International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Development and Psychometric Validation of a Positively Worded German Version of the System Usability Scale (SUS)

Sebastian A. C. Perrig, Nick von Felten, Beat Vollenwyder & Klaus Opwis

To cite this article: Sebastian A. C. Perrig, Nick von Felten, Beat Vollenwyder & Klaus Opwis (11 Dec 2024): Development and Psychometric Validation of a Positively Worded German Version of the System Usability Scale (SUS), International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2024.2434720

To link to this article: https://doi.org/10.1080/10447318.2024.2434720

Published online: 11 Dec 2024.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2024.2434720

Development and Psychometric Validation of a Positively Worded German Version of the System Usability Scale (SUS)

Sebastian A. C. Perrig, Nick von Felten, Beat Vollenwyder, and Klaus Opwis

Center for General Psychology and Methodology, University of Basel, Basel, Switzerland

ABSTRACT

Despite its popularity, knowledge of the System Usability Scale's (SUS) psychometric quality in German is limited. This article developed a positively worded German SUS version and investigated two existing versions. In a preregistered first study ($N = 250$), positive alternatives for negative items were evaluated using item analyses, exploratory factor analyses, and language expert suggestions. The two existing versions were also compared. A preregistered second study ($N = 877$) involved participants interacting with 12 websites and completing the different SUS versions to validate the new positive-only and the original versions. Analyses included item analyses, confirmatory factor analyses, correlations with related scales, and error analyses. Findings indicated that the SUS-DE-Pos, a positive version based on Rummel, performed best. Researchers should use this version or Rummel's version if negative items are needed. Overall, this work provides validated German SUS versions, with and without negative items, addressing a critical gap for German-speaking HCI researchers and practitioners.

KEYWORDS

Usability; System Usability Scale (SUS); measurement; questionnaire; survey scale; user experience (UX); websites; reverse-coded items; negative items

1. Introduction

Among the survey scales used in Human–Computer Interaction (HCI) and user experience (UX) research, the System Usability Scale (SUS) by Brooke (1996) is among the most popular (Bargas-Avila & Hornbæk, 2011; Perrig, Aeschbach, et al., 2024; Pettersson et al., 2018). Initially described as a "quick and dirty usability scale" (Brooke, 1996), the 10-item SUS has proven to be a quick but not "dirty" way to assess the perceived usability of users after interacting with a system (Lewis, 2018b). Numerous translations of the SUS into German have been proposed and are currently being employed, while no consensus has been reached regarding which German version of the SUS to use. In an effort to clarify the landscape of varying German SUS translations, Brix et al. (2023) identified two versions to be promising contenders: Rummel (2016) and Gao et al. (2020). However, both versions differ in their degree of validation and naturalness of wording (Brix et al., 2023). Therefore, a comprehensive comparison between these two proposed versions of the SUS in German is necessary to determine which version German-speaking HCI researchers and practitioners should use.

In addition to translation issues, numerous studies have tried to investigate what theoretical model most adequately fits the data collected with the SUS. These studies either confirmed the proposed unidimensional structure (Bangor et al., 2008), suggested alternative two-dimensional structures (Lewis & Sauro, 2009), or suggested that the alternating use of positively and negatively worded items may introduce artificial two-dimensionality (Lewis & Sauro, 2017). Negatively worded items are known to cause various psychometric issues, including lower reliability, reduced cross-cultural applicability, distorted factor structures, and higher error rates (Barnette, 2000; Sauro & Lewis, 2011; Schmitt & Stuits, 1985; Wong et al., 2003). To remedy this artificial distortion due to item wording, Sauro and Lewis (2011) developed an alternative version of the SUS devoid of negative items (henceforth called SUS-Pos). This positive scale version yields comparable results to the original version, while circumventing associated problems with negative items. Despite these advancements, a positively worded version of the SUS in German has not yet been created. Thus, this study seeks to develop and validate a German SUS with positive-only items.

Another area of concern is how survey instruments perform across different contexts. Simply translating the English versions of the SUS may not adequately capture the cultural and contextual differences that can impact scale reliability and validity (Furr, 2011). This highlights the importance of validating translations to adapt the scale for cross-cultural contexts. Additionally, the impact of different device types as a context deserves attention. Given the rapid rise in smartphone use worldwide (Tenzer, 2024), it is crucial to ensure that the psychometric properties of the SUS remain consistent across device contexts, such as desktops

CONTACT Sebastian A. C. Perrig sebastian.perrig@unibas.ch Center for General Psychology and Methodology, University of Basel, Basel, Switzerland

and mobile devices. To evaluate whether the scale's properties hold up for the translations and device types, we investigated the stability of its psychometric properties in these contexts.

In summary, the lack of an agreed-upon German version of the SUS motivated us to investigate and compare the psychometric quality of two promising translations identified in past research (Brix et al., 2023). Furthermore, the problems caused by reverse-coded items and the apparent benefits of a positive-item version of a survey scale encouraged the development of an alternative version of the SUS without negatively worded items. Additionally, we examined whether the quality of all versions remained stable across two different device contexts: desktop and mobile.

2. Related work

2.1. Usability and the System Usability Scale

Usability is commonly defined following the International Organization for Standardization (2018) 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." Furthermore, a frequently made distinction is between subjective and objective usability. While subjective usability is concerned with the perceptions and attitudes of users regarding a system, objective usability describes those properties of a system not dependent on a person's perception (Hornbæk, 2006).

However, previous research has questioned the value of the usability construct. Specifically, Tractinsky (2018) argued that usability is an umbrella construct, an overly broad concept that should be replaced by more precise constructs. Additionally, there seems to be a mismatch between the concept of usability and how it is empirically measured (Tractinsky, 2018). In contrast, Lewis (2018a) questioned whether usability truly is an umbrella construct. By summarizing past evidence, they showed that different usability scales are measuring the same underlying construct. Likewise, Borsci et al. (2019) acknowledged the importance of the issues brought up by Tractinsky (2018) but disagreed that usability is a dead end.

Despite these discussions, usability remains one of the most measured constructs in HCI research (Perrig, Aeschbach, et al., 2024), and while there are different ways to assess usability, arguably one of the most frequently used is the SUS by Brooke (1996). Originally designed to measure one construct across ten items, reflecting participants' subjective assessment of usability, sometimes also referred to as perceived usability, the SUS was initially assumed to be a unidimensional scale. However, work published in the years following the original publication of the SUS has challenged this assumed unidimensionality. In particular, Lewis and Sauro (2009) analyzed the factor structure of the SUS based on multiple independent SUS data sets. Results suggested that the SUS actually measures two factors: usability and learnability. Several years later, the authors revisited their argument (Lewis & Sauro, 2017), analyzing additional data collected with the SUS. Based on insights from this new data, they concluded that the SUS does not measure two distinct factors. Instead, the two factors observed result from item tone, with positively and negatively formulated items loading onto two distinct factors. Lewis and Sauro (2017) further argued that such a distinction based on item tone is of little theoretic or practical interest. Thus, they recommended that researchers should treat the scale as a unidimensional measure of perceived usability.

2.2. German System Usability Scale(s) and cross-cultural psychometric evaluation

Survey scales require thorough psychometric evaluation to be considered valid and reliable across different contexts. However, even if a survey scale has undergone rigorous assessments, a careful translation does not guarantee its psychometric stability across different languages (Furr, 2011). For this reason, it is important not only to translate but also to validate translated survey scales to ensure the constructs and understanding of the items are comparable across cultures (van de Vijver & Hambleton, 1996). This seems especially important for a construct like usability, which is heavily context-dependent (Clemmensen et al., 2009). Thus, the SUS, as one of the most popular survey scales to measure usability, has seen numerous language translations (Gao et al., 2020).

Recently, a scoping review sought to clarify the landscape of varying German SUS versions. In their work, Brix et al. (2023) aimed to identify and compare all currently available German translations of the SUS. They found four versions and examined them regarding the translation process and wording. Among these, two versions were determined to be particularly promising. The first version identified was by Rummel (2016), but while its wording is promising, this SUS variant has not yet undergone proper validation. Conversely, the second version recommended by Brix et al. (2023) is the SUS translation by Gao et al. (2020). However, while this version has been systematically validated, it was criticized for containing unnatural wording. These shortcomings led Brix et al. (2023) to refrain from recommending one version over the other, resulting in ambiguity regarding which scale to use. However, if too many different German versions of the SUS are being used across the research field, the comparability of findings across research projects is reduced (Perrig, Aeschbach, et al., 2024). Thus, it is crucial for German-speaking HCI researchers to know which version of the SUS to use and, ideally, to agree on one version.

2.3. Negative item tone

The items of a survey scale are usually designed to reflect various aspects of the construct to be measured, based on the idea that the unobservable construct itself is affecting how participants respond to the items (DeVellis, 2017). Further, a special kind of items are so-called negative or reverse-coded items, which are worded opposite to the other items of the scale (ie, the positive items). Negative items are

typically created either by adding negations (eg, "I thought the system was not easy to use" instead of "I thought the system was easy to use") or by working with antonyms (eg, "I found the system unnecessarily complex" rather than "I found the system to be simple") (Suárez-Álvarez et al., 2018). These items are introduced into survey scales to counteract response biases, such as extreme response bias (ie, the tendency to select the extreme answering options of the scale) or acquiescence bias (ie, the tendency to agree with most or all items in a scale) (Dalal & Carter, 2014; Sauro & Lewis, 2011).

Yet, past work has highlighted several issues with negatively worded items, including lower scale reliability (Barnette, 2000), distorted factor structure (Schmitt & Stuits, 1985), and reduced cross-cultural applicability (Wong et al., 2003). Furthermore, negatively worded items come with a risk of misinterpretation and mistakes by participants, in addition to miscodings by researchers (Sauro & Lewis, 2011). Sauro and Lewis (2011) thus recommended that HCI researchers avoid negative item formulation when developing new scales. This recommendation was empirically tested in a modified version of the SUS, where the negatively worded items were replaced with positive ones (Sauro & Lewis, 2011). This revised scale, referred to in this article as the SUS-Pos, was found to perform as well as the original version while avoiding several of its disadvantages in both the original study by Sauro and Lewis (2011) and a replication study by Kortum et al. (2021). A similar project on a website aesthetics scale found that using only positive item wording improved scale quality (Perrig, von Felten, et al., 2023).

2.4. Device contexts

Smartphone use is developing rapidly worldwide. While the global number of smartphone users amounted to around 4.7 billion in 2022, around 500 million more than in the previous year, an increase to 5.1 billion is forecast for 2028 (Tenzer, 2024). The psychometric quality of a survey scale is likely to differ depending on the context of use (Furr, 2011). Despite the popularity of smartphone devices, information on the psychometric quality of the two German SUS versions in a mobile device context has yet to be reported to the extent of our knowledge. At the same time, past work has already employed the original SUS to measure the usability of different mobile device applications (Kortum & Sorber, 2015). There are substantial differences between mobile and desktop interfaces, including smaller screen sizes and differences in precision between clicking and touch interfaces (Nielsen & Budiu, 2013). Thus, knowing whether the SUS performs equally well across different device types is crucial. Therefore, the present work also investigated whether the psychometric quality of the different German SUS versions would remain stable when comparing data collected in a desktop/laptop computer setting to responses collected in a mobile device context.

Finally, we acknowledge that emerging technologies, such as AI chatbots and machine learning, are increasingly gaining importance. However, we decided to focus on the more traditional device context of mobile devices and desktop websites, given that these contexts are still crucial in everyday life and will likely continue to be so for the foreseeable future.

3. Study overview and objectives

The original SUS consists of ten items, half of which are negatively worded. In preparation for the present project, the five negative items of the SUS were reworded to create positive alternatives (three to five positive options per negative item). These alternatives were then sent to a German language expert from the language center at the University of Basel for review. After the language center review, the new set of items was investigated in a first preregistered online study alongside the German SUS versions proposed by Gao et al. (2020) and Rummel (2016) to identify those positively worded items psychometrically closest to the negative counterparts. The two German SUS versions were chosen based on the recommendations by Brix et al. (2023). As a next step, the new positive versions of the German SUS, containing only positive items (SUS-DE-Gao-Pos, SUS-DE-Rummel-Pos), were compared to the other two German versions of the scale (SUS-DE-Gao, SUS-DE-Rummel) in a second preregistered online study, situated in both a desktop/laptop and mobile setting, following current best practices for scale quality investigation. Results from this second study showed that all four SUS versions achieved satisfactory psychometric quality, which remained stable across device types (desktop vs. mobile), although the positive version based on Rummel (2016) performed best. In addition, results suggested that this positive version delivered comparable results to the original version proposed by Rummel (2016) while reducing the number of mistakes participants make in their responses by avoiding negative item formulation. Thus, this scale version is likely most suitable for German-language HCI research.

Based on the information summarized thus far, we formulated the following three research objectives for the present project:

Objective 1:

Investigation of the psychometric quality of the German SUS version by Rummel (2016) following current psychometric practice.

Objective 2:

Investigation of the psychometric quality of the German SUS version by Gao et al. (2020) following current psychometric practice.

Objective 3:

Creation of a positively worded version of the SUS in German (SUS-DE-Pos) that performs psychometrically at least as good or better than the other two German versions of the SUS.

