlated with the relevant subscales of Turkish Computer System Usability Questionnaire. Our results suggest that our translated SUPR-Q items are capable of distinguishing between the sites as well as mobile shopping apps, and more sensitive than the UMUX applied in English, while its sensitivity is comparable with the Turkish T-CSUQ-SV. Considering that SUPR-Q has fewer items, it can be a good choice for researchers who need a brief tool for assessing web sites in terms of subjective user experience.

Keywords: *web usability, hedonic quality, pragmatic quality, PLS-CFA.*

WEB VE MOBİL SERVİSLERİN KULLANILABİLİRLİĞİNİ ÖLÇMEK ÜZERE SUPR-Q ÖLÇEĞİNİN TÜRKÇE'YE ADAPTASYONU

ÖZ

SUPR-Q (Standardized User Experience Percentile Rank Questionnaire) alışveriş ve e-ticaret hizmeti veren siteler gibi, kullanıcıdan gelen verileri işlemeye dayalı web siteleri ve mobil uygulamaları değerlendirmek için son derece uygun bir kullanılabilirlik ölçeğidir. 8 maddelik SUPR-Q'nun Türkçe uyarlaması, Türk alışveriş siteleri kullanıcılarından toplanan 120 yanıt üzerinden değerlendirilmiştir. 4 faktörlü Görünüm, Sadakat, Kullanılabilirlik ve Güven modeli, Kısmi En Küçük Karelere dayalı Doğrulayıcı Faktör Analizi kullanılarak veri kümesi bazında kısmen doğrulanmıştır. Bu nedenle, dijital

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

ürünlerin hedonik ve pragmatik kalitesi arasındaki yakın ilişki hakkında devam eden tartışmaları ve literatürde güven ile sadakat arasındaki kurulan güçlü nedensel ilişkiyi göz önünde bulundurarak, Ergonomi ve Güvenilirlik boyutlarından oluşan bir ölçüm modeli önerilmiştir. Hem dört faktörlü hem de iki faktörlü modeller, iyi düzeyde bir iç tutarlılık göstermektedir ve web sitelerinin yanı sıra mobil uygulamalar arasındaki farklılıklara karşı da hassastır. Her iki ölçeğin alt boyut puanları Türkçe Bilgisayar Sistemleri Kullanılabilirlik Anketi'nin (T-CSUQ) alt ölçekleri ile korelasyon içerisindedir. Sonuçlar, Türkçe'ye çevrilmiş SUPR-Q öğelerimizin siteler veya uygulamalar arasında ayrım yapabildiğini ve İngilizce olarak uygulanan UMUX'tan daha hassas olduğunu, duyarlılığının ise Türkçe T-CSUQ-SV ile karşılaştırılabilir olduğunu göstermiştir. SUPR-Q'nun daha az öğeye sahip olduğu düşünüldüğünde, web sitelerini öznel kullanıcı deneyimi açısından değerlendirmek için kısa bir araca ihtiyaç duyan araştırmacılar için önerilmektedir.

Anahtar kelimeler: *kullanılabilirlik, hedonik kalite, pragmatik kalite, PLS-CFA.*

INTRODUCTION

The concept of usability, which has been defined through three components; efficiency, effectiveness, and satisfaction (ISO9211 – 11, 2010) is operationalized through users' actions and attitudes. The measures based on actions, such as time on task or error rate are suggested as indicators of usability, which are depending on users' actions, for efficiency and effectiveness, respectively. Furthermore, biometric measures, such as facial movements detected through electromyography, are employed as indicators of satisfaction. On the other hand, self-report measures are also widely used to assess perceived usability, in the form of usability questionnaires. These questionnaires, which are applied following an experience with the system, aim to assess users' attitudes towards this experience. Although they are not capable of detecting the issues detrimental to experience, subjective usability questionnaires are valuable tools to assess the adequacy of computer systems in providing an acceptable user interaction.

SUPR-Q - Standardized User Experience Percentile Rank Questionnaire (Sauro, 2015) is a research instrument which can be used to generate reliable scores in benchmarking websites. While it is developed to considering all kinds of web sites with many different purposes, it is highly suitable for assessing websites and mobile applications which have transactional capabilities, such as shopping and e-commerce services, as it is designed to be "technology agnostic", with items that are appropriate to assess users' attitudes for different platforms or devices.

Online shopping has already been a growing market in Turkey, before the CoviD-19 pandemic, as the number of e-shoppers were increase in percentage, the share of e-retail is expanding within the total retail sales and the number of credit card owners are on the rise (Göl et al., 2019). As the pandemic limits the physical mobility of the people, even the simplest fast-moving consumer goods are bought online, leading to surplus in e-commerce, as reported by Yılmaz & Bayram, (2020). It can be projected that the post-pandemic consumers are likely to "prefer digital over physical in the future" and "may also be more accepting of further technological innovation in the delivery of consumption experiences" (Barnes, 2020). For this reason, we believe that researchers and the e-commerce industry would require tools to assess online shoppers' attitudes towards online platforms where the e-shoppers purchase through. Instead of developing a scale from scratch, we decided to adapt an existing standardized research tool, since the online digital marketing and content distribution is a global industry. The local establishments and global companies exit together in almost every country, while the local players also evolve into global entities through innovation.

Although there are several alternatives that are thoroughly mentioned in "Related Studies", which are specifically developed for measuring user attitudes on web sites, we have chosen SUPR-Q since it is developed through a broader conceptual framework compared to its predecessors. It is generalizable, which means that its item phrasing is generic enough so the same items can be used to evaluate different websites. Being multidimensional, it encompasses "the most well-defined factors for measuring website quality as uncovered in the review of existing instruments", and brief to be responded quickly by

participants, while it also fits easily to the mobile devices' screens. Additionally, it is designed to be used with a normative database for knowing where a website scores relative to its peers.

Despite our study being limited with online shopping services, it should be remembered that SUPR-Q is designed to be employed to assess other types of websites and applications, such as social media platforms, content distribution services or non-profit websites. We limited our scope with the online shopping services as a starting point for localizing SUPR-Q through psychometric evaluation. This should also be taken as an early attempt to form a normative database, which "will provide additional information to researchers who administer the instrument in isolation", as suggested by (Sauro, 2015).

RELATED STUDIES

Web Site Usability Scales

Multidimensional scales, which are standardized through psychometric methodologies, have been widely used to assess the users' subjective attitude related to their experience of using a computer system since the late 80's, starting with the QUIS (Questionnaire for User Interface Satisfaction). However, it took a decade until (Kirakowski & Cierlik, 1998) published on their WAMMI (Website analysis and measurement inventory), which is the earliest successful attempt for a scale developed for measuring perceived website usability. Using 20 items in the final version to produce scores on five factors; attractiveness, efficiency, controllability, helpfulness, and learnability; WAMMI was developed based on opinions of "designers, users, and webmasters about typical positive and negative experiences encountered when visiting and using websites". It should be remembered that these experts formed their opinions through their experiences of information-oriented websites of the late 90's. As a result, the scale contains items like "All the material is written in a way that is easy to understand." or "There is a lot of information on this web site.", which are querying some aspects of the web sites that are currently "taken for granted" by the users.

WEBQUAL (Website Quality) (Loiacono et al., 2002) was also a proprietary research tool like its predecessor WAMMI, with its 36 items for assessing 12 dimensions. In a later study (Loiacono et al., 2007), they criticize an identically named research tool (Barnes & Vidgen, 2000) for being poorly developed with a small sample size.

WQ (Website quality) (Aladwani & Palvia, 2002) is another web usability scale of early 2000's, employing 25 items to evaluate websites through four factors: Appearance, specific content, content quality, technical adequacy. Items in the content quality dimension queries the user on aspects of content such as accuracy, completeness, or usefulness, while the specific content is about availability of contact information or customer policies. Technical adequacy is concerned with issues such as availability of search or speed of page loads. For appearance, users are queried on their level of agreement for proper use of fonts and colors as well as site's attractiveness. Similar to WAMMI, the scale is suitable to assess information-oriented web sites rather than transactional sites.

