Pictorial System Usability Scale (P-SUS): Developing an Instrument for Measuring Perceived Usability

Juergen Baumgartner[ab], Naomi Frei[a], Mascha Kleinke[a], Juergen Sauer[a] and Andreas Sonderegger[c]

[a]University of Fribourg,
Department of Psychology
Rue Faucigny 2, 1700 Fribourg,
Switzerland
juergen.baumgartner@unifr.ch

[b]We Are Cube, Puzzle ITC
Belpstrasse 37, 3007 Bern,
Switzerland

[c]EPFL+ECAL Lab, EPFL
11 av. du 24-Janvier, 1020 Renens,
Switzerland

ABSTRACT

We have developed a pictorial multi-item scale, called P-SUS (Pictorial System Usability Scale), which aims to measure the perceived usability of mobile devices. The scale is based on the established verbal usability questionnaire SUS (System Usability Scale). A user-centred design process was employed to develop and refine its 10 pictorial items. The scale was tested in a first validation study (*N*=60) using student participants. Psychometric properties (convergent validity, criterion-related validity, sensitivity, and reliability), as well as the motivation to fill in the scale were assessed. The results indicated satisfactory convergent validity for about two-thirds of the items. Furthermore, strong correlations were obtained for the sum scores between verbal and pictorial SUS, and the pictorial scale was perceived as more motivating than the verbal questionnaire. The P-SUS represents a first attempt to provide a pictorial usability scale for the evaluation of (mobile) devices.

CCS CONCEPTS

• Human-centered computing → HCI design and evaluation methods

KEYWORDS

pictorial scale; consumer product; perceived usability; mobile device evaluation

ACM Reference format:

Juergen Baumgartner, Naomi Frei, Mascha Kleinke, Juergen Sauer and Andreas Sonderegger. 2019. Pictorial System Usability Scale (P-SUS): Developing an Instrument for Measuring Perceived Usability. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA. 11 pages. https://doi.org/10.1145/3290605.3300299

ACM ISBN 978-1-4503-5970-2/19/05.
DOI: https://doi.org/10.1145/3290605.3300299

1 INTRODUCTION

Coined over 30 years ago, usability is a concept of considerable importance for practitioners as well as for scientists in the domain of human-computer interaction (HCI). Describing two different qualities of the interaction of a user with a technical artefact, definitions of usability usually differentiate between a subjective consequence (satisfaction) and objective behavioural indicators (effectiveness and efficiency) of a user-system interaction. The International Organization for Standardization defines usability as 'the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use' [29; p. 6]. Although this conceptual breadth is the reason why some scientists argue that the usability construct is a dead-end from a theoretical point of view [59], a growing community of practitioners is out there adhering to the concept, driven by digitalisation of our lives and the continuous need to make the astounding technological progress accessible for the common user. This is generally done by evaluating usability throughout the development cycle of interactive artefacts allowing the development of technical systems that match the requirements of a user, also referred to as user-centred design (ISO 9241-210; [30]). While the concept of user experience (UX) has brought a slight shift in the scope of the user-centred design process by including also emotional experience as an important outcome measure, the assessment of subjective attitudes remains a predominant indicator of the interaction quality within the user-centred design process - a measure generally assessed by means of usability questionnaires. However, completing multi-item usability questionnaires

can be tedious and time-consuming in usability testing. According to Brooke [10], test users are reluctant to reply to long usability questionnaires after an extended usability test and tend to lose focus, leading to insufficient and biased data. In this context, a pictorial scale might provide a more effective, motivating, and pleasant alternative to verbal usability questionnaires.

1.1 Verbal Usability Questionnaires

A broad choice of instruments assessing the subjective experience of users interacting with technical devices is available (for an overview, see [4]), differing considerably with regard to their scope, detail, and length. With regard to the scope of measurement, it can be distinguished between instruments that assess exclusively usability as a subjective attitude towards the interaction with a system and others that open the scope towards the evaluation of user experience. A more fine-tuned categorisation can be made by the criteria put forward by Christophersen and Konradt [14], consisting of (a) evaluation purpose (summative vs. formative evaluation), (b) evaluated system (device type, domain), (c) dimensionality (single vs. multiple dimensions), (d) answer format (Likert-type, free text answers, 'no opinion'-options), and (e) item format (statement, question, semantic differential). Further criteria might be important to consider as well, such as length of the questionnaire, how established and frequently used an instrument is and of which quality the psychometric properties of the instrument are. An additional important criterion, especially for non-English researchers and practitioners is if the desired instrument exists in a validated form in the target language.

A well-known tool in usability practice and research is the System Usability Scale (SUS, [10,40]). This ten-item scale has been referred to as an industry standard [11] and is considered to be one of the most widely used questionnaires in usability practice [38]. The scale has excellent psychometric properties with measures of reliability scoring over .90 (e.g. [3,55]), good indicators of validity (e.g. [3,34]) and sensitivity [55]. In addition, several norming studies (e.g. [3,54,56]) have presented normative data providing an empirical basis for the interpretation of SUS scores. According to such normative data, a SUS score of 65 can be interpreted as a marginally acceptable result. This corresponds to a D grade ranging between the adjectives "OK" and "good" [3] and represents approx. the 40% percentile rank [54]. While the SUS is described as a "quick and dirty" instrument [10] due to its brevity, a plethora of other scales have been developed in order to assess usability and related constructs.