Overall, the present project aimed to make a clear recommendation to German HCI researchers and practitioners which version of the SUS to best employ in their research

while offering them a positive alternative version of the scale, allowing them to avoid issues associated with using negatively formulated survey items.

4. Item development

As a first step, a collection of alternative items to the negatively formulated SUS items in German was developed. The goal was to create a set of appropriate items that retained the meaning of the five original reverse-scored SUS items while avoiding negative wording. To create these new non-reversed items, the first and second authors, both native German speakers, formulated positive alternatives to the original negatively worded SUS items by working independently of each other with multiple dictionaries and by using their knowledge. Next, they discussed this first set of possible alternatives, combining items, formulating new alternatives, and deciding on favorites. In these initial steps, items were formulated both by translating the English items of the original (Brooke, 1996) and positive SUS (Sauro and Lewis, 2011), as well as by looking for antonyms and opposite formulations to the German negative items from the SUS versions by Gao et al. (2020) and Rummel (2016). After settling on a set of initial candidate items, the third and fourth authors gave feedback on the item pool and made additional suggestions, resulting in a preliminary set of 21 items, with three to five alternatives per original negative SUS item. Next, items were sent to a German language expert offering academic editing at the authors' University for an expert review. The first author met with the language expert for a discussion of the items, resulting in a final selection of ten items, with two positively worded alternatives per negative SUS item. All item pools from the individual steps are provided in the supplementary materials on the Open Science Framework (OSF), while the final selection of items is provided in Table 1.

5. Study One: Item selection

After an initial set of alternatives to the reverse-coded German SUS items was developed, this set had to be further reduced and refined, ideally to one alternative positive item per originally negative item, to create a new positive German SUS version. For this, a first within-subjects online experiment was conducted, in which the newly developed alternative items were employed alongside the two German SUS versions by Gao et al. (2020) and Rummel (2016). The procedure of this first study was inspired by the methodology employed for validating the German SUS version by Gao et al. (2020). This study was preregistered on OSF (https://doi.org/10.17605/OSF.IO/GJ7D5) to make our process transparent, clearly distinguishing between predictions and postdictions, which ultimately increases the credibility and replicability of research (Nosek et al., 2018; Simmons et al., 2021).

5.1. Methods

5.1.1. Stimuli

Participants were asked to rate a series of well-known products taken from previous work on the SUS (Gao et al., 2020; Kortum and Bangor, 2013): Excel, Word, Amazon, microwaves, and Google search. These products were chosen so that we would be able to compare the ratings collected in our study with results reported in past work employing the original English version of the SUS. Furthermore, we saw this selection of products as fitting for our study because we assumed most people would know them and still interact with them often. They span across different domains and, therefore, test the versatility of the SUS. Further, the market share of Amazon, Google, and Microsoft and the wide adoption of their products in business (eg, Excel and Word) make it unlikely that these products will completely disappear in the foreseeable future. Prior to rating a product,

Table 1. Positive alternative items for the German SUS, investigated in study one.

Original item from Rummel (2016)	Original item from Gao et al. (2020)	Positive alternative
Ich fand das Produkt unnötig komplex.	Ich fand das Produkt unnötig komplex.	**Ich fand die Komplexität des Produkts angemessen.** Ich fand das Produkt angemessen komplex.
Ich glaube, ich würde die Hilfe einer technisch versierten Person benötigen, um das Produkt benutzen zu können.	Ich denke, dass ich die Unterstützung einer technischen Person brauche, um dieses Produkt nutzen zu können.	**Ich glaube, dass ich das Produkt ohne Unterstützung einer technisch kompetenten Person benutzen könnte.** Ich glaube, dass ich das Produkt ohne Unterstützung einer technisch versierten Person benutzen könnte.
Ich denke, das Produkt enthielt zu viele Inkonsistenzen.	Ich dachte, dass dieses Produkt nicht konsistent genug war.	**Ich denke, bei diesem Produkt passt alles zusammen.** Ich denke, das Produkt ist in sich konsistent.
Ich fand das Produkt sehr umständlich zu nutzen.	Ich fand dieses Produkt sehr umständlich zu benutzen.	**Ich fand das Produkt sehr einfach zu benutzen.** Ich fand die Benutzung des Produkts intuitiv verständlich.
Ich musste eine Menge lernen, bevor ich anfangen konnte das Produkt zu verwenden.	Ich musste eine Menge Dinge lernen, bevor ich mit diesem Produkt loslegen konnte.	**Ich musste wenig lernen, bevor ich anfangen konnte das Produkt zu nutzen.** Ich konnte, ohne viel zu lernen, mit der Benutzung des Produkts beginnen.

Note: Items marked before the study as favorites are presented in bold.

participants were asked whether they knew the products they were asked to rate. If they indicated that they had not interacted with a particular product in the past, they were given the opportunity to name a comparable alternative and, if they had never interacted with that kind of product, were allowed to skip the rating of that particular product.

5.1.2. Participants

Participants were recruited on Prolific. We chose to recruit over Prolific based on our past experience with the platform (eg, Perrig, Scharowski, et al., 2024; Perrig, von Felten, et al., 2023), which showed us that it allows for the collection of large samples needed for psychometric evaluation while delivering a satisfactory level of data quality. A total of 253 participants currently living in Germany were recruited on Prolific and reimbursed £1.30 for completing the study. Participants were required to be fluent in German and were screened for the absence of colorblindness and visual impairment, given that this might impact their perception of a set of websites used as stimuli in the second study, to keep the recruitment criteria constant across the two studies. During data cleaning, following recommendations by Brühlmann et al. (2020), two participants were removed because of an interrupted survey, and one participant was removed for indicating their current country of residence outside of Germany, going against the recruitment criteria. This left a final sample of 250 responses (126 women, 123 men, and one non-binary person; mean age 31.50 years, $SD = 9.61, \min = 18, \max = 67$).

5.1.3. Measures

5.1.3.1. SUS. For each product, participants were asked to respond to a total of 28 different SUS items: ten items from Gao et al. (2020), ten items from Rummel (2016), and ten newly developed positive alternatives. Two items that are identical for Gao et al. (2020) and Rummel (2016) were only asked once per product. Responses were collected on a five-point Likert-type response scale. All items were presented in randomized order on one page, with the response scale repeated after every 11 items.

5.1.3.2. Single item adjective rating. In addition to the SUS items, participants rated each product using a single-item adjective rating of usability taken from Bangor et al. (2009). The exact wording of this and all other items is provided in the supplementary materials. This item was chosen to follow the procedure used in past work on the SUS (eg, Gao et al., 2020) and to be able to compare our results to that past work, if needed, for example in future meta-research. The item was translated from English into German by the first author, and the accuracy of the translation was ensured through a back-translation by the second author.

5.1.4. Procedure

Participants provided informed consent on the first page. Next, they were given instructions for the task to be completed. On the following pages, the five stimuli products were rated in randomized order using all measures. Next, participants were asked to provide demographic information (age, gender, country of residence).[1] Finally, participants had the opportunity to give feedback before being redirected to Prolific for payment. Following recommendations by Brühlmann et al. (2020), two instructed response items (IRI, as suggested by Brühlmann et al., 2020, based on Curran, 2016) embedded among the SUS items and a single item for self-reported data quality (SRSI UseMe, Meade & Craig, 2012) at the end of the survey were used to ensure sufficient response quality. The entire survey was presented in German. Completing the survey took participants an average of 12.37 min ($SD = 6.53$, $\min = 2.85$, $\max = 55.30$). An overview of the study procedure is given in Figure 1. Before preregistration, a small-sample pre-study was run ($N = 10$) to ensure the procedure worked as expected without technical difficulties. The data from this pre-study were included in the final analysis because no issues with the study design were observed based on this sample, and no changes were made to the survey.

5.2. Results

Reporting of the results focuses on various psychometric analyses to compare the different SUS items and versions regarding reliability and validity. Reliability is concerned with "how accurately a test measures the thing which it does measure" (Kelley, 1927, p. 14) while validity considers "whether a test really measures what it purports to measure" (Kelley, 1927, p. 14). For the quality of the individual items, item descriptive statistics, item difficulty and variance, discriminatory power, and inter-item correlations were considered. Exploratory factor analysis (EFA) was used for item selection, while confirmatory factor analysis (CFA) was then used to investigate the theoretical models and the construct validity (ie, the theoretical relationship of a variable to other variables (DeVellis, 2017)) of the different SUS versions. In addition, the individual product ratings and the correlations among ratings collected with the different SUS versions were considered to see how close the different versions are to one-another. Regarding reliability, indicators of internal consistency, namely coefficients α (Cronbach, 1951) and ω (McDonald, 1999), were calculated. All analyses can be found in the supplementary materials on OSF. Results were obtained using the statistical software R (R Core Team, 2024, version 4.4.0).

5.2.1. Item analysis and correlations

The analysis started with calculations of descriptive statistics, item-difficulty and -variance, discriminatory power, and inter-item correlations for all 28 SUS items (items from Gao et al. (2020), items from Rummel (2016), and 10 newly developed positive alternatives). In summary, item analysis showed that most of the SUS items exhibited values considered to be satisfactory across these analyses (details provided in the supplementary materials).

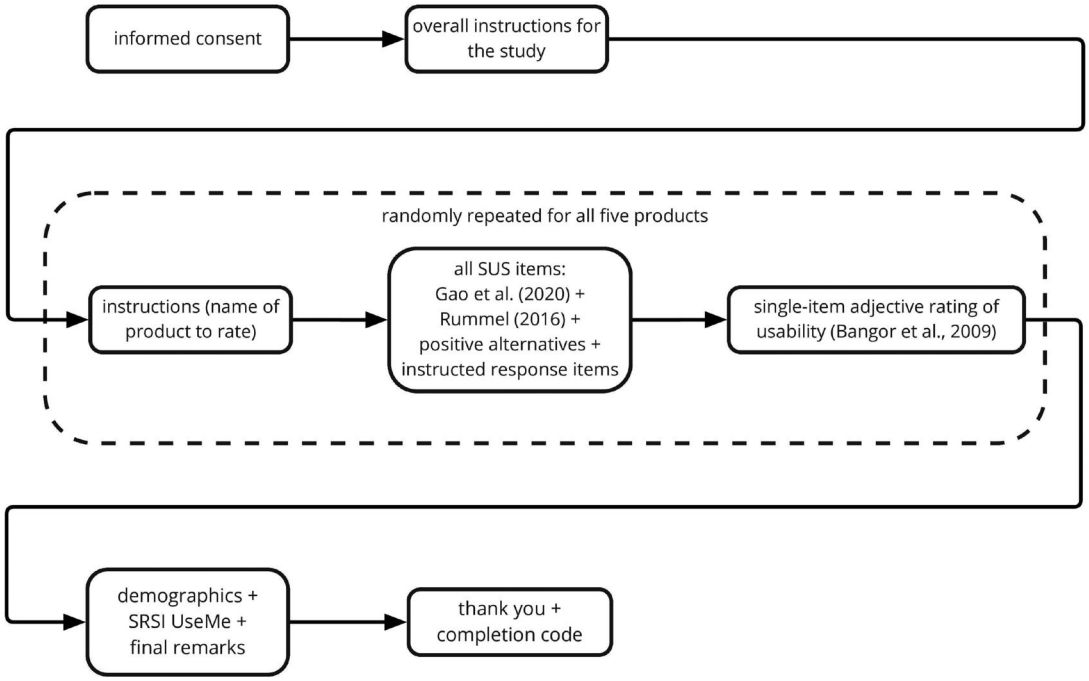

Figure 1. Flowchart depicting the procedure of study one.

In addition, we looked at correlations between the different item versions to see if any of the positive alternatives would correlate stronger with the negative item versions than the other alternatives, which was not the case for most items (see supplementary materials for details). Given the results, we saw no need to exclude any items at this point. Thus, the analysis continued while keeping the results thus far in mind.

5.2.2. Exploratory factor analyses

Next, we performed EFA with the SUS data. EFA is a method used to identify the underlying constructs or dimensions that a set of items may measure, examining which constructs or dimensions could explain the correlations observed among the measured variables (Kline, 1994). For the interpretation of factor loadings, we followed the 0.40–0.30–0.20 rule by Howard (2016), asserting that items should load at least 0.40 on a primary factor without secondary loadings >0.30 and differences of at least 0.20 between the primary and any secondary loadings. Concerning the interpretation of communality (ie, variance that is common to the measured variables), we considered values <0.30 as problematic (Hair et al., 2010). We used principal axis factoring as our factor analytic method (see de Winter & Dodou, 2012, for details) for all EFAs, as recommended by Howard (2016), because multivariate normality was not given.