(Wang & Senecal, 2008) proposed the WU (Web Usability) scale with 8 items for Ease of Navigation, Speed and Interactivity, which are first order constructs of the underlying construct of Web Usability. Although they used a relatively larger sample size in their study, the participants were limited to undergraduate students, who evaluated a single website.

In addition to these scales, SUS (System Usability Scale) (Brooke, 1996) and UMUX (Usability Metric for User Experience) (Finstad, 2010) are often used in subjective website usability evaluations, because of their technology-agnostic structure. SUS is available in Turkish (Demirkol and Şeneler, 2018) and it has been successfully used for evaluation of web sites (Demirkol et al., 2020). However, both SUS and UMUX are unidimensional and focus on ease of use, efficiency, effectiveness, satisfaction and interface ergonomics. They do not address the issues related with users' motivations for using the system. Within the context of web, users' decision to use a web site not solely related with its interface and interaction design, but also the reliability of the organization and services provided by the website is highly important for the perceived quality of the online service.

Criticizing the previous scales for not being generalizable and not being brief to be responded quickly, Sauro, (2015) proposed SUPR-Q, which is designed to be also employed as a benchmark instrument.

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

Development and psychometrics of SUPR-Q

SUPR-Q is developed into its final 8-item version through three studies. In all the studies, participants are derived from the US population and websites assessed by the participants are mostly originated from the United States.

The first study was started with the construction of a pool of 33 items, which are selected from the literature corresponding to the four constructs of usability, loyalty, trust, and appearance. These constructs are determined based on previous research, regarding their ability to describe website quality through a "technology agnostic perspective". Except the item "How likely are you to recommend the website to a friend?", which is scaled from 0 to 10 as of Net Promoter Score (Reichheld, 2003), all items are responded through a 5-point Likert Scale of 1 for "strongly agree" to 5 for "strongly disagree". Based on the data collected from 91 participants who reported to use 51 different shopping sites, 7 items were eliminated through a series of factor analysis. Seeking for a frugal research instrument, "items with item-total correlations less than .5 and with cross-loadings on multiple factors within .2 were deleted", leaving 13 items, with four or three items per dimension. Cronbach's alpha values of each dimension and the overall items indicate an adequate level of reliability, which are given in detail on Table 1. All factor scores and the overall score positively correlated with SUS (system usability scale) (Brooke, 1996) between .59 to .75, indicating the convergent validity of the scale.

Table 1. Cronbach's alpha values indicating reliability obtained in different studies. Sauro (2015)

	Appearance	Loyalty	Usability	Trust	Overall
Study 1	.83	.83	.87	.93	.87
Study 2	.82	.63	.94	.89	.91
Study 3	.78	.64	.88	.85	.86

On Study 2, (Sauro, 2015) collected data from 484 participants who attempted one predefined task on one of predefined 40 websites. Sites were selected to represent a spectrum of usability, from sites that are anecdotally identified to have poor usability by its users to most visited sites from a wide range of industries and were not limited to shopping sites, while the given tasks were specifically described according to each site's purpose of use. Results revealed that 13 items still fit a four-factor structure reasonably well with most loadings above .6., except two items which are dropped to further reduce the scale length. The overall score for retaining 11 items correlated strongly with WAMMI (Kirakowski & Cierlik, 1998) and SUS (Brooke, 1996), and each subscale had a moderate to high correlation with the scales they converged with. As it can be followed on the Cronbach's alpha values given on Table 1, each subscale and the overall scale is adequately reliable. A series of one-way ANOVAs revealed that SUPR-Q is also capable of discriminating between the websites., i.e., providing significantly different scores on some of the web sites evaluated in the study. Its discrimination capability for its overall score is similar to WAMMI and SUS scores obtained from the same participants. For the usability, trust and loyalty factors, it is also capable of discriminating the websites, but appearance dimension did not reveal significantly different scores between the sites evaluated in the study.

3,891 responses for 51 sites were assessed in the third study, where each response is given by a different participant. Participants are representing a large variety in terms of age, gender (57% female), educational background, occupation, and prior experience with the web site they evaluated. Of the 11 items, three items were dropped to form the final version of SUPR-Q. The overall score of this final 8-item version reveals a relatively correlation with SUS, while the usability subscales has also a strong convergent validity with SUS, as the others showed moderate correlations. SUPR-Q also revealed the same discriminating power with SUS scores collected in the study, although it has two fewer items than the 10-item SUS. Cronbach's alpha values for study 3 can also be seen on Table 1.

SUPR-Q is frugally short, developed through a elaborative process involving a very large group of participants, revealing the same factor structure in different studies, being capable of producing discriminating scores, having an acceptable level of reliability even based on Cronbach's alpha which

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

is known as a conservative indicator and providing evidence for convergent validity through Pearson correlations with SUS and WAMMI.

For these reasons, we believed that SPR-Q is a good choice to be translated and localized through a psychometric evaluation involving a Turkish population sample, who evaluate their experiences on Turkish origin websites.

METHODOLOGY

The research methodology regarding translation and data collection is inspected and approved by the Ethics Board of the university where the authors are employed.

Translation

In the translation process, we checked the prior good practices of translation for other usability scales and followed the guidelines suggested for localization.

One of the authors is an expert in human-computer interaction while the other is on marketing and advertising. They are natively Turkish speakers but have been teaching in English for more than a decade. They translated the items independently and compared, discussed, and revised their translations through an online word processor, using audio chat. The revised items are back-translated into English by a licensed educator of English language. The original items, back translations and Turkish items were given to three independent evaluators who are also licensed English teachers. They inspected the match between the original and back-translation by scoring them 1 for a good match, 2 for partial consistency and 3 for a poor match. For the items that they scored as 2 or three, they were also asked to suggest a better translation by reviewing the translated item.

Based on the average of three evaluators, each of the items had an average score less than 1 or 1.6, indicating that at least two of the evaluators agreed on a good match between original and back translated items. However, the alternative translations suggested by evaluators are reviewed in an online session in which the authors and the back-translating language expert had participated. It is decided that there is no need to make any changes on the initial translations, given on Table 2.

Table 2. Original Items and their Turkish translations.

		Original item	Turkish translation
Usability	U1	The website is easy to use.	Bu web sitesini/uygulamayı kullanmak kolaydır.
	U2	It is easy to navigate within the website.	Bu web sitesi/uygulama içerisinde gezinmek kolaydır.
Trust	C1	I feel comfortable purchasing from the website.	Bu siteden/uygulamadan alışveriş yaparken kendimi rahat hissediyorum.
	C2	I feel confident conducting business on the website.	Bu site/uygulama ile iş yaparken kendimi güvende hissederim.
Loyalty	L1	How likely are you to recommend the website to a friend or colleague?	Bu siteyi/uygulamayı bir arkadaşınıza ya da meslektaşınıza ne ölçüde tavsiye edersiniz?
	L2	I will likely return to the website in the future.	Bu siteyi/uygulamayı ilerleyen zamanlarda büyük ihtimalle tekrar ziyaret ederim / kullanırım.
Appearance	A1	I find the website to be attractive.	Bu siteyi/uygulamayı albenili (çekici) buldum.
	A2	The website has a clean and simple presentation.	Bu web sitesinin/uygulamanın açık ve sade bir sunumu var.

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

Research Instruments and Data Collection

The survey is executed through the soscisurvey.de website (Leiner, 2021). In addition to translated SUPR-Q items, the survey form included the T-CSUQ-SV (Erdinç & Lewis, 2013), the Turkish version of Computer System Usability Questionnaire (Lewis, 1995) and the UMUX (Usability Metric for User Experience) (Finstad, 2010) items in English. Participants are asked to respond to UMUX items only if they think that they have fluency in English, otherwise skip the corresponding part. In addition to the scales, demographic questions on age and gender were included. The participants are asked to join the survey if they shopped online within the last month using a web site or a mobile app. At the beginning of the survey. They choose the site they shopped among the list of 13 popular shopping sites, including the "other" option. When "other" is selected, they are asked to type in the name of the site or the app on an auto-complete text field, with 426 preset options. If the site or app is not available in the given list, they may complete typing. Participants were also queried on when they used the site or app, the device and platform (desktop, laptop, tablet computer or mobile phone; web browser or app), their frequency of using the app or site, their frequency of online shopping (see Participants for details). The survey invitation is sent via social media and social messaging groups, within an easy sampling approach and embracing the snowball sampling method, participants are asked to send the survey to social messaging groups and share it on social media.