We split up the SUS data into four item sets, two containing all items for Rummel (2016), once with and once without the positive alternatives, and two including all items from Gao et al. (2020) with and without the positive item versions. For each item set, we considered three types of EFAs: First, we tested a single-factor solution, following the recommended unidimensional structure of the SUS, and a two-factor solution to account for positive vs. negative item tone. Second, we consulted results from parallel analysis and a scree plot to search for the suggested number of factors to be extracted from the data. We employed an oblique rotation method (*oblimin*) for these first two types of EFAs, thus allowing the factors to be correlated. Finally, we conducted EFAs using an orthogonal rotation method (*varimax*), extracting two uncorrelated factors to exclude all items not loading on the primary factor (based on the proposed single-factor model). In addition, we calculated several EFAs across all 28 SUS items employed in the study. For the sake of brevity, the reporting of results focuses on the single-factor EFAs for Gao et al. (2020) and Rummel (2016), including the positive alternatives. Interested readers are referred to the supplementary materials for the complete results.

The single-factor EFA for the items from Gao et al. (2020) was able to explain 46.5% of variance. Table 2 contains all factor loadings >0.20 and communalities for this single-factor EFA. Results for the single-factor EFA of the items from Rummel (2016), which was able to explain 46.9% of variance, are provided in Table 3. For the selection of items for the SUS-DE-Pos, we considered both differences in loadings for the positive alternative SUS items as well as dissimilarities in their communality.

5.2.3. Item selection for the SUS-DE-Pos

Consulting all results until here, we were able to select the new items for the SUS-DE-Pos. We selected the following positive item alternatives:

- Item pos_01, based on the higher factor loadings and communalities in both single-factor EFAs, compared to

Table 2. Factor loadings >0.20 and communalities from study one in the single-factor EFA for all SUS items from Gao et al. (2020) and the positive alternatives.

Item name	Item	PA1	h2
gao_01	Ich denke, dass ich dieses Produkt häufig verwenden möchte.	0.54	0.30
rummel_gao_02	Ich fand das Produkt unnötig komplex. (R)	0.74	0.55
pos_01	**Ich fand die Komplexität des Produkts angemessen.**	0.60	0.36
pos_02	Ich fand das Produkt angemessen komplex.	0.37	0.14
gao_03	Ich dachte, das Produkt war einfach zu bedienen.	0.85	0.72
gao_04	Ich denke, dass ich die Unterstützung einer technischen Person brauche, um dieses Produkt nutzen zu können. (R)	0.66	0.44
pos_03	Ich glaube, dass ich das Produkt ohne Unterstützung einer technisch kompetenten Person benutzen könnte.	0.61	0.37
pos_04	**Ich glaube, dass ich das Produkt ohne Unterstützung einer technisch versierten Person benutzen könnte.**	0.63	0.40
rummel_gao_05	Ich fand, die verschiedenen Funktionen in diesem Produkt waren gut integriert.	0.61	0.37
gao_06	Ich dachte, dass dieses Produkt nicht konsistent genug war. (R)	0.48	0.23
pos_05	**Ich denke, bei diesem Produkt passt alles zusammen.**	0.66	0.44
pos_06	Ich denke, das Produkt ist in sich konsistent.	0.51	0.26
gao_07	Ich würde mir vorstellen, dass die meisten Leute sehr schnell lernen würden, dieses Produkt zu benutzen.	0.81	0.65
gao_08	Ich fand dieses Produkt sehr umständlich zu benutzen. (R)	0.72	0.52
pos_07	Ich fand das Produkt sehr einfach zu benutzen.	0.89	0.80
pos_08	**Ich fand die Benutzung des Produkts intuitiv verständlich.**	0.85	0.72
gao_09	Ich habe mich sehr selbstsicher gefühlt, dieses Produkt zu verwenden.	0.77	0.60
gao_10	Ich musste eine Menge Dinge lernen, bevor ich mit diesem Produkt loslegen konnte. (R)	0.73	0.53
pos_09	Ich musste wenig lernen, bevor ich anfangen konnte das Produkt zu nutzen.	0.50	0.25
pos_10	**Ich konnte, ohne viel zu lernen, mit der Benutzung des Produkts beginnen.**	0.81	0.65

Note: Selected items for the SUS-DE-Pos are marked in bold and reverse-coded items with (*R*). PA1 = factor loadings; h2 = communality.

Table 3. Factor loadings >0.20 and communalities from study one in the single-factor EFA for all SUS items from Rummel (2016) and the positive alternatives.

Item name	Item	PA1	h2
rummel_01	Ich denke, dass ich das Produkt gerne häufig benutzen möchte.	0.52	0.27
rummel_gao_02	Ich fand das Produkt unnötig komplex. (R)	0.75	0.56
pos_01	**Ich fand die Komplexität des Produkts angemessen.**	0.60	0.36
pos_02	Ich fand das Produkt angemessen komplex.	0.37	0.14
rummel_03	Ich fand das Produkt einfach zu benutzen.	0.89	0.79
rummel_04	Ich glaube, ich würde die Hilfe einer technisch versierten Person benötigen, um das Produkt benutzen zu können. (R)	0.67	0.46
pos_03	Ich glaube, dass ich das Produkt ohne Unterstützung einer technisch kompetenten Person benutzen könnte.	0.60	0.37
pos_04	**Ich glaube, dass ich das Produkt ohne Unterstützung einer technisch versierten Person benutzen könnte.**	0.63	0.40
rummel_gao_05	Ich fand, die verschiedenen Funktionen in diesem Produkt waren gut integriert.	0.61	0.37
rummel_06	Ich denke, das Produkt enthielt zu viele Inkonsistenzen. (R)	0.50	0.25
pos_05	**Ich denke, bei diesem Produkt passt alles zusammen.**	0.66	0.44
pos_06	Ich denke, das Produkt ist in sich konsistent.	0.52	0.27
rummel_07	Ich kann mir vorstellen, dass die meisten Menschen den Umgang mit diesem Produkt sehr schnell lernen.	0.80	0.63
rummel_08	Ich fand das Produkt sehr umständlich zu nutzen. (R)	0.74	0.55
pos_07	Ich fand das Produkt sehr einfach zu benutzen.	0.89	0.80
pos_08	**Ich fand die Benutzung des Produkts intuitiv verständlich.**	0.85	0.72
rummel_09	Ich fühlte mich bei der Benutzung des Produkt sehr sicher.	0.74	0.54
gao_10	Ich musste eine Menge Dinge lernen, bevor ich mit diesem Produkt loslegen konnte. (R)	0.75	0.56
pos_09	Ich musste wenig lernen, bevor ich anfangen konnte das Produkt zu nutzen.	0.50	0.25
pos_10	**Ich konnte, ohne viel zu lernen, mit der Benutzung des Produkts beginnen.**	0.81	0.65

Note: Selected items for the SUS-DE-Pos are marked in bold and reverse-coded items with (R). PA1 = factor loadings; h2 = communality.

the item pos_02. Furthermore, this item was also preferred by the language expert over the other alternative.

- Item pos_04, given that factor loadings and communalities were slightly better than for the other alternative (pos_03).
- Item pos_05, because it was preferred by the language expert while also showing higher factor loadings and communalities in the EFAs compared to the other alternative (pos_06).
- Item pos_08 was selected over item pos_07 although the item pos_07 performed slightly better in the EFAs, given that the wording of item pos_07 is very close to the original item 3, especially in the version by Rummel (2016).
- Item pos_10, due to the higher factor loadings and communalities in both EFAs compared to the item pos_09.

5.2.4. Initial comparison of SUS versions

After selecting the five alternative items most suitable for a positive German SUS version, we wanted to see how versions including these items would compare to the originally proposed versions by Gao et al. (2020) and Rummel (2016). We thus employed a collection of statistical procedures to compare four SUS versions: Gao et al. (2020)'s original version (SUS-DE-Gao), a positive version of Gao et al. (2020) using the newly developed positive items (SUS-DE-Gao-Pos), Rummel (2016)'s original version (SUS-DE-Rummel), and a positive version of Rummel (2016) (SUS-DE-Rummel-Pos).

Table 4. Coefficients α and ω (including 95% CIs) from study one for the four SUS versions.

Scale version	coefficient α	coefficient ω
SUS-DE-Gao	0.90 [0.89, 0.91]	0.90 [0.89, 0.91]
SUS-DE-Gao-Pos	0.91 [0.90, 0.92]	0.92 [0.91, 0.92]
SUS-DE-Rummel	0.90 [0.89, 0.91]	0.91 [0.90, 0.92]
SUS-DE-Rummel-Pos	0.91 [0.90, 0.92]	0.91 [0.90, 0.92]

5.2.4.1. Reliability. We started by looking at internal consistency as an indicator of reliability using coefficients α

Table 5. Fit indices for CFA models of the SUS versions in study one.

Tested model	χ^2	df	*p*-value χ^2	RMSEA	SRMR	CFI	AIC	BIC
SUS-DE-Gao, single factor	490.35	35	<0.001	0.125	0.066	0.895	29497.78	29599.58
SUS-DE-Gao, two factors (positive/negative)	438.28	34	<0.001	0.118	0.064	0.908	29389.56	29496.45
SUS-DE-Gao-Pos, single factor	522.61	35	<0.001	0.127	0.066	0.905	28445.56	28547.36
SUS-DE-Rummel, single factor	504.61	35	<0.001	0.127	0.065	0.896	29764.06	29865.86
SUS-DE-Rummel, two factors (positive/negative)	432.75	34	<0.001	0.118	0.062	0.912	29682.33	29789.22
SUS-DE-Rummel-Pos, single factor	546.46	35	<0.001	0.129	0.068	0.903	28463.45	28565.26

(Cronbach, 1951) and ω (McDonald, 1999) including 95% confidence intervals, based on recommendations by Dunn et al. (2014). Results are reported in Table 4. All values were equal to or above 0.90, indicating very high internal consistency. Furthermore, all values for the positive versions were equal to or slightly higher than for the original SUS versions, while values were comparable for Rummel (2016) and Gao et al. (2020).

5.2.4.2. Model fit. In the next step, we worked with CFA, which tests how well hypothesized theoretical models fit the data collected with a scale (Brown, 2015). Thus, the method allowed us to see how well the data collected with the different SUS versions would fit the originally proposed unidimensional SUS model. In addition, we considered two-factor models accounting for item wording (positive vs. negative) as alternative models for the two original SUS versions (SUS-DE-Gao, SUS-DE-Rummel). The following criteria were considered as an indication of good model fit: Low χ^2 value and $p > 0.05$ for the χ^2 test, RMSEA < 0.06, SRMR $\leq$ 0.08 and $0.95 \leq$ CFI ≤ 1 (Hu & Bentler, 1999).[2] Results are listed in Table 5. Furthermore, as the five-point scale used for the SUS could be interpreted as ordinal, we also tested polychoric correlation matrices with a WLSMV-estimator in both this first and the second study. Results from these CFAs are provided in the supplementary materials on OSF.

Generally speaking, none of the fit indices except for the *SRMR* suggested a good fit for any of the models. The *CFI* values were slightly worse for the single-factor solutions for the original versions compared to the two-factor versions and the positive alternatives. The *RMSEA* values suggested a slightly better fit for the two-factor models, followed by the original single-factor models and the positive alternatives. However, no remarkable differences between the models were observed based on these results.

5.2.4.3. Product ratings. Next, we looked for significant differences in SUS ratings between the five products and the different scale versions. For this, we calculated a two-way Analysis of Variance (ANOVA) with the factors "product" (Excel, Word, Amazon, microwaves, Google search) and "SUS version" (SUS-DE-Gao, SUS-DE-Gao-Pos, SUS-DE-Rummel, SUS-DE-Rummel-Pos). Descriptive statistics for the ratings are provided in Table 6.

We found a statistically significant difference in SUS ratings by product ($F(4, 4780) = 876.62$, $p > 0.001$, $\eta^2_{partial} = 0.423$) and by SUS version ($F(3, 4780) = 2.63$, $p > 0.048$, $\eta^2_{partial} = 0.002$). The interaction was non-significant ($F(12, 4780) = 0.26$, $p > 0.994$, $\eta^2_{partial} < 0.001$). In addition, Kruskal–Wallis rank sum tests were performed because equal variances and normal distribution were not always given, which was likewise significant for the ratings by product ($\chi^2(4) = 1837.60$, $p > 0.001$) but not by SUS version ($\chi^2(3) = 6.30$, $p > 0.098$). Using Tukey's HSD Test for post-hoc comparison, we found significant individual differences ($p > 0.001$) between all of the products except for the pair "microwave-amazon" ($p > 0.051$) and no significant differences between any of the scale versions ($p_{min} = 0.159$ to $p_{max} = 0.999$, see OSF for detailed results). Results thus showed that while there were no notable differences in SUS ratings depending on the different scale versions, there were significant differences in ratings between the different products. All ratings are visualized in Figure 2.

When comparing the collected ratings to SUS scores for the five products reported in past work (Gao et al., 2020), it can be said that results matched, with the lowest ratings for Excel, followed by Word, with high ratings for Amazon and Google. Only values for the Microwave were higher in our sample across all four SUS versions compared to prior research.[3]

In addition, we also looked at if the single-item ratings differed between the five products. Results were comparable to those for the different SUS versions, with statistically significant differences in mean single-item ratings between multiple products (see OSF for details). Furthermore, the order of product ratings from lowest to highest was identical to those gathered with the four SUS versions.

5.2.4.4. Correlations among product ratings. Finally, we considered correlations among the product ratings to see how close the ratings of the different SUS versions for the various products were. All correlations are shown in Tables 7 and 8. Only participants who rated all five products were considered in this analysis ($n = 215$).