Data Analysis

Participants who stated that "they use the same shopping service for almost always" were excluded, since they may skew the results. We think that their lack of experience with other shopping services may lead them to misjudge their experience on the site or app they evaluated. Also, the responses that repeatedly give the same score to the SUPR-Q items were excluded from the sample. Participants who indicated that their age is below 18 are also excluded.

The factor structure of SUPR-Q is confirmed using PLS based CFA (confirmatory factor analysis) (Hair et al., 2017; Henseler et al., 2016), using the Smart-PLS version 3.3.3 (Ringle et al., 2015). Analysis was made based on responses given for mobile and desktop web sites, excluding the mobile applications, because SUPR-Q was originally developed for evaluating web sites. Compared with the prevalent variance-based algorithms, PLS methods provide a higher statistical power, and they have "almost no limiting assumptions regarding the model specifications and data" (Hair et al., 2011) such as sample size or distribution of data (Hair et al., 2017). Although there are several "golden rules" about the sample size required for covariance-based factor analysis methods (Schreiber, 2021), PLS based analysis "can be a very sensible methodological choice if sample size is restricted" (Reinartz et al., 2009).

The data collected from evaluators of mobile apps are used for comparison with data collected on website users through student t-tests, to explore the applicability of SUPR-Q for assessing user experience in mobile apps. Furthermore, the scale's capability of providing different scores for different apps, i.e., sensitivity, is explored through a series of ANOVA on three most evaluated services. Pearson correlations of SUPR-Q with factor and overall scores of UMUX and T-CSUQ-SV are inspected to assess the convergent validity.

Participants

All the analysis were conducted on a sample size of 290 participants. Their ages were varying between 18 to 71 (N=223; M=36.1; SD=11.7). As it can be followed on Table 3, many of the participants were female shoppers. Participants who evaluated mobile phone apps were more than the participants who evaluated a website. Even for web site users, the smartphones are preferred over other devices.

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

Table 3. Participants by gender and the devices and agents they used for shopping

	Female	Male	N/A	ALL
App with smartphone	102	50	18	170
App with tablet	-	-	-	-
APP TOTAL	**102**	**50**	**18**	170
Site with desktop	7	14	0	21
Site with laptop	16	10	4	30
Site with smartphone	47	12	9	68
Site with tablet	1	-	-	1
SITE TOTAL	**71**	**36**	**13**	120
ALL PARTICIPANTS	**173**	**86**	**31**	**290**

Participants have evaluated 27 distinct websites and 15 distinct mobile apps. Some of the evaluated apps and sites are managed by the same shopping service. The services with relatively large numbers of evaluations are detected for both app and site evaluations. Service names are not given explicitly, since the research does not aim to make a comparison with commercial entities. As given on Table 4, most of our participants evaluated their experience with the website or app to make a purchase from Service A. Of the 27 sites evaluated, only 35% of the data is collected regarding 23 of these sites. For the 15 apps evaluated, 18% of the data is collected for 11 of these apps. Rest of the data is based on the evaluations of four shopping services.

Table 4. Evaluations on sites and apps by services

	Site	N	%	App	N	%
Service A	1	40	33%	1	95	56%
Service B	1	19	16%	1	19	11%
Service C	1	15	13%	1	7	4%
Service D	1	3	3%	1	19	11%
with 2 to 7 evaluations	5	24	20%	6	25	15%
with only one evaluation	18	19	16%	5	5	3%
	27	120		15	170	

Most of the participants reflected on a shopping experience within the last week as shown on Table 5. Given the percentages, the ratio of users with different recency of experience are similarly distributed, except for the users who indicated that they shopped online within the last month earlier than three weeks ago.

Table 5. The recency of evaluated shopping experience

	Websites (N)	Apps (N)	Websites %	Apps %
Today	7	16	6%	9%
Yesterday	18	29	15%	17%
A few days ago	27	50	23%	29%
Last week	23	39	19%	23%
Two weeks ago	15	19	13%	11%
Three weeks ago	7	7	6%	4%
Earlier in last month	23	10	19%	6%

Comparing the frequencies of amount spent by the website and app shoppers, the ratio of shoppers who spent within the given limits are similar for purchases up to 1000 ₺. For purchases above this limit, our participants preferred to use websites rather than mobile apps as seen in Figure 1.

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

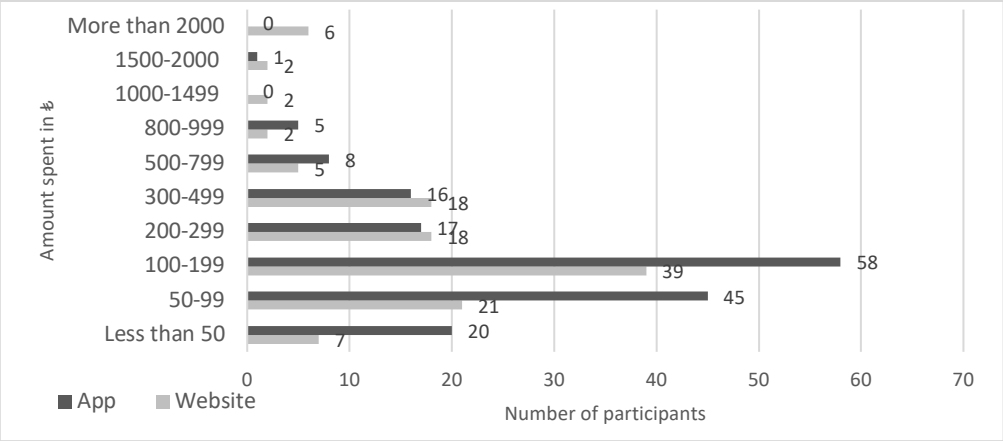

Figure 1. Number of participants based on amount (₺) spent in the evaluated shopping experience. $1 was approximately ₺8 during the time of the study.

Participants were also asked to rate the importance of several aspects related with the shopping experience through a 7-point Likert scale from 1 - not important to 7 - very important.

Table 6. Importance of aspects related with evaluated shopping experience.

	Websites		Apps		
	M	SD	M	SD	t-test
Availability of product comparison	4.98	1.81	5.51	1.50	t(224.7)=-2.59,p<.05
Free Shipping	5.48	1.95	5.97	1.42	t(288)=-2.46, p<.05
Correctness of product information	6.42	0.87	6.28	1.02	t(288)=1.23,p>.05
Security of personal data	6.31	1.14	6.29	1.17	t(261.41)=0.15, p>.05
Privacy of personal data	6.28	1.22	6.22	1.25	t(260.48)=0.35, p>.05
Reputation of site/app	6.06	1.27	6.22	0.98	t(213.66)=-1.2, p>.05
Pricing	6.01	1.16	6.05	1.18	t(259.96)=-0.32, p>.05
Product info given in detail	5.98	1.28	5.98	1.13	t(234.54)=0.05, p>.05
Ease of entering the payment information	5.95	1.33	6.14	0.99	t(288)=-1.36, p>.05
Return options	5.90	1.64	6.04	1.58	t(288)=-0.71, p>.05
Delivery time	5.89	1.36	5.99	1.22	t(239.1)=-0.66, p>.05
Clearness of product images	5.88	1.48	6.01	1.08	t(288)=-0.86, p>.05
Discount	5.88	1.51	5.86	1.43	t(247.17)=0.06, p>.05
Abundancy of choices	5.81	1.46	5.98	1.14	t(288)=-1.1, p>.05
Number of product images	5.65	1.49	5.85	1.30	t(233.54)=-1.17, p>.05
Ease of use of the system interface	5.53	1.44	5.76	1.15	t(288)=-1.48, p>.05
Promotions	5.49	1.63	5.69	1.44	t(235.75)=-1.09, p>.05
Availability of getting in touch with shopping service directors	5.38	1.75	5.52	1.48	t(227.65)=-0.68, p>.05
Availability of getting in touch with salesperson	5.34	1.85	5.51	1.57	t(229.51)=-0.79, p>.05
Simplicity of the app/site design	5.26	1.54	5.39	1.35	t(234.46)=-0.78, p>.05
Shipping Service (Company)	5.04	1.96	5.24	1.65	t(228.23)=-0.88, p>.05
Look and feel of system design	4.92	1.83	5.19	1.32	t(288)=-1.5, p>.05
Owner of the shopping service (person or entity)	4.65	1.91	4.91	1.71	t(238.64)=-1.17, p>.05
Payment on instalments	4.14	2.31	3.88	2.22	t(249.87)=0.96, p>.05

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

Given the results on Table 6, a series of t-tests revealed that there is not a significant difference between the website shoppers and mobile app shoppers regarding to importance of these aspects on the shopping experience they evaluated, except for "Availability of product comparison" and "Free shipping", which were significantly more important for mobile app shoppers.

The online shoppers who evaluated the web sites do not visit and use the site they evaluated as frequently as the app users, as given on Table 7. There are relatively more first-time users on web sites, and a large majority of app evaluators are frequent users. These findings are expected regarding the fact that installing an app on a personal mobile device requires more dedication than shopping on a website, which leads to more frequent use of the app.

It should be remembered that the data provided regarding our participants is not collected to represent the status of web shoppers in Turkey. It is given to report on the attributes of the sample in which we collected data to assess the Turkish translation of SUPRQ. In addition, data shows that web shoppers and app shoppers are not completely different groups, sharing similar shopping habits, spending similar amounts and being concerned about similar aspects of online shopping.

Table 7. Frequency of use for the evaluated website

	Websites		Apps	
	N	%	N	%
First time	9	7.5%	3	1.8%
Once	10	8.3%	7	4.1%
A few times	27	22.5%	24	14.1%
Occasionally	36	30.0%	58	34.1%
Frequently	38	31.7%	78	45.9%

RESULTS & DISCUSSION

Based on the data collected through 120 web shoppers, the psychometric qualities of the scale SUPR-Q were assessed for its internal consistency, unidimensionality of constructs, and discriminant validity.

Internal Consistency

Given in Table 8, the SUPR-Q subscales revealed an adequate reliability for confirmatory purposes, based on the composite reliability and rho_A measure. Since Cronbach's α is a conservative measure which can be taken as the lower bound for reliability, we suggest that the score of .63 for Loyalty can be ignored, considering that the factor depends on two reflective observed variables. For composite reliability being highly liberal, we also report the rho_A measure (Dijkstra & Henseler, 2015) that is known to be usually providing a score between composite reliability and Cronbach's alpha.

Table 8. Indicators of internal consistency

	Appearance	Trust	Loyalty	Usability
Cronbach's α	0.72	0.76	0.63	0.80
rho_A	0.81	0.79	0.75	0.80
Composite Reliability	0.87	0.89	0.84	0.91
AVE	0.77	0.80	0.72	0.83

Since there is not any composite reliability score exceeding .95, items are not redundant. Based on our dataset, the Turkish translation SUPR-Q is not proven to be reliable as the original scale, but still meets the criteria for an adequately reliable research instrument. You may see Table 1 to compare our results with the original studies by (Sauro, 2015). Given the AVE (Average Variance Extracted) values in Table

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

8 which are above .5 for each dimension, our data verified that these factors explain at least half the variance of their respective indicators.

Cross-loadings for convergent validity

When the loadings and cross-loadings of the items are checked (see Table 9), it is observed that the highest load of each item is on its intended factor, i.e., the factor that the item has loaded on the final stage of the original study. These loadings are above .708, indicating that they explain at least the half of variance, as the AVE does. On the other hand, an observed variable should not have a high loading on a latent variable other than its own.

Table 9. Cross-loadings of items on factors

	Appearance	Trust	Loyalty	Usability
A1 - attractive | albenili/çekici	**0.825**	0.333	0.444	0.534
A2 - clean and simple | açık ve sade	**0.930**	0.547	0.656	0.696
C1 - comfortable purchasing | alışveriş rahat	0.542	**0.923**	0.660	0.416
C2 - confident in business | iş yaparken güvende	0.372	**0.868**	0.577	0.274
L1 - Net Promoter Score	0.643	0.715	**0.924**	0.602
L2 - return in the future | gelecekte tekrar	0.414	0.408	**0.766**	0.342
U1 - easy to use | kullanması kolay	0.631	0.359	0.538	**0.914**
U2 - easy to navigate | gezinmesi kolay	0.668	0.361	0.528	**0.912**

The items are partially worded as a reminder of concepts. See Table 2 for full wording of items.

However, the A2 item of Appearance loads on to Usability moderately, while also the A1 has a relatively high load on the same factor. Besides, U1 and U2 of Usability are also moderately loading Appearance. Likewise, items of Loyalty and Trust are also moderately loading on each other's intended factor. While there is not a "rule of thumb" criteria for cross-loadings, items cross-loading highly to another factor may defy the discriminant validity.

Discriminant Validity

Although the recent studies (Hair et al., 2019; Henseler et al., 2015) suggest using the HTMT (hetero-trait-to-mono-trait) ratio as an indicator of discriminant validity instead of Fornell-Larcker criterion, we reported both in this study. According to Fornell-Larcker criterion, the square-root of AVE (given in bold on Table 10) of a factor should be higher than the correlations of the factor with other factors in the scale (given in italics on Table 10), as an evidence for discriminant validity. As it can be observed on Table 10, our data collected with the Turkish translation of SUPR-Q supports the Fornell-Larcker criterion, providing evidence on discriminant validity.

Table 10. Correlations between factors compared with AVE for each factor.

	Appearance	Trust	Loyalty	Usability
Appearance	**0.879**			
Trust	*0.521*	**0.896**		
Loyalty	*0.645*	*0.694*	**0.849**	
Usability	*0.711*	*0.394*	*0.584*	**0.913**

On the other hand, when explored through the HTMT ratios, discriminant validity of Turkish-translated SUPR-Q is not supported. Since the HTMT ratios (given on Table 11) are above the liberal threshold (Hair et al., 2017; Hair et al., 2019) of HTMT.90, we could not provide evidence that Usability is empirically distinct from Appearance, and Trust is from Loyalty. The HTMT ratio of Appearance to Loyalty is also above the conservative threshold of HTMT.85.

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

Table 11. HTMT (Hetero-trait to mono-trait) ratios

	Appearance	Trust	Loyalty	Usability
Trust	0.66			
Loyalty	0.89	0.94		
Usability	0.92	0.49	0.78	

Discussion on Psychometric Attributes of Turkish translation of SUPR-Q

Although the factor loadings of the items were strong and cross-loadings were very low in three studies reported by (Sauro, 2015), our dataset did not support the original factor structure of SUPR-Q.

First, we suspected that our data is highly skewed, since we asked our participants to evaluate the website that they recently used for purchasing goods, similarly with the Study 1 reported in the original study. As the site they evaluated is of their own choice, they might have strong positive bias on their evaluation. However, our data is moderately skewed; that is .9 for Appearance, .84 for Usability, .77 for Trust and .71 for Loyalty, for the scores calculated as means of manifest variables. It is known that PLS based factor analytic methods can tolerate skewness of the data.

Another point is that 33% of the evaluations are made for the website of the same service, and the other 32% of the data is evaluations on sites of three other services, while 35% is evaluations of 18 other websites. As the 65% of the data is collected for evaluation of four sites only, these sites may not have triggered different scores as of the 51 sites evaluated by the participants in Study 1. Unfortunately, the original study did not report the number of evaluations based per site.