Results showed that correlations among SUS ratings were generally higher among ratings of the same product, compared to correlations with SUS ratings of another product. Furthermore, while the SUS ratings showed high correlations with the single-item ratings, these correlations were smaller than those among the SUS ratings of the same product. Overall, these results further demonstrated that the different SUS versions delivered comparable results.

5.3. Discussion

After the first study, we concluded that we now had an initial set of positive alternative items for the German SUS comparable to the original negatively formulated items used in Rummel (2016) and Gao et al. (2020), producing almost identical results. This new alternative set consisted of five

Table 6. Descriptive statistics from study one for the product ratings collected using the different SUS versions and the single item from Bangor et al. (2009).

	SUS-DE-Gao				SUS-DE-Gao-Pos				SUS-DE-Rummel				SUS-DE-Rummel-Pos				Single item			
	Mean	SD	min	max	Mean	SD	min	max	Mean	SD	min	max	Mean	SD	min	max	Mean	SD	min	max
Amazon	83.03	13.43	20.00	100.00	81.89	13.44	22.50	100.00	83.22	13.29	27.50	100.00	81.89	13.42	17.50	100.00	5.50	0.89	1.00	7.00
Google search	89.64	11.44	45.00	100.00	89.29	11.21	42.50	100.00	89.23	12.19	45.00	100.00	88.84	11.50	42.50	100.00	6.04	0.81	2.00	7.00
Word	74.56	15.34	17.50	100.00	73.86	15.07	22.50	100.00	74.36	15.92	20.00	100.00	73.94	15.22	25.00	100.00	5.32	0.90	2.00	7.00
Excel	53.81	18.19	7.50	100.00	52.24	18.62	2.50	100.00	54.09	18.51	7.50	100.00	52.28	18.52	2.50	100.00	4.51	1.14	1.00	7.00
Microwave	85.09	14.26	27.50	100.00	83.10	14.72	15.00	100.00	85.80	13.71	27.50	100.00	83.39	14.41	22.50	100.00	5.61	0.96	2.00	7.00

Note: Values could range from 0 to 100 for the SUS scores and from 1 to 7 for the single-item ratings.

positively formulated SUS items, which we used to replace the negative items in either German version of the SUS employed in the first study to create two positive versions: SUS-DE-Gao-Pos and SUS-DE-Rummel-Pos. As a next step, the psychometric quality of these German SUS items was to be investigated with a larger sample, comparing it to the other two versions of the SUS in German. In addition, we wanted to compare the two German SUS versions, SUS-DE-Gao and SUS-DE-Rummel, to see if one was superior to the other, given that results from the first study did not allow for picking a clear favorite. Furthermore, while using a set of products from past research allowed us to compare our results to ratings found in other studies of the SUS, we wanted to have participants interact with an actual product before rating it for increased ecological validity of our findings. We also assumed that some of the less favorable results concerning the SUS versions' psychometric quality, such as the unsatisfactory model fit in the CFAs, might be due to participants not interacting with an actual product prior to filling out the scale. Finally, we wanted to see how the different SUS versions would perform across two different device contexts. Namely, we wanted to see if the psychometric quality of the various German SUS versions would remain stable when employing them in both a desktop/laptop computer scenario and a mobile device context. Therefore, further research was necessary.

6. Study Two: Psychometric evaluation

Given the above-described limitations of the first study, we wanted to investigate and compare the psychometric quality of the four German SUS versions with a larger sample. In addition, we wanted to see how the SUS versions would perform when used to collect user ratings after interacting with existing websites rather than remembering past experiences with a product. Finally, we were interested in how the different SUS versions would perform across different device types, namely desktop/laptop computers vs. mobile devices. Thus, a second between-subjects online study with a considerably larger sample using a set of existing websites as stimuli was conducted. The procedure and design of this second study closely followed prior research on the development of a positive version of a visual aesthetics scale (Perrig, von Felten, et al., 2023). Alike to the first study, this study was preregistered on OSF (https://doi.org/10.17605/OSF.IO/Z7YBR).

6.1. Methods

6.1.1. Stimuli

Following the procedure from Perrig, von Felten, et al. (2023), we prepared a set of 12 existing Swiss websites with content in German as stimuli. We chose to use Swiss websites to minimize the risk that participants were already familiar with the sites while still allowing the German participants to be able to understand the website's language. In addition, participants were asked if they

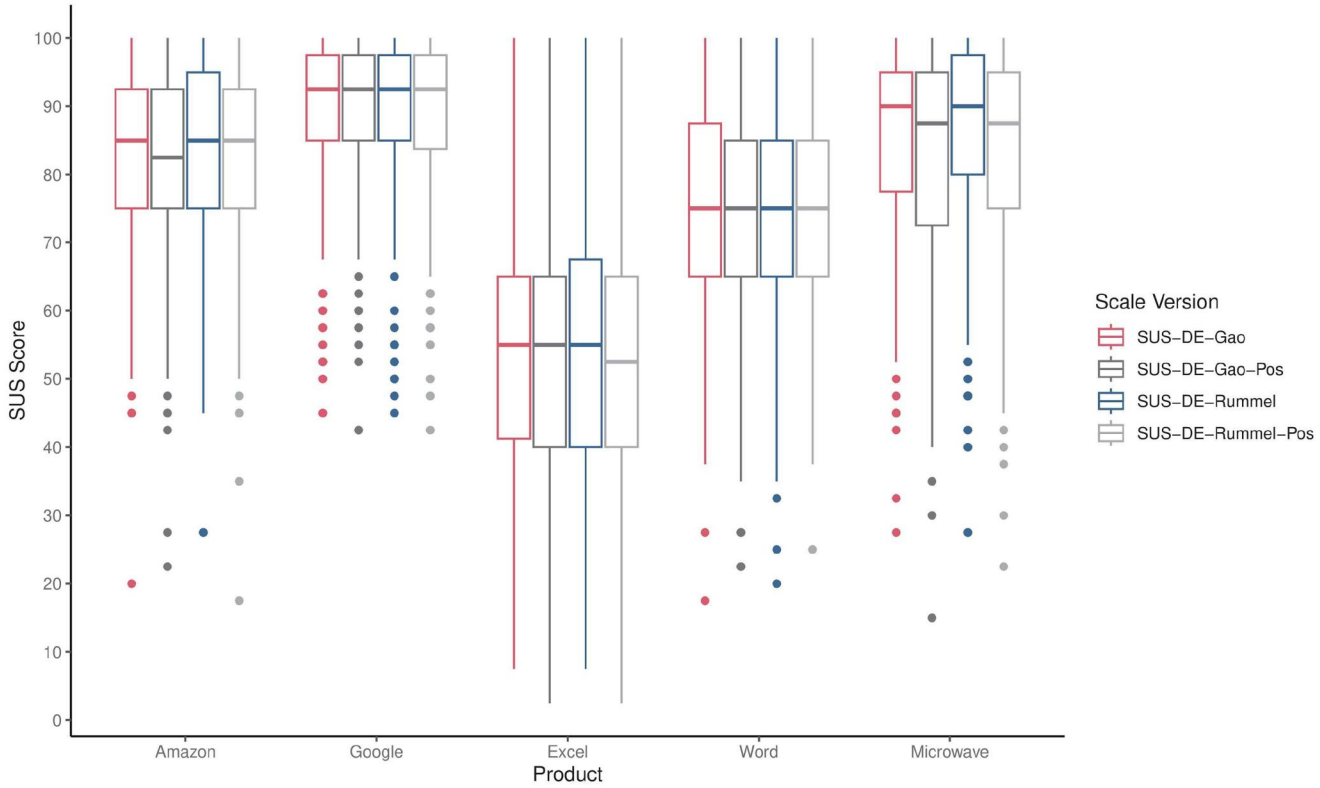

Figure 2. SUS scores for the five products rated in study one, using the four SUS versions.

already knew the visited site to allow us to consider possible differences among those unfamiliar with the site compared to previous users in our analyses. The 12 sites were chosen to cover six content areas (arts and entertainment, law & government, news & media publishers, science & education, food & drink, and lifestyle), which were taken from Similarweb.com to include various types of websites. One unpopular and one popular website was chosen for each content area using rankings from Similarweb.com. Popular sites were among the top 300 in Switzerland and highly ranked within their respective category, while unpopular sites had substantially lower ranks. All 12 websites used, including their content area and overall rank, are presented in Table 9.

To ensure that participants actually interacted with the websites, they were asked to answer two unique content-based questions developed for the individual sites. Of these questions, the first asked about the general content of the website (eg, news, weather, fashion), while the second one was a more site-specific task to be completed (eg, "Sie möchten die Website gerne finanziell unterstützen. Welche Abo-Möglichkeit gibt es?" ["You would like to support the website financially. What subscription option is available?"]). For each question, four answering options were provided, of which only one was correct. All tasks were designed to be comparable in difficulty by ensuring that answers could not be found on the landing page of the website and by requiring a minimum of one to four clicks to be completed. The supplementary materials on OSF contain all tasks. The mean percentage of correct responses across all items was 89.40% ($SD = 15.56\%$), with a range from 39.44% correct answers to 100.00%. Across the 12 websites, most participants (95.38%) reported not knowing the site they interacted with. Thus, responses from all participants, familiar and unfamiliar, were analyzed together.

6.1.2. Participants

For the study, 902 participants currently living in Germany were recruited on Prolific using the same criteria for screening and reimbursement as in the first study. Furthermore, they were not allowed to participate if they had already partaken in the first study. In addition, participants were recruited for one of two versions of the study, either completing it on a desktop/laptop computer or on a mobile device. During data cleaning, ten observations were removed based on the SRSI UseMe item, two due to an interrupted survey, one for getting both website questions wrong, one for not responding to both content questions, and one for indicating their current country of residence outside of Germany, going against the recruitment criteria. After data cleaning, a final sample of 887 respondents remained (428 women, 438 men, 21 non-binary people; mean age 28.41 years, $SD = 7.67$, min = 18, max = 64). Participants were evenly spread across the two SUS versions ($n = 446$ for Gao et al., 2020, $n = 441$ for Rummel, 2016), the two device types ($n = 447$ for desktop, $n = 440$ for mobile), and the 12 stimuli websites ($n_{min} = 71$, $n_{max} = 77$).

Table 7. Pearson's product–moment correlations from study one between the product ratings for Amazon, Google search, and Word from the SUS versions.

	Rummel Amazon	Rummel-Pos Amazon	Gao Amazon	Gao-Pos Amazon	SI Amazon	Rummel Google	Rummel-Pos Google	Gao Google	Gao-Pos Google	SI Google	Rummel Word	Rummel-Pos Word	Gao Word	Gao-Pos Word
Rummel-Pos Amazon	0.90****													
Gao Amazon	0.94****	0.86****												
Gao-Pos Amazon	0.88****	0.97****	0.90****											
SI Amazon	0.69****	0.71****	0.69****	0.70****										
Rummel Google	0.55****	0.57****	0.53****	0.55****	0.36****									
Rummel-Pos Google	0.50****	0.55****	0.48****	0.54****	0.35****	0.88****								
Gao Google	0.50****	0.49****	0.48****	0.48****	0.28****	0.91****	0.82****							
Gao-Pos Google	0.45****	0.49****	0.42****	0.49****	0.31****	0.84****	0.96****	0.85****						
SI Google	0.31****	0.40****	0.28****	0.36****	0.44****	0.53****	0.54****	0.49****	0.54****					
Rummel Word	0.30****	0.31****	0.28****	0.31****	0.15*	0.26****	0.20**	0.25****	0.18**	0.06				
Rummel-Pos Word	0.31****	0.38****	0.28****	0.37****	0.18**	0.29****	0.27****	0.25****	0.25****	0.13	0.91****			
Gao Word	0.32****	0.34****	0.32****	0.35****	0.18**	0.27****	0.20**	0.26****	0.18**	0.07	0.94****	0.89****		
Gao-Pos Word	0.32****	0.40****	0.31****	0.40****	0.20**	0.28****	0.26****	0.24****	0.23****	0.12	0.87****	0.97****	0.91****	
SI Word	0.16*	0.22**	0.15*	0.21**	0.20**	0.13	0.14*	0.11	0.13	0.12	0.65****	0.70****	0.66****	0.69****

Note: ****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

6.1.3. Measures

Participants rated the assigned stimuli website by responding to all items of one of the two SUS versions, Gao et al. (2020)'s version, Rummel (2016)'s version, in addition to the newly developed positive items. Furthermore, the single-item adjective rating of usability by Bangor et al., (2009) and German versions of the User Experience Questionnaire (UEQ, Laugwitz et al., 2008) and the Web-CLIC-S (Thielsch & Hirschfeld, 2021) were used to assess the SUS versions' convergent and divergent validity.