Aside from the differences between the sample characteristics and evaluated sites, we think that the reason for factors not being discriminated is conceptual. Trust and Loyalty are highly related with each other, while appearance is closely affined with usability.

Many studies reported explored and reported correlation between the appearance of the websites and perceived usability (e.g. (Alharoon & Gillan, 2020; Hartmann et al., 2007; Lavie & Tractinsky, 2004; Lindgaard et al., 2011; Skulmowski et al., 2016; van Schaik & Ling, 2003)

According to Chang et al., (2014) , aesthetic formality and aesthetic appeal of a shopping website have a significant and positive effect on control, which is defined in three dimensions: behavioral (direct action on the environment), cognitive (the interpretation of events), decisional (having a choice among alternative courses of action). Of these three dimensions, factor loading of cognitive control was the highest in composition of the control construct. Clearly, the concept of control strongly corresponds to the usability within the context of interactive systems, while appearance is corresponding with aesthetic attributes. Surprisingly, their study did not reveal an effect of aesthetics on pleasure. Concluding on user ratings of shopping websites, (Hassenzahl & Monk, 2010) points out that the “expressive aesthetics”, stated as “creative”, “fascinating”, “original”, or “sophisticated” is overlapping with the hedonic quality, while the classic aesthetics is specified as “aesthetic” “pleasant” “clear” “clean” and “symmetric”. Thus, he claims that potential explanation would be to understand classic aesthetics as a form of “visual” usability (i.e., “clear,” “clean,” and “symmetric”) complementing the usability of interaction. As the aesthetic formality of (Chang et al., 2014) is explored through a set of manifest variables on legibility, order and being organized, it corresponds to Hassenzahl & Monk’s (2010) classic aesthetics. Aesthetic appeal, explored through being fascinating/monotonous, conventional/creative, and impressive/unremarkable refers to “expressive aesthetics”.

In SUPR-Q, the manifest variable of Appearance querying on the attractiveness of the website is conceptually related with “expressive aesthetics” and the item referring to clarity and simplicity is conceptualized as a measure of classic aesthetics. (Hassenzahl & Monk, 2010) showed that the relationship between beauty and hedonic quality is quite direct, but the relationship between beauty and pragmatic quality, observed and reported as relationship between usability and appearance of an artefact in many studies, should be explained by a mediator, the “goodness” of the evaluated product. On the other hand, their highly controlled sample of participants and well-planned experimental conditions with

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

selection of websites for their diversity in picture and text content, color, density of information, and layout, still lead to significant correlations between beauty and pragmatic quality.

Although the Appearance dimension of SUPR-Q is conceptually sound, we believe that the heavy influence of visual appeal on perceived usability, which was observed in many previous studies, had affected our results.

Trust is the behavioral intention or behavior of belief to one side against uncertainty and vulnerabilities (Moorman et al., 1992). Even if the consumer thinks that a commercial party can be trusted, if he is unwilling to trust this business, there is no customer trust. Trust is also associated and explained in the Theory of Reasoned Action (TRA) (Fishbein & Ajzen, 1977) as it directly influences attitudes, and the higher the level of trust, the better the attitude. (Suleman & Zuniarti, 2013) reports several empirical studies in marketing and consumer behaviour, showing this positive relationship (Hsu et al., 2014; Indarsin & Ali, 2017; Özkan & Kanat, 2011) Some other studies explain the trust by relying on the online purchasing behavior of consumers. It has also been investigated that trust is a very positive and significant influence on consumer purchasing decisions (Hsu et al., 2013) In accordance with the results of the same research in terms of trust will affect the attitudes and decisions of customers towards online shopping.

Another point of view taken on the concept of trust is concerned with technology. According to Lee & See (2004), technological trust is the belief that a technology is supportive of one's goals in situations where the user cannot have complete knowledge. New technologies, websites and online product-services are some of these fields, which are associated and defined with the dimensions of trust (Benbasat & Wang, 2005; Bhattacherjee, 2002; Chen & Dhillon, 2003; Cyr, 2008; Gefen, 2000).

Studies about the relationship between trust and loyalty are exploring the topic within the context of offline and online behavior. Company trustworthiness perceptions can increase customer intentions to return to a company both offline and online (Diamantopoulos & Winklhofer, 2001; Gefen, 2002; Lynch et al., 2001). It is explained that customers who trust a company are more likely to use the website, whether for a repeat visit to the site or to make an actual purchase and the more a consumer trusts a service provider, the more likely they are to continue the relationship (Cyr, 2008). Furthermore, e-loyalty has been defined as customers' favorable attitudes toward online sellers, which results in repeated purchasing behavior (Srinivasan et al., 2002). E-loyalty appears to be strongly related to customers' perceived website service quality, trustworthiness, and its consequences. E-loyal customers are profitable and comparably less price sensitive (Porter, 2001; Reichheld & Schefter, 2000).

As discussed in the studies mentioned above, there is a causal relationship between trust and loyalty, either in offline or online customer behavior. Trust is one of the prerequisites for customers to return to a company to take further services and make more purchases, through its website or app, as also through other channels. In addition, the technology trust is also involved while the trust on a web site or an app is inquired. Thus, the responses to the items to the Turkish translated SUPR-Q were highly similar to each other for trust and loyalty dimensions, since the participants have evaluated the websites and apps of their last online purchase, which are mainly the sites that they often use, due to their loyalty to these sites, built on a positive level of trust on these companies and their digital channels.

With the sampling method that we have used, which is asking the participants to evaluate a web site of or app of their own choice, it was not possible to clearly discriminate the concepts of appearance from usability and trust from loyalty. Further studies need to explore this issue through an experimental approach where researchers assign a website or app to the participants that they did not have any prior knowledge, experiences, and attitudes towards.

It should be noted that there is evidence on discriminant validity of four factor Turkish-translated SUPR-Q through Fornell-Larcker criterion, but not through the HTMT criterion. For the researchers who are willing to use SUPR-Q in Turkish, we also provided an analysis for two-constructs: Ergonomics which involves items for Appearance and Usability, and Credibility, which is a combination of Trust and Loyalty items. From this point of the paper, the four-construct version of Turkish translated SUPR-Q will be referred as TR4 - SUPR-Q, and the two-construct version will be named as TR2-SUPR-Q.

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

Below, you may find the psychometric qualities for TR2-SUPR-Q. Afterwards, the results on convergent validity and sensitivity will be given and discussed for both versions.

Psychometrics for Alternative Two-Factor Model of Turkish SUPR-Q

The two-factor model we constructed with Ergonomics and Credibility dimensions is also evaluated through PLS based confirmatory method. The items have higher loadings on their intended dimensions, but their load on the other dimension is also moderately strong, as seen on Table 12.

Table 12. Cross-loadings of items on two -factor model TR2-SUPR-Q

	Ergonomics	Credibility
A2	**0.883**	0.659
U2	**0.849**	0.492
U1	**0.830**	0.497
A1	**0.740**	0.430
L1	0.674	**0.898**
C1	0.520	**0.852**
L1	0.410	**0.659**
C2	0.351	**0.761**

Reliability indicators given in Table 13 are slightly higher than TR4-SUPR-Q (see Table 7), but do not indicate any redundancy. Considering the AVE values, the factors in the two-item model are less capable explaining the variance of their respective indicators, but still above the satisfactory level of .5.

Table 13. Indicators of internal consistency for two-factor model TR2-SUPR-Q

	Ergonomics	Credibility
Cronbach's Alpha	0.846	0.807
rho_A	0.873	0.867
Composite Reliability	0.896	0.874
Average Variance Extracted (AVE)	0.684	0.637

The √AVE for Ergonomics is .827 and for Credibility, it is .798, which are higher than the .641 correlation between the two factors, providing evidence for discriminant validity through Fornell-Larcker criterion. The HTMT ratio of the factors is .728, also supporting the discriminant validity.

Although there is not a consensus on the use of model fit indicators in PLS literature (J. Hair et al., 2017; Henseler et al., 2016), we also reported model fit measures on Table 14, the SRMR (standardized root mean square residual) and NFI (normal fit index), and exact fit measures of d_ULS and d_G along with the Chi-square values, for future reference.

Table 14. Model fit indicators for four-factor and two-factor models

	TR2-SUPR-Q	TR4-SUPR-Q
SRMR	0.092	0.096
d_ULS	0.306	0.332
d_G	0.110	0.221
Chi-Square	75.135	170.134
NFI	0.841	0.640

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

Comparison of Results Acquired via Four-factor and Two-Factor Models

Sensitivity Shopping Platform: Website and App Shopping

Through a series of independent samples t-tests comparing the means for dimensions TR4-SUPR-Q, TR2-SUPR-Q and overall score of 8 items, we detected significant differences for the means scores of app and website shoppers, as given in Table 15. The results are not unexpected, since the app users installed the app on their own device with a future intention to use it. As they have confidence in the shopping service, they agreed to install a software provided by the service. As they did not experience any usability issue that may lead them to quit the use of the app and uninstall it, their repeated use also enhanced their abilities in the app. They also get used to the "look and feel" of the app interface. Consequently, all indicators and the overall score given by the app users are higher than website users. However, the mean differences are varying between .18 to .32, which are quite close to each other. The standard deviations reported for the website scores are higher than the app scores, suggesting that app users are more likely to give similar scores while evaluating their long-term choice of app, as the score for websites are likely to be in a wider spectrum, as web shoppers are less dedicated, regarding to their frequency of use for the medium they evaluated (see Table 7).

Table 15. Differences for the means scores of app and website shoppers

		Websites		Apps		
		M	SD	M	SD	t-test
TR4	Appearance	3.60	.90	3.93	.62	t(288)=-3.63, p<.001
	Usability	4.14	.81	4.41	.63	t(288)=-3.23, p<.01
	Loyalty	4.30	.79	4.53	.62	t(288)=-2.85, p<.01
	Trust	3.90	.78	4.09	.71	t(239.92)=-2.02, p<.05
TR2	Credibility	4.10	.72	4.31	.59	t(288)=-2.72, p<.01
	Ergonomics	3.87	.79	4.17	.56	t(288)=-3.77, p<.001
Overall		3.99	.67	4.24	.50	t(288)=-3.66, p<.001

Sensitivity to differences between websites or apps

We also would like to assess the sensitivity of Turkish SUPR-Q when the same shopping service is evaluated by its users who prefer different platforms. For this reason, we executed a mean comparison on scores given to Service A, Service B and C, which were evaluated by 74 websites and 120 app users (see Table 4).

Table 16. Mean comparison of three services on website and app scores

	Service	Websites			Apps		
		M	SD	Effect (ANOVA)	M	SD	Effect (ANOVA)
Appearance	A	3.68	1.00		3.98	0.59	
	B	3.42	0.96	F(2, 71) =0.627, p >.05	3.76	0.48	F(2, 118) =1.197, p >.05
	C	3.73	0.56		3.86	0.75	
Usability	A	4.30	0.74		4.54	0.51	
	B	3.89	0.92	F(2, 71) =1.956, p >.05	4.08	0.65	F(2, 118) =5.588, p <.05
	C	4.03	0.64		4.29	0.91	
Loyalty	A	4.19	0.81		4.56	0.58	
	B	4.26	0.75	F(2, 71) =2.604, p >.05	4.50	0.69	F(2, 118) =0.101, p >.05
	C	4.70	0.48		4.61	0.54	
Trust	A	3.73	0.84	F(2, 71) =3.261, p <.05	4.12	0.67	F(2, 118) =0.568, p >.05

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

	B	3.92	0.65		3.97	0.84	
	C	4.30	0.56		4.29	0.81	
Credibility	A	3.96	0.74		4.34	0.54	
	B	4.09	0.63	F(2, 71) =3.625, p <.05	4.24	0.73	F(2, 118) =0.386, p >.05
	C	4.50	0.46		4.45	0.66	
Ergonomics	A	3.99	0.78		4.26	0.48	
	B	3.66	0.91	F(2, 71) =1.172, p >.05	3.92	0.53	F(2, 118) =3.743, p <.05
	C	3.88	0.52		4.07	0.80	
Overall	A	3.97	0.70		4.30	0.46	
	B	3.88	0.71	F(2, 71) =1.03, p >.05	4.08	0.57	F(2, 118) =1.62, p >.05
	C	4.19	0.37		4.26	0.60	
T-CSUQ Overall	A	2.33	0.95		2.02	0.72	
	B	2.25	1.01	F(2, 71) =0.081, p >.05	2.45	0.80	F(2, 117) =3.899, p <.05
	C	2.25	0.65		2.52	0.49	
T-CSUQ SysUse	A	2.10	0.99		1.87	0.71	
	B	2.16	0.99	F(2, 71) =0.063, p >.05	2.37	0.92	F(2, 117) =5.091, p <.01
	C	2.04	0.71		2.45	0.42	
T-CSUQ InfoQual	A	2.56	1.14		2.23	0.90	
	B	2.42	1.21	F(2, 71) =0.313, p >.05	2.75	1.05	F(2, 117) =3.04, p >.05
	C	2.31	0.68		2.67	0.86	
T-CSUQ IntQual	A	2.58	1.08		2.19	0.92	
	B	2.33	1.10	F(2, 71) =0.362, p >.05	2.39	0.80	F(2, 117) =0.746, p >.05
	C	2.51	0.68		2.52	0.63	
UMUX Overall	A	64.88	17.80		68.75	18.92	
	B	58.01	17.88	F(2, 42) =0.656, p >.05	60.07	19.50	F(2, 55) =1.035, p >.05
	C	62.12	14.00		63.89	17.41	

Websites: N=74 except for UMUX N=44; Apps: N=120 except for UMUX N=58. Participants were asked to respond to UMUX only if they speak English.

A series of one-way ANOVA revealed a significant difference on Trust dimension of four-factor version and Credibility dimension or two-factor version, when the mean scores on the websites of three services were compared, given on Table 15. Bonferroni post-hoc test showed that effect was due to the significant differences between Service A and Service C.

A similar analysis was made for the mean scores of the same three services, which only lead to a significant difference on Usability dimension of TR4-SUPR-Q, but the effect was not significant on Ergonomics dimension of TR2-SUPR-Q. A Bonferroni post-hoc test showed that the Usability score of Service A is significantly higher than Service B. Results can be observed on Table 15.

Based on our data set, it is possible to claim that TR2 or TR4 SUPR-Q is capable of distinguishing between different websites or apps. However, the two-factor version is less sensitive to the differences between sites. On the other hand, it should be considered that the scores reported are given for highly popular e-commerce services. These services do not trade goods on their own behalf, but they provide online infrastructure for merchants, in exchange of profit shares. One of the services is owned by a global company originated from China, one is the localized operation for a US based online shopping service, and the last one is the oldest online retail shopping service provider owned by one of the largest holdings in Turkey. Thus, it is not surprising that these sites and apps are created and operated by a team of professionals who monitor the user feedback and enhance the services continuously. As given

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

on Table 16, the mean scores vary between 3.4 to 4.6, which are relatively high, and close to each other for all three services.

Furthermore, we observed similar effects on the scores of T-CSUQ and UMUX scores. There is not a significant difference on UMUX score neither between apps nor websites. Evaluations on three websites did not reveal a significant difference based on T-CSUQ scores. For apps, there is an effect of evaluated site on the mean score on T-CSUQ SysUse dimension and the overall CSUQ score. The Bonferroni post-hoc test indicates that the effect is due to the significant difference between Service A and C, as it was observed on the Usability dimension of TR4-SUPR-Q.