6.1.3.1. User Experience Questionnaire. In addition to one of the SUS versions, participants responded to the 26 items of the German version of the UEQ (Laugwitz et al., 2008), a semantic differential scale for the measurement of several UX-related constructs, both convergent and divergent to the SUS. The UEQ is comprised of 26 adjective pairs, to which responses are collected on a seven-point scale. After data collection, the original authors recommend re-coding all values to range from −3 to +3. Items are grouped across six dimensions, with four items each for the three goal-oriented pragmatic quality dimensions (ie, perspicuity, efficiency, dependability) and the two non-goal-directed hedonic quality dimensions (ie, stimulation, novelty), as well as six items for overall product attractiveness. Internal consistency was good for the overall UEQ ($\alpha = 0.95$, 95% CI[0.95, 0.96], $\omega = 0.95$, 95% CI[0.94, 0.95]), and for the individual subscales (values for α and ω from 0.77 to 0.93, see OSF for details). We decided to have respondents fill out the UEQ because it covers a broad selection of constructs relevant to UX. The hedonic quality constructs of the UEQ were expected to correlate to a lesser extent with the SUS ratings than the attractiveness rating, with the highest correlations expected for the pragmatic quality dimensions.

6.1.3.1.1. Web-CLIC-S. As a further measure, participants were asked to judge the stimuli websites' content using the German version of the Web-CLIC-S (Thielsch and Hirschfeld, 2021), a four-item short version of the Web-CLIC scale. Across its four items, the Web-CLIC-S measures four content areas: clarity, likability, informativeness, and credibility. Responses are collected on a seven-point Likert-type scale. Mean values across the four items were formed after data collection to measure participants' subjective content perception of the stimuli websites. Internal consistency for the scale was good ($\alpha = 0.78$, 95% CI[0.75, 0.80], $\omega = 0.79$, 95% CI[0.77, 0.82]). Lower correlations between the SUS ratings and the Web-CLIC-S were expected compared to the ratings of pragmatic quality from the UEQ and the correlations among the three SUS versions.

6.1.4. Procedure

Two identical versions of the study were created, one to be filled out on desktop/laptop computers and a second version for mobile devices. Recruitment for the two study versions was handled via two distinct calls for participants on Prolific, and the survey was set up to filter out participants accessing the study with the wrong device type.

Table 8. Pearson's product-moment correlations from study one between the product ratings for Excel and the microwave from the SUS versions.

	Rummel Excel	Rummel-Pos Excel	Gao Excel	Gao-Pos Excel	SI Excel	Rummel microwave	Rummel-Pos microwave	Gao microwave	Gao-Pos microwave
Rummel-Pos Excel	0.94****								
Gao Excel	0.96****	0.93****							
Gao-Pos Excel	0.91****	0.98****	0.94****						
SI Excel	0.70****	0.73****	0.72****	0.74****					
Rummel microwave	0.24***	0.19**	0.21**	0.17*	0.15*				
Rummel-Pos microwave	0.21**	0.21**	0.19**	0.19**	0.16*	0.91****			
Gao microwave	0.22**	0.18**	0.19**	0.17*	0.14*	0.94****	0.91****		
Gao-Pos microwave	0.21**	0.19**	0.18**	0.17*	0.15*	0.90****	0.98****	0.93****	
SI microwave	0.05	0.08	0.03	0.07	0.11	0.67****	0.73****	0.70****	0.74****

Note: ****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

Table 9. Websites used as stimuli in study two.

Nr.	Website Name	Link	Content Area	Rank
1	SRF Schweizer Radio und Fernsehen	https://www.srf.ch/	Arts & entertainment	14
2	Schweizerische Eidgenossenschaft	https://www.admin.ch/	Law & government	15
3	Neue Zürcher Zeitung	https://www.nzz.ch/	News & media publishers	6
4	ETH Zürich	https://ethz.ch/	Science & education	67
5	Betty Bossi	https://www.bettybossi.ch/	Food & drink	86
6	Globus	https://www.globus.ch/	Lifestyle	259
7	myfilm online	https://www.myfilm.ch/	Arts & entertainment	16,257
8	Kanton Uri	https://www.ur.ch/	Law & government	2,202
9	bajour.ch	https://bajour.ch/	News & media publishers	4,389
10	Freies Gymnasium Zürich	https://www.fgz.ch/	Science & education	239,023
11	Paul Ullrich AG	https://ullrich.ch/	Food & drink	4,725
12	PKZ	https://www.pkz.ch/	Lifestyle	2,095

Note: Website popularity ranks and content areas were retrieved on March 14. 2024 from https://www.similarweb.com/

Participants provided informed consent on the first page of the survey. Next, they were randomly assigned to one of the 12 websites and asked to explore the website while responding to questions related to the website content. Afterward, they were randomly assigned to one of the two SUS versions, filling out all ten items of that version plus the five positive alternative items. On the following pages of the survey, participants responded to the additional measures in randomized order. Next, participants provided demographic information and had the opportunity to give feedback before being redirected to Prolific for payment. Data quality was ensured following the same procedure as for the first study. The survey was presented to participants in German. Completing the survey took participants an average of 7.63 minutes ($SD = 3.45$, min = 2.70, max = 27.98). Figure 3 gives an overview of the study procedure.

6.1.5. Pre-study

Prior to preregistering the study, we tested the procedure, stimuli, and tasks with a small-sample pre-study ($N = 71$; $n = 36$ for desktop, $n = 35$ for mobile; five to seven participants per stimuli website). The goal of this pre-study was to examine if the majority of participants could complete the tasks and if there were any major issues with the study procedure. Most tasks were answered correctly by the participants, with a mean of 90.48% correct answers ($SD = 18.09\%$). However, some minor adjustments were made for the final study based on insights from this pre-study. Given these adjustments, the data from this pre-study was not used for the final analysis.

6.2. Results

The following section contains different forms of psychometric quality investigation for the four SUS versions: Gao et al. (2020)'s version, Rummel (2016)'s version, and versions of the German SUS using the newly developed positive items. All results were obtained using the statistical software R (R Core Team, 2024, version 4.4.0). The complete analysis can be found in the supplementary materials on OSF.

6.2.1. Item analysis

Starting with item analysis, descriptive statistics, item difficulty, item variance, discriminatory power, and inter-item correlations for all 30 SUS items were considered (10 original and 5 positive alternatives each for Rummel, 2016, and Gao et al., 2020). In summary, these analyses showed that while values sometimes differed between the different versions of the SUS items, most items exhibited values not considered to be overtly problematic. Exceptions were 13 out of 30 items exhibiting low item variance (< 1) and inter-item correlations turning out slightly higher than desirable, indicating that the items may be capturing only part of the construct (Piedmont, 2014). The analysis thus continued without removing any items but keeping those conspicuous items in mind.

6.2.2. Confirmatory factor analyses

Next, we used CFA to check how well the data collected with the different SUS versions would fit different theoretical models. Here, we considered unidimensional models for all four SUS versions while calculating additional two-factor

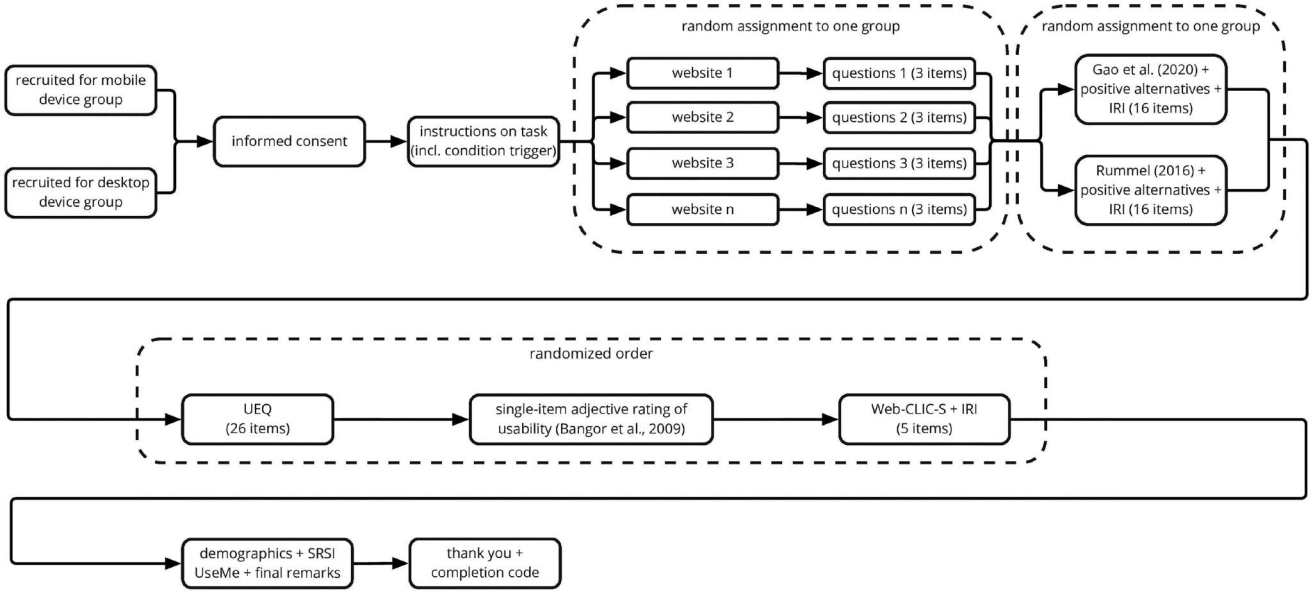

Figure 3. Flowchart depicting the procedure of study two.

Table 10. Fit indices for CFA models of the SUS versions in study two.

Tested model	χ^2	df	*p*-value χ^2	RMSEA	SRMR	CFI	AIC	BIC
SUS-DE-Gao, single factor	104.74	35	<0.001	0.074	0.048	0.958	10790.17	10872.18
SUS-DE-Gao, two factors (positive/negative)	97.67	34	<0.001	0.072	0.047	0.961	10783.86	10869.96
SUS-DE-Gao-Pos, single factor	138.76	35	<0.001	0.088	0.046	0.950	10386.44	10468.44
SUS-DE-Rummel, single factor	124.26	35	<0.001	0.087	0.054	0.948	10427.68	10509.46
SUS-DE-Rummel, two factors (positive/negative)	117.71	34	<0.001	0.085	0.052	0.952	10420.59	10506.46
SUS-DE-Rummel-Pos, single factor	94.78	35	<0.001	0.067	0.038	0.970	10220.54	10302.32

models accounting for positive vs. negative item wording for the two original SUS versions (SUS-DE-Gao, SUS-DE-Rummel). As in Study One, the criteria by Hu and Bentler (1999) were considered to judge model fit, and a robust maximum likelihood estimator was used because multivariate normality was not given. Results are listed in Table 10.

The χ^2 test was significant for all scale versions and models. However, the test is known to be influenced by larger sample sizes (> 200) and deviations from multivariate normality (Whittaker & Schumacker, 2022), both present in the current study. Therefore, we focused on the different fit indices to compare model fits. Still, it must be noted that compared to the first study, χ^2 values were substantially smaller despite the larger sample size, and model fits were generally improved compared to those in Study One.

Starting with the SUS-DE-Gao, both the unidimensional and the two-factor model were favored over the positive version by the *RMSR* and the *CFI* results, although the two-factor model fit the data slightly better. The *SRMR*, however, favored the SUS-DE-Gao-Pos. Concerning the versions based on Rummel (2016), the positive SUS-DE-Rummel-Pos exhibited the most favorable values out of all models, with a high *SRMR*, lowest *RMSEA*, and low *CFI*, followed by the original SUS-DE-Rummel, which showed comparable values for both the unidimensional and the two-factor models. Generally speaking, the SUS-DE-Rummel-Pos exhibited the most favorable values out of all, with only the *RMSEA* above the cutoff but still below 0.80, thus showing a reasonable error of approximation (MacCallum et al., 1996). Interestingly enough, results were flipped for the versions based on Gao et al. (2020), with the positive version performing slightly worse than the SUS-DE-Gao, with the unidimensional and the two-factor models.

Information criteria help compare model fits by balancing their complexity and goodness of fit, with lower values indicating better fit. We used both the Akaike information criterion (AIC, Akaike, 1987) and the Bayesian information criterion (BIC, Schwarz, 1978) to compare the model fits of the different CFAs. Results are also reported in Table 10 and favored the SUS-DE-Rummel-Pos over all other models, followed by the SUS-DE-Gao-Pos, and the two-factor and single-factor models for the SUS-DE-Rummel, with the worst results for both models of the SUS-DE-Gao (Brown, 2015).

6.2.3. Measurement invariance across device types

In addition to comparing multiple fit indices to judge model fit, we used scaled χ^2 difference tests to see if the model fit would differ significantly between the two device types, desktop and mobile. Across all four scale versions, none of the tests were statistically significant, suggesting that the model's structure, factor loadings, item intercepts, and item residual variances did not change substantially between the two device types for any of the four SUS versions. Thus, we conclude that configural, metric, scalar, and strict invariance was achieved and that the SUS' model was stable across device types. Results are reported in Table 11.

Table 11. Tests of measurement invariance across device types (desktop vs. mobile) for the CFA models of the SUS versions in study two.