Concurrent Validity

We explored the Pearson correlations of the SUPR-Q scores with UMUX and T-CSUQ scores for the web site evaluations, given on Table 17. All correlations were statistically significantly different than zero at the $p < .01$ level. UMUX score, obtained from 71 participants who responded to UMUX in English, moderately correlates with the scores obtained via Turkish translated SUPR-Q items. On the other hand, T-CSUQ scores obtained from 120 website evaluators strongly correlate with scores obtained via TR2-SUPR-Q and TR4-SUPR-Q. Negative correlations are due to the reverse item structure of CSUQ, that the lower mean scores indicate a better user experience. While the observed correlations are slightly weaker than the correlations observed between SUS and SUPR-Q in the original study (Sauro, 2015), they provide evidence that Turkish translated items and suggested alternative two factor structure has concurrent validity with other scales developed for evaluating usability. On the other hand, it should be considered that neither T-CSUQ nor UMUX queries users about trust and loyalty towards the system, although they correlated with related dimensions of SUPR-Q, both in the original study and our study.

Table 17. Correlations between scores obtained through website evaluations.

	UMUX	T-CSUQ Overall	T-CSUQ SysUse	T-CSUQ InfoQual	T-CSUQ IntQual
Appearance	0.503	-0.725	-0.694	-0.62	-0.698
Usability	0.423	-0.68	-0.698	-0.54	-0.644
Loyalty	0.314	-0.594	-0.576	-0.497	-0.578
Trust	0.44	-0.517	-0.476	-0.495	-0.487
Credibility	0.42	-0.61	-0.578	-0.545	-0.585
Ergonomics	0.497	-0.764	-0.755	-0.632	-0.729
Overall	0.503	-0.77	-0.748	-0.659	-0.737

When the Pearson correlations of dimensions were explored for apps given on Table 18, we observed that UMUX score obtained from 85 app evaluators did not correlate statistically significantly different than zero with Trust and Loyalty besides Credibility score obtained from the same item set with the other two.

Table 18. Correlations between scores obtained through app evaluations.

	UMUX Overall	T-CSUQ Overall	T-CSUQ SysUse	T-CSUQ InfoQual	T-CSUQ IntQual
Appearance	.276*	-.476**	-.404**	-.410**	-.478**
Usability	.280**	-.560**	-.503**	-.484**	-.515**
Loyalty	.022	-.410**	-.367**	-.335**	-.390**
Trust	.127	-.346**	-.299**	-.278**	-.343**
Credibility	.092	-.425**	-.374**	-.344**	-.413**
Ergonomics	.309**	-.576**	-.505**	-.498**	-.553**
Overall	.234*	-.568**	-.499**	-.477**	-.548**

*Significant at .05 level ** Significant at .01 level

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

Furthermore, other dimensions of SUPR-Q have weak correlations with UMUX score, although significantly different from zero. Only the Ergonomics score is correlating moderately, but not very different from Appearance and Usability. As UMUX is known to be strongly correlating with SUS (Lewis, 2018; Lewis et al., 2015; Berkman et al., 2016), the results are unexpected. However, we observed that TR2-SUPR-Q and TR4-SUPR-Q overall score and dimensions are moderate correlating with T-CSUQ, which is applied in Turkish, while UMUX was applied in English in our study.

CONCLUSION

Our psychometric evaluation of Turkish translated SUPR-Q items provided strong evidence for the internal consistency of its subscales, suggesting it is a reliable measure of user experience.

While some of our findings suggest that the four-factor measurement model of the original scale can be re-constructed based on Turkish items, discriminant validity of this structure was not completely verified in this study. For this reason, a two-factor model is suggested as an alternative. However, it should be considered that previous studies suggest a strong statistical relationship between the "hedonic quality" and "pragmatic quality" of information technology artefacts, since the appearance of them. Likewise, "trust" is associated with "loyalty" as a pre-condition, as the trusted product or service is used repeatedly, creating loyalty.

Furthermore, the factor analysis in our study is conducted on the data collected for the websites of online retail services. An analysis on data collected from a larger diversity of sites, such as sites that users can purchase services like tickets, digital content or insurance may provide a richer variety in data. In addition, aggregator sites, where users can browse goods or services from multiple providers can be included. Another limitation of this study is concerned with the easy sampling methodology, in which the participants were asked to evaluate a site or app of their choice. Data collected for a selection of sites and apps with a variety of interaction and interface design styles may provide a dataset that may reveal clear evidence for discriminant validity. We suggest researchers to report the results for both versions, until there is clear evidence for the validity of either measurement model.

Our results suggest that our translated SUPR-Q items are capable of distinguishing between the sites or apps, and more sensitive than the UMUX applied in English, while its sensitivity is comparable with the Turkish T-CSUQ-SV. Considering that SUPR-Q has fewer items, it can be a good choice for researchers who need a brief tool. Furthermore, the evidence we provided on sensitivity suggest that Turkish translation of SUPR-Q can be used to "generate a database used to produce percentile ranks and make scores more meaningful to researchers and practitioners", as it is suggested for the original SUPR-Q.

Although SUPR-Q is not developed as a tool for assessing user experience on mobile applications, our results suggest that it can be used for this purpose. Considering that most of the modern mobile apps are sharing similar interface and interaction design patterns with websites and have the same purpose, the content of SUPR-Q is applicable to these apps. Along with the concurrent validity with T-CSUQ which is observed both on apps and websites, different results that are obtained for three apps evaluated in our study shows that SUPR-Q is applicable for shopping apps as well as shopping sites. However, web and app scores are not comparable since their context of use can be different as well as the user motivations and the interface design.

Future studies for providing further evidence on validity of the Turkish translation should embrace a more controlled approach for data collection, to provide data from a variety of sites and apps. Along with the popular mainstream instances, a large amount of data should also be collected through less known sites and apps, which are more likely to be free from positive user bias and may yield to a higher variability in the data.

REFERENCES

Aladwani, A. M., & Palvia, P. C. (2002). Developing and validating an instrument for measuring user-perceived web quality. *Information and Management*, *39*(6). *doi.org/10.1016/S0378-7206(01)00113-6*

Alharoon, D., & Gillan, D. J. (2020). The Relation of the Perceptions of Aesthetics and Usability. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1). *doi.org/10.1177/1071181320641452*

Barnes, S. J. (2020). Information management research and practice in the post-COVID-19 world. *International Journal of Information Management*, *55*. *doi.org/10.1016/j.ijinfomgt.2020.102175*

Barnes, S.J., & Vidgen, R. (2000). WebQual : An Exploration of Web-site Quality. *Communications*, *1*, 298–305. *doi.org/10.1.1.107.5463*

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the association for information systems, 6(3), 4. doi.org/10.17705/1jais.00065*

Bhattacherjee, A. (2002). Individual trust in online firms: Scale development and initial test. *Journal of Management Information Systems, 19, 1, 211–241. doi.org/10.1080/07421222.2002.11045715*

Berkman, M. I., & Karahoca, D. (2016). Re-Assessing the Usability Metric for User Experience (UMUX) Scale. *Journal of Usability Studies*, *11*(3), 89–109. *dl.acm.org/citation.cfm?id=2993221*

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

Chang, S. H., Chih, W. H., Liou, D. K., & Hwang, L. R. (2014). The influence of web aesthetics on customers' PAD. *Computers in Human Behavior*, *36*. *doi.org/10.1016/j.chb.2014.03.050*

Chen, S. C., & Dhillon, G. S. (2003). Interpreting dimensions of consumer trust in e-commerce. *Information technology and management, 4(2), 303-318. doi.org/10.1023/A:1022962631249*

Cyr, D. (2008). Modeling web site design across cultures: relationships to trust, satisfaction, and e-loyalty. *Journal of management information systems, 24(4), 47-72. doi.org/10.2753/MIS0742-1222240402*

Demirkol, D., & Şeneler, Ç. (2018). A Turkish translation of the system usability scale: The SUS-TR. *Uşak Üniversitesi Sosyal Bilimler Dergisi, 11(3), 237-253.*

Demirkol, D., Seneler, C., Daim, T., & Shaygan, A. (2020). Measuring perceived usability of university students towards a student information system (SIS): A Turkish university case. *Technology in Society, 62, 101281.*

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of marketing research, 38(2), 269-277. doi.org/10.1509/jmkr.38.2.269.18845*

Dijkstra, T. K., & Henseler, J. (2015). Consistent Partial Least Squares Path Modeling. *MIS Quarterly*, *39*(2), 297–316. *www.jstor.org/stable/26628355*

Erdinç, O., & Lewis, J. R. (2013). Psychometric Evaluation of the T-CSUQ: The Turkish Version of the Computer System Usability Questionnaire. *International Journal of Human-Computer Interaction*, *29*(5). *doi.org/10.1080/10447318.2012.711702*

Finstad, K. (2010). The Usability Metric for User Experience. *Interacting with Computers*, *22*(5). *doi.org/10.1016/j.intcom.2010.04.004*

Fishbein, M., & Ajzen, I. (1975). Belief, Attitude, Intention and Behavior: An Introductionto Theory and Research. *Reading, MA: Addison-Wesley*.

Gefen, D. (2002). Customer loyalty in e-commerce. *Journal of the association for information systems, 3(1), 2.*

Gefen, D. (2000). E-commerce: the role of familiarity and trust. *Omega, 28(6), 725-737. doi.org/10.1016/S0305-0483(00)00021-9*

Göl, H., İlhan, E., Ot, İ., Döm, İ., & Çakır, İ. (2019). *E-Ticaretin Gelişimi, Sınırların Aşılması ve Yeni Normlar*. *www.eticaretraporu.org/wp-content/uploads/2019/05/DD-TUSIAD-ETicaret-Raporu-2019.pdf*

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Hair, J.F., Hollingsworth, C. L., Randolph, A. B., & Chong, A. Y. L. (2017). An updated and expanded assessment of PLS-SEM in information systems research. *Industrial Management & Data Systems*, *117*(3), 442–458. *doi.org/10.1108/IMDS-04-2016-0130*

Hair, J.F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152. *doi.org/10.2753/MTP1069-6679190202*

Hair, J.F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, *31*(1). *doi.org/10.1108/EBR-11-2018-0203*

Hartmann, J., Sutcliffe, A., & de Angeli, A. (2007). Investigating attractiveness in web user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07. doi.org/10.1145/1240624.1240687*

Hassenzahl, M., & Monk, A. (2010). The Inference of Perceived Usability From Beauty. *Human-Computer Interaction*, *25*(3). *doi.org/10.1080/07370024.2010.500139*

Henseler, J., Hubona, G., & Ray, P. A. (2016). Using PLS path modeling in new technology research: updated guidelines. *Industrial Management & Data Systems*, *116*(1), 2–20. *doi.org/10.1108/IMDS-09-2015-0382*

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135. *doi.org/10.1007/s11747-014-0403-8*

Hsu, M. H., Chuang, L. W., & Hsu, C. S. (2014). Understanding online shopping intention: the roles of four types of trust and their antecedents. *Internet Research, 24(3), 332-352. doi.org/10.1108/IntR-01-2013-0007.*

Hsu, C. L., Lin, J. C. C., & Chiang, H. S. (2013). The effects of blogger recommendations on customers' online shopping intentions. *Internet Research. 23(1), 69–88. doi.org/10.1108/10662241311295782*

Indarsin, T., & Ali, H. (2017). Attitude toward Using m-Commerce: The Analysis of Perceived Usefulness Perceived Ease of Use, and Perceived Trust: Case Study in Ikens Wholesale Trade, Jakarta – Indonesia. *Saudi Journal of Business and Management Studies, 2(11), 995-1007.*

Kirakowski, J., & Cierlik, B. (1998). Measuring the usability of web sites. *Proceedings of the Human Factors and Ergonomics Society*, *1*. *doi.org/10.1177/154193129804200405*

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies*, *60*(3). *doi.org/10.1016/j.ijhcs.2003.09.002*

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors, 46(1), 50-80. doi.org/10.1518/hfes.46.1.50_30392*

Leiner, D. J. (2021). *SoSci Survey (soscisurvey.de)* (Version 3.2.24).

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

Lewis, J. R. (2018). Measuring Perceived Usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction*. *doi.org/10.1080/10447318.2017.1418805*

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Investigating the correspondence between UMUX-LITE and SUS scores. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9186*. *doi.org/10.1007/978-3-319-20886-2_20*

Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., & Noonan, P. (2011). An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction*, *18*(1). *doi.org/10.1145/1959022.1959023*

Loiacono, E. T., Watson, R. T., & Goodhue, D. L. (2002). WebQual ^{TM} : A Measure of Web Site Quality. *Marketing Theory and Applications*, *13*(3).

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate

Loiacono, E. T., Watson, R. T., & Goodhue, D. L. (2007). WebQual: An instrument for consumer evaluation of web sites. *International Journal of Electronic Commerce, 11(3). doi.org/10.2753/JEC1086-4415110302*

Lynch, P. D., Kent, R. J., & Srinivasan, S. S. (2001). The global internet shopper: evidence from shopping tasks in twelve countries. *Journal of advertising research, 41(3), 15-23. doi.org/10.2501/JAR-41-3-15-23*

Moorman, C., Zaltman, G., and Deshpande, R. (1992). Relationships between providers and users of market research: The dynamics of trust within and between organizations. Journal of marketing research, 29(3), 314-328.

Özkan, S., & Kanat, I.E. (2011). e-Government adoption model based on theory of planned behavior: Empirical validation. *Government Information Quarterly, 28(4), 503-511. doi.org/10.1016/j.giq.2010.10.007.*

Porter, M. (2001), Strategy and the internet, *Harvard Business Review, Vol. 97 No. 3, pp. 62-78.*

Reichheld, F. F., & Schefter, P. (2000). E-loyalty: your secret weapon on the web. *Harvard business review, 78(4), 105-113.*

Reichheld, F. F. (2003). The One Number You Need to Grow. In *Harvard Business Review 81(12)*

Reinartz, W., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing*, *26*(4). *doi.org/10.1016/j.ijresmar.2009.08.001*

Ringle, C., Wende, S., & Becker, J.-M. (2015). *SmartPLS3* (3.3.2). Boenningstedt: SmartPLS GmbH. *smartpls.com*

Sauro, J. (2015). SUPR-Q: A Comprehensive Measure of the Quality of the Website User Experience. *Journal of Usability Studies*, *10*(2), 68–86. *doi.org/10.5555/2817315.2817317*

Schreiber, J. B. (2021). Issues and recommendations for exploratory factor analysis and principal component analysis. *Research in Social and Administrative Pharmacy*, *17*(5). *doi.org/10.1016/j.sapharm.2020.07.027*

Skulmowski, A., Augustin, Y., Pradel, S., Nebel, S., Schneider, S., & Rey, G. D. (2016). The negative impact of saturation on website trustworthiness and appeal: A temporal model of aesthetic website perception. *Computers in Human Behavior*, *61*. *doi.org/10.1016/j.chb.2016.03.054*

Srinivasan, S. S., Anderson, R., & Ponnavolu, K. (2002). Customer loyalty in e-commerce: an exploration of its antecedents and consequences. *Journal of retailing, 78(1), 41-50. doi.org/10.1016/S0022-4359(01)00065-3*

Suleman, D., & Zuniarti, I. (2019). Consumer Decisions toward Fashion Product Shopping in Indonesia: The effects of Attitude, Perception of Ease of Use, Usefulness, and Trust. Management *Dynamics in the Knowledge Economy, 7(2), 133-146.*

van Schaik, P., & Ling, J. (2003). The effect of link colour on information retrieval in educational intranet use. *Computers in Human Behavior*, *19*(5). *doi.org/10.1016/S0747-5632(03)00004-9*

Wang, J., & Senecal, S. (2008). Measuring perceived website usability. *Journal of Internet Commerce*, *6*(4). *doi.org/10.1080/15332860802086318*

Yılmaz, Ö., & Bayram, O. (2020). COVID-19 pandemi döneminde Türkiye'de e-ticaret ve e-ihracat. *Kayseri Üniversitesi Sosyal Bilimler Dergisi*. *doi.org/10.51177/kayusosder.777097*

Submit Date: 29.06.2021, Acceptance Date: 12.08.2021, DOI NO: 10.7456/11104100/009

Research Article - This article was checked by iThenticate