Tested model	df	AIC	BIC	χ^2	χ^2-difference	df-difference	*p*-value χ^2
SUS-DE-Gao							
Configural invariance	70	10826	11072	176.30			
Weak invariance	79	10821	11030	189.02	12.40	9	0.192
Strong invariance	88	10810	10982	196.34	7.31	9	0.605
Strict invariance	98	10813	10944	219.43	11.04	10	0.355
SUS-DE-Gao-Pos							
Configural invariance	70	10424	10670	205.33			
Weak invariance	79	10419	10628	218.15	12.64	9	0.180
Strong invariance	88	10410	10582	227.16	8.94	9	0.443
Strict invariance	98	10409	10540	245.98	11.09	10	0.350
SUS-DE-Rummel							
Configural invariance	70	10478	10724	226.99			
Weak invariance	79	10465	10674	231.74	4.41	9	0.883
Strong invariance	88	10452	10624	236.64	4.92	9	0.842
Strict invariance	98	10446	10577	251.09	8.13	10	0.616
SUS-DE-Rummel-Pos							
Configural invariance	70	10265	10510	148.44			
Weak invariance	79	10258	10467	159.49	11.13	9	0.267
Strong invariance	88	10249	10420	168.18	9.00	9	0.437
Strict invariance	98	10239	10370	178.51	6.03	10	0.813

Note: The χ^2 column contains standard (non-scaled) test statistics. The χ^2 difference test used a Satorra and Bentler (2001) correction.

6.2.4. Differences in SUS scores

Next, the analysis looked for statistical differences in SUS ratings between the twelve stimuli websites to investigate the SUS versions' sensitivity and criterion-related validity (ie, are the different SUS versions able to distinguish between different stimuli). Descriptive statistics for all SUS scores, sorted by condition, can be found in Table 12, and a visualization of the website ratings with the four SUS versions is provided in Figure 4.

We started by looking for significant differences in SUS scores between the twelve website conditions, also considering the two device types (desktop vs. mobile) and the scale versions (SUS-DE-Gao, SUS-DE-Gao-Pos, SUS-DE-Rummel, SUS-DE-Rummel-Pos). We refrained from making predictions concerning possible differences between the different websites and device types. The primary purpose of the analyses here was to see if the four SUS versions would deliver comparable results. However, the factors "website" and "device type" were considered to look for possible interactions with the scale versions and to see how the magnitude of the effects of the different factors and interactions would compare.

A three-way ANOVA with the factors *device type*, *website condition*, and *scale version* found a significant difference in SUS scores between the device types ($F(1, 1678) = 7.97$, $p = 0.005$, $\eta^2_{partial} = 0.005$), the websites ($F(11, 1678) = 13.14$, $p = 0.001$, $\eta^2_{partial} = 0.079$), and the scale versions ($F(3, 1678) = 2.70$, $p = 0.044$, $\eta^2_{partial} = 0.005$). Furthermore, the interaction between device type and website condition was also significant ($F(11, 1678) = 2.32$, $p = 0.008$, $\eta^2_{partial} = 0.016$). Thus, results suggest that the device type and the scale version had a negligible effect on the SUS scores, in comparison to a medium-sized effect for the website condition and a small effect for the interaction between website condition and device type. Thus, based on the effect sizes, we concluded that the ratings differed between the stimuli websites, sometimes depending on whether participants interacted with the website on mobile or desktop. In contrast, ratings were comparable across the four SUS versions. In addition, we performed a Kruskal-Wallis rank sum test because equal variances and normal distribution were not given, which was likewise significant for the device types ($\chi^2(1) = 9.71$, $p = 0.002$), the website conditions ($\chi^2(11) = 141.70$, $p = 0.001$), and for the scale versions ($\chi^2(3) = 7.82$, $p = 0.050$).

Based on the results thus far, we concluded that the four SUS versions delivered comparable ratings. At the same time, the SUS-DE-Rummel-Pos was most favored based on the psychometric analyses, in particular, the CFA. We thus decided to focus our reporting on this version of the scale, in addition to the original version of it (ie, the SUS-DE-Rummel), for the remainder of the results. For completeness sake, all results for the versions based on Gao et al. (2020) are provided on OSF, although it has to be said that none of these results clearly favored the versions based on Gao et al. (2020) over those based on Rummel (2016).

6.2.5. Convergent and divergent validity

Two approaches were used for the assessment of convergent and divergent validity: First, Pearson's product-moment correlations between the SUS versions and other measures were calculated. Second, CFAs were performed with all scales included, calculating error-free correlations among them (Eid et al., 2017).

All correlations were as expected and pre-registered, thus favoring the SUS' convergent and divergent validity for both the SUS-DE-Rummel and the SUS-DE-Rummel-Pos (see Table 13 for exact values). Specifically, correlations among the SUS scores between the original and positive versions based on Rummel (2016) were very high ($r = 0.95$, 95% CI[0.95, 0.96]). In addition, both versions correlated highly with the single-item usability rating and the pragmatic quality score from the UEQ. In contrast, correlations were lower between the SUS versions and the UEQ's attractiveness score, and between the SUS scores and the Web-CLIC-S. Finally, correlations with the UEQ's hedonic quality score

Table 12. Descriptive statistics from study two for the four SUS versions by website condition (popular = 1–6, unpopular = 7–12) and device type (desktop = top, mobile = bottom).

		SUS-DE-Gao				SUS-DE-Gao-Pos					SUS-DE-Rummel				SUS-DE-Rummel-Pos			
website	*n*	*mean*	*SD*	*min*	*max*	*mean*	*SD*	*min*	*max*	*n*	*mean*	*SD*	*min*	*max*	*mean*	*SD*	*min*	*max*
desktop 1	20	70.75	19.42	30.0	92.5	66.00	21.34	25.0	92.5	18	71.39	16.94	37.5	97.5	70.28	14.92	37.5	87.5
desktop 2	16	73.12	15.75	32.5	92.5	72.50	13.60	40.0	92.5	22	73.41	16.23	37.5	92.5	69.77	16.11	30.0	87.5
desktop 3	24	65.10	19.83	22.5	95.0	63.44	20.69	22.5	97.5	14	74.82	14.36	52.5	95.0	73.93	15.98	45.0	97.5
desktop 4	12	70.62	19.43	40.0	97.5	66.88	19.75	37.5	97.5	27	69.81	21.64	20.0	100.0	68.43	19.02	22.5	100.0
desktop 5	14	70.54	17.38	37.5	100.0	66.61	19.45	32.5	100.0	21	71.31	15.24	47.5	95.0	69.17	16.42	32.5	92.5
desktop 6	26	66.83	19.84	20.0	97.5	65.19	19.12	22.5	97.5	12	66.04	22.19	22.5	97.5	65.21	23.92	20.0	95.0
desktop 7	18	72.92	20.28	17.5	92.5	70.69	20.93	17.5	92.5	18	79.03	13.99	37.5	95.0	75.00	13.99	40.0	97.5
desktop 8	17	63.09	16.55	35.0	90.0	60.88	17.23	30.0	87.5	20	63.12	18.37	25.0	90.0	63.75	19.07	25.0	92.5
desktop 9	19	58.42	16.16	32.5	90.0	57.76	18.35	30.0	100.0	18	56.94	20.17	0.0	95.0	53.06	19.51	0.0	95.0
desktop 10	23	67.83	20.08	12.5	95.0	68.15	20.45	7.5	95.0	15	77.83	12.85	52.5	100.0	77.50	11.61	57.5	100.0
desktop 11	15	75.67	16.30	42.5	92.5	74.17	17.62	40.0	92.5	20	73.12	21.38	12.5	95.0	70.50	21.89	7.5	95.0
desktop 12	21	75.95	15.50	42.5	95.0	72.02	17.24	32.5	92.5	17	81.03	14.79	50.0	97.5	77.35	16.33	42.5	97.5
mobile 1	20	75.00	15.52	40.0	92.5	73.88	14.29	45.0	92.5	13	74.42	18.68	37.5	92.5	72.50	13.35	50.0	87.5
mobile 2	19	62.89	15.99	25.0	82.5	60.66	15.18	32.5	82.5	20	70.50	19.84	35.0	97.5	67.50	18.25	37.5	92.5
mobile 3	19	63.68	18.77	17.5	92.5	62.11	18.55	17.5	90.0	16	70.94	20.25	32.5	97.5	68.28	21.29	30.0	92.5
mobile 4	21	72.86	17.49	37.5	97.5	71.19	17.81	35.0	97.5	16	75.00	11.25	50.0	87.5	71.25	10.12	45.0	82.5
mobile 5	17	65.74	19.92	27.5	92.5	63.97	18.20	22.5	92.5	20	60.50	18.58	25.0	92.5	58.12	20.82	17.5	90.0
mobile 6	22	70.68	20.59	17.5	90.0	69.32	20.79	20.0	90.0	14	64.11	23.44	10.0	97.5	64.64	22.12	17.5	100.0
mobile 7	19	74.87	14.10	32.5	95.0	69.34	14.01	30.0	92.5	18	75.69	14.29	40.0	92.5	76.39	12.58	42.5	97.5
mobile 8	15	62.83	16.98	40.0	97.5	61.17	17.78	37.5	97.5	19	56.18	20.99	25.0	92.5	57.89	19.94	25.0	92.5
mobile 9	17	53.53	18.71	17.5	82.5	53.24	18.87	17.5	77.5	23	59.57	15.31	35.0	87.5	56.74	16.64	30.0	87.5
mobile 10	13	60.77	14.56	42.5	92.5	61.15	16.38	35.0	90.0	24	65.10	18.44	32.5	95.0	61.46	18.44	32.5	90.0
mobile 11	21	68.10	15.87	35.0	90.0	67.14	17.45	30.0	90.0	18	68.89	16.70	35.0	95.0	68.19	18.29	32.5	95.0
mobile 12	18	75.97	12.55	45.0	92.5	72.08	11.35	47.5	90.0	18	75.14	17.18	45.0	100.0	72.64	16.39	50.0	97.5

Note: Participants filled out items for both the original and negative versions of either SUS variant. Thus, the sample size is only provided once each.

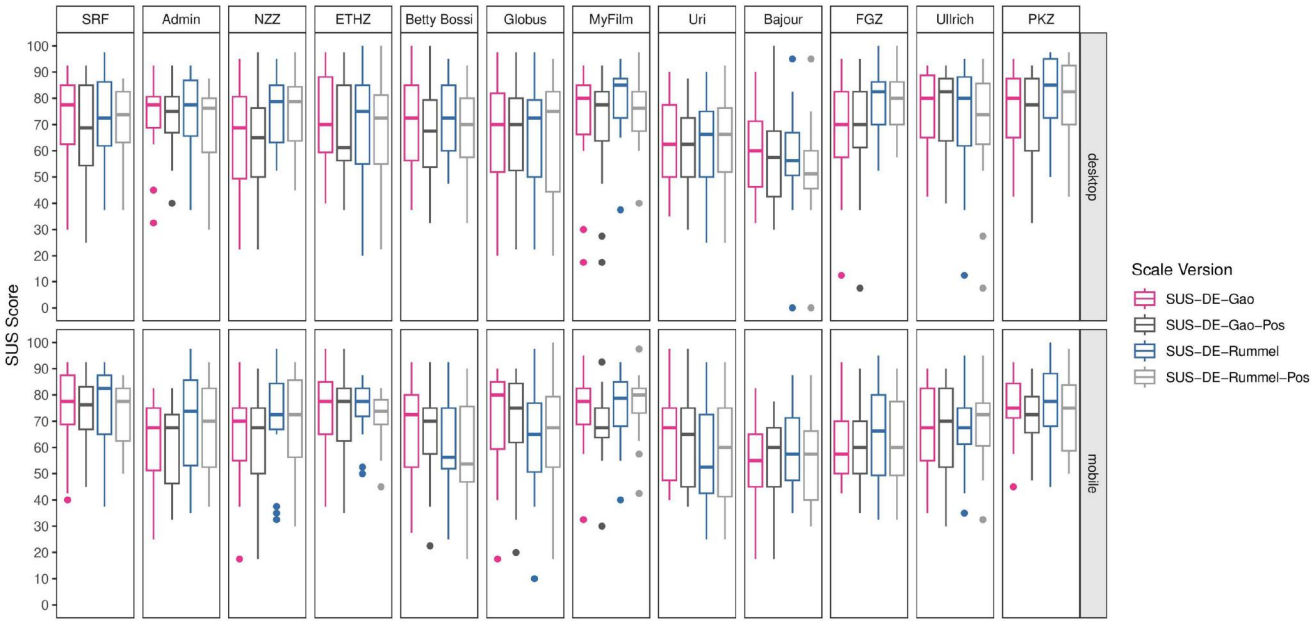

Figure 4. SUS scores for the twelve websites rated in study two, using the four SUS versions, separated by device type (desktop vs. mobile).

were lowest for both scale versions. Comparable results to the Pearson's correlations were found with the large CFA, thus providing additional evidence for the scales' convergent and divergent validity. These measurement-error-free correlations are provided in the supplementary materials on OSF.

6.2.6. Reliability

For both the SUS-DE-Rummel (α = 0.90, 95% CI[0.89, 0.92], ω = 0.91, 95% CI[0.90, 0.92]) and the SUS-DE-Rummel-Pos (α = 0.91, 95% CI[0.90, 0.92], ω = 0.91, 95% CI[0.90, 0.93]) coefficient α and ω values indicated high internal consistency.

6.2.7. Answering mistakes by participants

Finally, we investigated whether participants would exhibit answering behavior that could be considered as mistakes. Following the procedure reported in Sauro and Lewis (2011), we flagged all participants who agreed with both the positive and negative items of a scale or disagreed with both

Table 13. Pearson's correlations (including 95% CIs) from study two for the SUS versions based on Rummel (2016) with the convergent and divergent scales.

	SUS-DE-Rummel	SUS-DE-Rummel-Pos
UEQ Attractiveness	0.71 [0.66, 0.75]	0.71 [0.67, 0.76]
UEQ Pragmatic Quality	0.85 [0.83, 0.88]	0.85 [0.83, 0.88]
UEQ Hedonic Quality	0.42 [0.35, 0.50]	0.44 [0.36, 0.51]
Web-Clic-S	0.68 [0.63, 0.73]	0.70 [0.65, 0.74]
Single Item Usability	0.78 [0.74, 0.82]	0.78 [0.74, 0.81]

types of items. In this way, all participants who exhibited at least three responses indicating an agreement to positively and negatively worded items or three responses showing disagreement to positively and negatively worded items were marked as mistaken. While 8 out of 441 participants (1.81%) were flagged for the SUS-DE-Rummel-Pos, 17 out of 441 (3.85%) were flagged for the SUS-DE-Rummel. Thus, more than twice as many participants exhibited an inconsistent response behavior that could be considered mistaken for the original version containing negative items compared to the positive SUS.

7. General discussion

The present work developed a German version of the SUS unburdened by reverse-coded items. While reverse-coded items are intended to counteract undesirable response behavior, results from past research have shown these items to do more harm than good (Dalal & Carter, 2014; Sauro & Lewis, 2011). These harms include lower scale reliability (Barnette, 2000), distortions of factor structures (Schmitt & Stuits, 1985), a reduction in cross-cultural applicability (Wong et al., 2003), as well as the risk of misinterpretation or mistakes by participants (Sauro & Lewis, 2011). Thus, the present work set out to improve the quality of the German SUS by developing a version containing only positive items, which was developed and then validated throughout two studies. As a first step, an initial set of positively formulated alternatives for the five reverse-coded items of the SUS were created based on the original SUS (Brooke, 1996), two previously proposed German versions of the SUS by Gao et al. (2020) and Rummel (2016), and the English positive version of the SUS by Sauro and Lewis (2011). A language expert then reviewed these items for further refinement. Next, a first study asking participants to rate recalled experiences with different products showed that most candidate items were viable alternatives, performing comparably to the original SUS items. Thus, five positive alternatives, one for each reverse-coded SUS item, were chosen based on the results from the first study and the language evaluation. Finally, a second study using existing websites was conducted to evaluate the psychometric quality of four different SUS versions, comparing both original and positive versions of the scale based on Gao et al. (2020) and Rummel (2016). Findings from this second study showed that the positive version based on Rummel (2016), henceforth called the SUS-DE-Pos, delivered the most favorable psychometric results. The other three scale versions achieved slightly worse but still satisfactory psychometric values, comparable to values from past work on other versions of the SUS (Gao et al., 2020; Lewis & Sauro, 2017).[4] Furthermore, model fits were generally improved compared to those in the first study. In this regard, it is plausible to assume that the actual interaction with the websites, in contrast to a recalled experience, had a favorable effect on the model fits. This difference between recalled and actual interaction is something to keep in mind when working with the SUS.

7.1. Recommendations for research and practice

Based on the present results, we recommend that researchers who want to use a German version of the SUS work with the newly developed SUS-DE-Pos, a positive version of the scale based on past work by Rummel (2016). We make this recommendation for multiple reasons: First, out of the four investigated German versions of the SUS, the SUS-DE-Pos based on Rummel (2016) exhibited the most satisfactory results, particularly concerning model fit in the CFA reported in Study Two. In addition, other analyses in the present work, such as analyses of possible significant differences among the SUS scores collected with the different scales and correlations among these scores, suggested that all four scale versions deliver results that are closely related to one another. Furthermore, the fact that the SUS-DE-Pos showed more favorable psychometric values compared to its original version is not surprising, given past reports on problems caused by negatively formulated items in past research. In particular, the slight improvement in internal consistency and the better model fit for the SUS-DE-Pos compared to the other German SUS versions is in line with past research suggesting a negative effect of reverse-coded items on scale reliability (Barnette, 2000) and factor structures (Schmitt & Stuits, 1985). Regarding the cross-cultural applicability of the SUS, a concern also related to negatively worded items (Wong et al., 2003), results further demonstrated that the German versions of the SUS are of comparable psychometric quality to the original English version, and when considering the product ratings from Study One, deliver similar results. Not only have reverse-coded items been shown to negatively affect the psychometric quality of survey scales (eg, Perrig, von Felten, et al., 2023; Salazar, 2015), they can also lead participants to misrespond if their wording leaves room for interpretation, resulting in comprehension issues among respondents (Sauro & Lewis, 2011; Weijters & Baumgartner, 2012). Evidence of such misresponses was also observed in the results of Study Two in the present work, with more than twice as many participants providing inconsistent responses to the original version of the SUS based on Rummel (2016) compared to the newly developed positive version. This leads us to the second reason for advising HCI researchers and practitioners to use this version of the scale: it allows them to avoid negatively formulated items and, thus, the issues associated with their use. This also reflects the recommendation by Sauro and Lewis (2011) to avoid negative items in HCI scales, which researchers and practitioners can now follow when using the SUS in German.

Should HCI researchers for any reason, such as consistency with (their own) prior work, want to stay close to the original SUS by Brooke (1996), thus not avoiding negatively formulated items, we recommend they use the version proposed by Rummel (2016), referred to as SUS-DE-Rummel in the present article. We make this recommendation mainly for the sake of consistency, avoiding that too many different German versions of the SUS are being used across the research field. Thus, despite the slightly better CFA results for the SUS versions based on Gao et al. (2020), we recommend working with the SUS-DE-Rummel given that its wording is closer to the best-performing scale in the second study, the SUS-DE-Pos, with the exception of the negative item formulation. We see this recommendation as justified given that the SUS-DE-Rummel provided satisfactory results in the psychometric analyses, which are furthermore comparable to results from past work on other versions of the SUS (Gao et al., 2020; Lewis & Sauro, 2017). In addition, all four versions of the SUS investigated in the present work provided ratings for the stimuli websites, which were very close to each other. Furthermore, the SUS version by Rummel (2016) has been deemed to use more natural wording compared to the one by Gao et al. (2020) in a prior review comparing different German SUS versions (Brix et al., 2023).

Finally, the present work also looked at the theoretical model behind the German versions of the SUS, particularly considering whether a two-dimensional model accounting for positive vs. negative item wording would fit the data better than the originally proposed unidimensional model. Hence, we looked at alternative models for those scale versions containing negative items (ie, SUS-DE-Gao, SUS-DE-Rummel), concluding that two-factor models resulted in an improved model fit. While for some survey scales, this might lead researchers to assume a two-dimensional model, for the SUS, this makes limited sense theoretically, as already pointed out by Lewis and Sauro (2017). Negative items are known to distort factor structures of scales by creating artificial factors that go beyond what might be reasonably assumed based on theoretical considerations (Lindwall et al., 2012; Zhang et al., 2016). Given this line of theoretical reasoning, alongside the fact that the psychometric results in the present work were still acceptable for the unidimensional models, we recommend using a unidimensional model when working with any version of the SUS, including those containing negative items, summarizing responses into a singular SUS score of perceived usability, as recommended by Brooke (1996).

7.2. Stability across device types

Another question that interested us was whether the model fit of the different SUS versions remains stable across the two device settings of desktop/laptop computers vs. mobile devices, ie, whether the scale would retain measurement invariance across the two device types. Given the popularity of smartphone devices (Tenzer, 2024), it is essential for HCI researchers to have survey scales that remain stable across various device contexts. Given the differences between mobile and desktop devices (Nielsen & Budiu, 2013), results from one context cannot just be assumed to be transferable to the other. Thus, research is needed that investigates if and how findings found in a desktop context also apply to a mobile setting (eg, Perrig, Ueffing, et al., 2023), or vice-versa. The same also goes for other contexts beyond mobile or desktop. For such research to be able to collect robust data, we have to be sure that our measures, such as the SUS, remain stable across varying contexts of interaction. To answer this question, we considered both results from χ^2 difference tests as well as an ANOVA where the type of device was included as a possible influencing factor on the collected SUS scores. While the χ^2 difference tests suggested that the model fit remained constant across the two device types for all four scale versions, results from the ANOVA showed that the recorded SUS scores for the stimuli websites differed between the device types. However, concerning the latter, it is reasonable to assume that the difference in SUS ratings did not result from differences in the psychometric quality of the measurement instrument used but rather from differences in perceived usability between the desktop and mobile versions of the stimuli sites used in Study Two. Thus, while the present results support the measurement invariance of all four investigated SUS versions, they also show that the investigated SUS versions are able to pick up differences in perceived usability between the desktop and mobile versions of websites. Still, we see an opportunity for future research to further delve into this aspect of the SUS' performance across device types, as well as across additional types of products and systems beyond desktop and mobile websites.

7.3. Administering the SUS-DE-Pos

Given that the German versions of the SUS are based on the original scale, researchers should follow the original guidelines from Brooke (1996) describing how the SUS should be administered and, in particular, how the resulting SUS score needs to be calculated. Responses to the items should be collected using the 5-point Likert-type scale ranging from 1 ("Stimme überhaupt nicht zu") to 5 ("Stimme vollständig zu") also employed in the present work. Table 14 contains the complete SUS-DE-Pos, including instructions and answering options. The placeholder [X] should be replaced, either with the correct declination of "das Produkt" ("the product"), "das System" ("the system"), "die Website" ("the website"), or with the actual name of the product to be rated. After data collection, items are summarized like the original SUS by transforming them into scores ranging from 0 to 100. In cases where the positive version of the SUS is used, no items need to be reversed during this process. Following the original recommendations by Brooke (1996), we suggest that researchers use the SUS mostly after participants had the opportunity to interact with a system but before any debriefing or discussion about the experienced interaction. In this regard, the SUS can either be used to assess the overall perceived usability of the current

Table 14. The SUS-DE-Pos, for use in a study.

Bitte beurteilen Sie auf einer Skala von 1 ("Stimme überhaupt nicht zu") bis 5 ("Stimme vollständig zu") inwieweit Sie den folgenden Aussagen in Bezug auf [X] zustimmen. Bitte denken Sie nicht zu lange über jede Aussage nach und stellen Sie sicher, dass Sie zu allen Aussagen eine Antwort geben.

	Stimme überhaupt nicht zu 1	2	3	4	Stimme vollständig zu 5
Ich denke, dass ich [X] gerne häufig benutzen würde.	O	O	O	O	O
Ich fand die Komplexität [X] angemessen.	O	O	O	O	O
Ich fand [X] einfach zu benutzen.	O	O	O	O	O
Ich glaube, dass ich [X] ohne Unterstützung einer technisch versierten Person benutzen könnte.	O	O	O	O	O
Ich fand, die verschiedenen Funktionen in [X] waren gut integriert.	O	O	O	O	O
Ich denke, bei [X] passt alles zusammen.	O	O	O	O	O
Ich kann mir vorstellen, dass die meisten Menschen den Umgang mit [X] sehr schnell lernen.	O	O	O	O	O
Ich fand die Benutzung [X] intuitiv verständlich.	O	O	O	O	O
Ich fühlte mich bei der Benutzung [X] sehr sicher.	O	O	O	O	O
Ich konnte, ohne viel zu lernen, mit der Benutzung [X] beginnen.	O	O	O	O	O

version of a system in terms of a summative evaluation, or to investigate and compare the perceived usability of different parts or versions of a system as part of an iterative design process and formative evaluations. In the present study, we validated the German SUS versions for use in rating both desktop and mobile versions of websites. Thus, research can use the scale confidently in these settings. However, the SUS and other usability evaluations are being used across varying contexts, such as a mental health setting (Hvidt et al., 2020) or government work (Gulliksen, 2021). Given the highly context-dependent nature of HCI research, it cannot be excluded that the context might impact the quality of the SUS-DE-Pos. We recommend that the psychometric quality of the scale is investigated again when working in other settings. Suppose such an investigation is impossible, for example, due to small sample sizes. In that case, we encourage researchers to share their data so future work can aggregate SUS data collected in German for additional psychometric quality investigations. Given that all SUS versions employed in the present study are based on prior work by Gao et al. (2020) and Rummel (2016), researchers should cite these sources alongside the present work when using any of the scales, in addition to citing the original paper on the SUS by Brooke (1996).

7.4. Limitations and future work

We used crowd-sourced samples for both reported user studies. Thus, it is not yet clear if the different German SUS versions hold up in their psychometric quality with other samples, such as students or volunteers. Nevertheless, crowd-sourced samples have been shown to be of comparable quality to others (Buhrmester et al., 2011). Furthermore, the participants in both studies were recruited to have Germany as their current country of residence. It remains to be investigated whether the German versions of the SUS also hold up with other German-speaking samples, such as from Austria or Switzerland, with German speakers outside Europe, or with people in German-speaking countries who do not speak German fluently. While the products rated in the first study were selected to allow for a comparison of the present results to past work, we acknowledge that other types of technology are increasingly important. Thus, we encourage future research to investigate how well the SUS-DE-Pos performs in such novel settings. In addition, the stimuli websites used in the second study were carefully selected to cover a broad spectrum of possible content areas. However, participants were purposefully presented with websites with which they were likely unfamiliar. Thus, future work could also work with websites with which participants are familiar, as well as other stimuli that go beyond websites. Further, although the first study considered a broad range of products, the ratings were based solely on remembered past interactions. While this procedure was adapted from previous work on the SUS (Gao et al., 2020; Kortum & Bangor, 2013), we addressed this limitation in the second study by having participants fill out the SUS versions directly after an interaction with a product. However, our selection of stimuli in the second study was limited to websites. Thus, future work should see how the German SUS versions investigated here perform after interacting with different types of products and systems, going beyond the products and websites used in the present work. In addition, different contexts should be considered in future work investigating the quality of the SUS-DE-Pos, such as comparing the same kind of system or device used both in a work and a leisure setting. Finally, the tasks used in the second study only had participants interact with the stimuli website over a limited time frame and not, for example, throughout multiple sessions. In genuine usage, participants are expected to return to specific products they use, return more than once, and use them over more extended periods. It thus is an open question if the German SUS versions can measure usability over time.

8. Conclusion

Two alternative versions of the SUS in German, which do not use reverse-coded items, were developed and validated through two pre-registered online studies. In addition, these two versions were compared to previously recommended German SUS versions, one by Gao et al. (2020) and one by Rummel (2016), on which they were based. Results showed that a positive version of the SUS based on Rummel (2016),

referred to in the present work as the SUS-DE-Pos, achieved the most favorable results. The other SUS versions delivered slightly worse but overall still satisfactory and comparable results. Overall, the present work provides German-speaking HCI researchers and practitioners with new insights into the psychometric quality of different SUS versions in German, helping them pick the right scale for their research. Based on the present results, we recommend using the newly developed positive version of the SUS based on Rummel (2016), the SUS-DE-Pos, or if researchers want to work with negatively formulated items, employing the originally proposed version by Rummel (2016), referred to in the present work as the SUS-DE-Rummel.

Notes

1. Demographic information was only asked of participants to allow for some comparability of our sample to others. However, participants were allowed to skip providing demographic information. Furthermore, in acknowledging societal discussions concerning gender categories, we allowed participants to select one of three options (woman, man, non-binary), state that they prefer not to disclose their gender, or self-describe their gender. The country of residence was asked from participants to ensure that they met our inclusion criteria.
2. *RMSEA* = Root Mean Square Error of Approximation; *SRMR* = Standardized Root Mean Square Residual; *CFI* = Comparative Fit Index.
3. Mean SUS scores reported for the five products in Gao et al. (2020), using the German version of the SUS, were as follows: 56.7 for Excel ($SD = 19.5$), 72.6 for Word ($SD = 17.0$), 83.6 for Amazon ($SD = 12.7$), 75.2 for Microwave ($SD = 17.5$), and 90.6 for Google ($SD = 9.6$).
4. In Lewis and Sauro (2017), fit measures for the English SUS using a unidimensional model showed a *CFI* of 0.799 and an *RMSEA* of 0.190, compared to a *CFI* of 0.958 and an *RMSEA* of 0.088 for a two-dimensional model accounting for item tone. In Gao et al. (2020), *RMSEA* values for a unidimensional model for the German SUS ranged from 0.093 to 0.134, and *SRMR* values ranged from 0.052 to 0.104.

Acknowledgments

Special thanks to Dr. Phil. Beatrice Mall-Grob for her language expertise concerning the development of the positive SUS items in German.

Author contributions

SP and NvF developed the initial item set with inputs from BV and KO. SP, NvF, and BV contributed to the conception and design of the item development and first study. SP and NvF designed the second study, selected the set of stimuli websites, and designed the tasks with inputs from BV. SP and NvF performed the statistical analysis for both studies. SP wrote the first draft of the manuscript. All authors contributed to the manuscript revision, reading and approving the submitted version.

Ethical approval

Both studies were reviewed and approved by the ethics committee of the Faculty of Psychology of the University Basel (approval number 002-24-1). In both studies, all respondents gave informed consent before proceeding with the online survey.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was financed entirely by our research group, as we received no additional funding. The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

ORCID

Sebastian A. C. Perrig http://orcid.org/0000-0002-4301-8206
Nick von Felten http://orcid.org/0000-0003-0278-9896
Beat Vollenwyder http://orcid.org/0000-0002-9516-2680
Klaus Opwis http://orcid.org/0000-0003-0509-8070

Data availability statement

The data, complete analysis (R scripts), and supplementary materials (printouts of online surveys) for this study are available on OSF (https://doi.org/10.17605/OSF.IO/XFNB4).

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332. https://doi.org/10.1007/BF02294359

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: A critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2689–2698). Association for Computing Machinery.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*(3), 361–370. https://doi.org/10.1177/00131640021970592

Borsci, S., Federici, S., Malizia, A., & De Filippis, M. L. (2019). Shaking the usability tree: Why usability is not a dead end, and a constructive way forward. *Behaviour & Information Technology*, *38*(5), 519–532. https://doi.org/10.1080/0144929X.2018.1541255

Brix, T. J., Janssen, A., Storck, M., & Varghese, J. (2023). Comparison of German translations of the system usability scale–which to take? German medical data sciences 2023—science. Close to people. In R. R. Ohrig (Eds.). *Studies in health technology and informatics* (pp. 96–101, Vol. 307). https://doi.org/10.3233/SHTI230699

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

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. (2nd ed.). The Guilford Press.

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, *2*, 100022. https://doi.org/10.1016/j.metip.2020.100022

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science: a Journal of the Association for Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q., & Yammiyavar, P. (2009, July). Cultural cognition in usability evaluation. *Interacting with Computers*, *21*(3), 212–220. https://doi.org/10.1016/j.intcom.2009.05.003

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Dalal, D. K., & Carter, N. T. (2014). Negatively worded items negatively impact survey research. In *More statistical and methodological myths and urban legends* (1st ed., pp. 112–132). Routledge.

de Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710. https://doi.org/10.1080/02664763.2011.610445

DeVellis, R. F. (2017). *Scale development: Theory and applications.* (4th ed., Vol. 26). SAGE Publications, Inc.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology (London, England: 1953)*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden [Statistics and research methods]* (5th ed.). Beltz.

Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology* (1st ed.). SAGE Publications, Ltd.

Gao, M., Kortum, P., & Oswald, F. L. (2020). Multi-language toolkit for the system usability scale. *International Journal of Human–Computer Interaction*, *36*(20), 1883–1901. https://doi.org/10.1080/10447318.2020.1801173

Gulliksen, J (2021). Digital work environment rounds – Systematic inspections of usability supported by the legislation. In C. Ardito, et al. (Eds.), *Human–computer interaction – Interact 2021.* (pp. 197–218). Springer International Publishing.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Prentice Hall.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human–Computer Studies*, *64*(2), 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human–Computer Interaction*, *32*(1), 51–62. https://doi.org/10.1080/10447318.2015.1087664

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

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 Ergonomics of human-system interaction—Part 11: Usability: Definitions and concepts [Computer software manual]. Vernier, Geneva, Switzerland. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

Kelley, T. L. (1927). *Interpretation of educational measurements* (1st ed.). World Book Co.

Kline, P. (1994). *An easy guide to factor analysis.* Routledge.

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., & Bangor, A. (2013). Usability ratings for everyday products measured with the system usability scale. *International Journal of Human–Computer Interaction*, *29*(2), 67–76. https://doi.org/10.1080/10447318.2012.681221

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

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *HCI and usability for education and work* (pp. 63–76). Springer.

Lewis, J. R. (2018a). Is the report of the death of the construct of usability an exaggeration? *Journal of Usability Studies*, *14*(1), 1–7.

Lewis, J. R. (2018b). 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., & Sauro, J. (2009). The factor structure of the system usability scale. In M. Kurosu (Ed.), *Human centered design.* (pp. 94–103). Springer.

Lewis, J. R., & Sauro, J. (2017). Revisiting the factor structure of the system usability scale. *Journal of Usability Studies*, *12*(4), 183–192.

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, *94*(2), 196–204. https://doi.org/10.1080/00223891.2011.645936

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

McDonald, R. P. (1999). *Test theory: A unified treatment.* (1st ed.). Psychology Press.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Nielsen, J., & Budiu, R. (2013). *Mobile usability*. New Riders.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018, March). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Perrig, S. A. C., Aeschbach, L. F., Scharowski, N., von Felten, N., Opwis, K., & Brühlmann, F. (2024). Measurement practices in user experience (UX) research: A systematic quantitative literature review. *Frontiers in Computer Science*, *6*, 1368860. https://doi.org/10.3389/fcomp.2024.1368860

Perrig, S. A. C., Ueffing, D., Opwis, K., & Brühlmann, F. (2023). Smartphone app aesthetics influence users' experience and performance. *Frontiers in Psychology*, *14*, 1113842. https://doi.org/10.3389/fpsyg.2023.1113842

Perrig, S. A. C., von Felten, N., Honda, M., Opwis, K., & Brühlmann, F. (2023). Development and validation of a positive-item version of the Visual Aesthetics of Websites Inventory: The VisAWI-pos. *International Journal of Human–Computer Interaction*, *40*(20), 6622–6646. https://doi.org/10.1080/10447318.2023.2258634

Perrig, S. A. C., Scharowski, N., Brühlmann, F., von Felten, N., Opwis, K., & Aeschbach, L. F. (2024). *Independent validation of the player experience inventory: Findings from a large set of video game players.* In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3613904.3642270

Pettersson, I., Lachner, F., Frison, A.-K., Riener, A., & Butz, A. (2018). A Bermuda triangle? A review of method application and triangulation in user experience evaluation. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–16). Association for Computing Machinery.

Piedmont, R. L. (2014). Inter-item correlations. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 3303–3304). Springer Netherlands.

R Core Team (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. https://www.R-project.org/

Rummel, B. (2016). System Usability Scale – jetzt auch auf Deutsch [System usability scale – Now also in German]. https://blogs.sap.com/2016/02/01/system-usability-scale-jetzt-auch-auf-deutsch/

Salazar, M. S. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192–199.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Sauro, J., & Lewis, J. R. (2011). *When designing usability questionnaires, does it hurt to be positive?*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2215–2224). Association for Computing Machinery. New York, NY. https://doi.org/10.1145/1978942.1979266

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373. https://doi.org/10.1177/014662168500900405

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. http://www.jstor.org/stable/2958889 https://doi.org/10.1214/aos/1176344136

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2021). January) Pre-registration: Why and How. *Journal of Consumer Psychology*, *31*(1), 151–162. https://doi.org/10.1002/jcpy.1208

Suárez-Álvarez, 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

Tenzer, F. (2024). Anzahl der Smartphone-Nutzer weltweit von 2016 bis 2022 und Prognose bis 2028 [Number of smartphone users worldwide from 2016 to 2022 and forecast to 2028]. https://de.statista.com/statistik/daten/studie/309656/umfrage/prognose-zur-anzahl-der-smartphone-nutzer-weltweit/

Thielsch, M. T., & Hirschfeld, G. (2021). Quick assessment of web content perceptions. *International Journal of Human–Computer Interaction*, *37*(1), 68–80. https://doi.org/10.1080/10447318.2020.1805877

Tractinsky, N. (2018). The usability construct: A dead end? *Human–Computer Interaction*, *33*(2), 131–177. https://doi.org/10.1080/07370024.2017.1298038

van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*(2), 89–99. https://doi.org/10.1027/1016-9040.1.2.89

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737–747. https://doi.org/10.1509/jmr.11.0368

Whittaker, T. A., & Schumacker, R. E. (2022). *A beginner's guide to structural equation modeling*. (5th ed.). Routledge.

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, *30*(1), 72–91. https://doi.org/10.1086/374697

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PloS One*, *11*(6), e0157795. https://doi.org/10.1371/journal.pone.0157795

About the authors

Sebastian A. C. Perrig is a researcher in the Human–Computer Interaction research group at the Center for General Psychology and Methodology at the University of Basel, Switzerland. His research interests include user and player experience, scale development and validation, visual aesthetics, online survey research, and online data quality.

Nick von Felten is an MSc student and student research assistant in the Human–Computer Interaction research group at the Center for General Psychology and Methodology at the University of Basel, Switzerland. His research interests include human-centered artificial intelligence, questionnaire development, statistical methods, and user experience.

Beat Vollenwyder is a user experience architect at Swiss Federal Railways and an external lecturer in the Human–Computer Interaction research group at the Center for General Psychology and Methodology at the University of Basel. He is particularly interested in how digital experiences can be made accessible.

Klaus Opwis was appointed head of the Center for General Psychology and Methodology at the University of Basel, Switzerland, in 1995. His research interests include applied cognitive psychology, visual aesthetics, research methods, and HCI research.