Elaborate usability questionnaires with more than 70 items such as the MPUQ (Mobile Phone Usability Questionnaire; [53]) or the IsoMetrics [23] have the advantage that they consist of multiple dimensions. This increases, in the context of a formative evaluation, their diagnostic value and provide (in contrast to short questionnaires or even single-item scales) more information of where usability flaws may hide. Drawbacks are that they are time-consuming and require a considerable investment of motivation and effort of the participants. Thereby, negative effects such as increased dropout rates, a reduced response rate, and undesirable participants' behaviour might be the consequence. Bosnjak and colleagues [41] found that participation rates were significantly lower when the expected survey length was 30 minutes or longer. Furthermore, survey length was a greater predictor of response rates than materialistic incentives. Similarly, Galesic and Bosnjak [22] demonstrated that participants were more likely to start and complete a questionnaire if it only took 10 minutes, compared to 20 and 30 minutes. For answers given near the end of the longer survey, they observed that participants generally spent less time and that the length of open question responses decreased significantly. Participants also showed a tendency of marking the same score on the response scale for questions later in the survey, resulting in reduced variability in responses. Another potential consequence of long questionnaires is negative survey experience, which can impact participants' motivation in following studies and results in a loss of response quality [46]. Creating a positive survey experience is therefore not only important for present study motivation but also participation in subsequent studies.

Research suggests that positive affect increases interest and enjoyment of interesting tasks and therefore results in a higher intrinsic motivation [32]. One way to increase participants' motivation in usability studies is to provide questionnaires in a way that goes beyond the usual verbal evaluation. This could be done by creating pictorial questionnaires.

1.2 Pictorial Usability Questionnaires

A limited number of pictorial instruments has been developed in the domain of HCI. These instruments are exclusively dedicated to the measurement of emotion or mood, such as SAM (Self Assessment Manikin; [7]),

PREMO (Product Emotion Measurement Tool; [16]), PAM (Pick-A-Mood; [18]), LEMtool (Layered Emotion Measurement Tool; [13]) and AniSAM/AniAvatar [57]. With regard to usability, a pictorial single-item scale has been developed recently [4].

Numerous advantages are associated with the use of pictorial instruments. It has been argued that the use of pictorial items increases motivation. This has been attributed to the fact that pictorial items are more intuitively understandable and cognitively less demanding [7,17]. This and the pictures in general make the processing quicker and more pleasant [25,58]. Another important advantage of pictorial questionnaires is that they are independent of language. Therefore, a pictorial instrument does not have to be translated when used in different linguistic zones and can be used across cultures [18]. Furthermore, the language independence simplifies the questioning of children and people with lower linguistic skills [12,24,50,57].

The use of pictorial scales also has some drawbacks. First of all, it has been argued that pictorial items may not be as intuitively understandable as suggested in the literature [5]. Pictorial items can be vague and not contain enough hints for a reliable judgement [36]. They also have been criticised to be too oversimplified [57]. This can lead to misinterpretation or confusion and thus to increased processing time and flawed answers [51,57]. Pictorial items could also contain elements such as gestures or facial expressions that are not understood equally across cultures [4]. A way to avoid such misunderstandings is to add verbal hints or verbal instructions. By consequence, the items are no longer completely language independent [5,9]. A further point to consider is the potential (subjective) influence of a picture on the participant. There is a risk that a subject could choose an item because it is subjectively attractive, and not because its meaning correctly reflects the subject's experiences or thoughts [25,51]. In addition, the development of valid pictorial scales can be very time-consuming and may demand several adaptions and pre-tests [57].

1.3 Research Goals

The purpose of this project was twofold. The first aim was to develop a pictorial usability questionnaire based on the well-established verbal usability questionnaire SUS [10] that facilitates gathering subjective data. The reason for creating a pictorial version of the SUS was to provide an inclusive, motivating and pleasant alternative to the most widely used usability questionnaire. The pictorial items were designed specifically with regard to the evaluation of smartphones, which represents a frequently used everyday device. Since most items of the SUS are formulated in a very general way, they can be easily applied to various devices that use a graphical user interface. Kortum and Bangor [35] for example demonstrated in an online study that the SUS is well suitable and valid for the evaluation of a wide range of devices such as mobile phones, navigation systems, audio players, etc.

The second aim was to gather initial data allowing us to determine psychometric properties of the newly developed pictorial scale with a first lab-based study. Additional goals were to gain further insights into how such scales are perceived. To our knowledge, there are no pictorial multi-item scales developed yet for the measurement of perceived usability.

2 SCALE DEVELOPMENT

2.1 Scale Development Process

The development process consisted of three phases. An overview of the development process is given in table 1.

Table 1. Overview of the iterative process of P-SUS development comprising phases, methods, number of involved team members or participants and steps.

Phase	Method	N	Step
Iteration 1	Association Elicitation Test	9	C
	Design Meeting	3	C
	Implementation	-	R
	Design Meeting	4	C
	Implementation	-	R
	Think-Aloud	5	E
Iteration 2	Design Meeting	3	C
	Implementation	-	R
	Design Meeting	4	C
	Implementation	-	R
	Think-Aloud	3	E
	Expert Survey	10	E
Iteration 3	Design Meeting	3	C
	Implementation	-	R
	Design Meeting	4	C
	Think-Aloud and Rapid Prototyping	5	R/E
	Validation Study	60	E

C=Conception, R=Realization, E= Evaluation

In each phase, a three-step procedure was applied, which was modelled on the Plan-Build-Run model used in IT projects [62], resulting in Conception-Realization-Evaluation. Several methods were applied for the three steps, consisting of (a) association elicitation test, (b) design meetings, (c) implementation of changes, (d) think-aloud protocols, (e) rapid prototyping, (f) expert survey

and (g) validation study. The methods are described in more detail in the following paragraphs.

2.1.1 Methods Used for Conception

Association elicitation test. An association elicitation test was applied at the beginning of the first phase to gain insights into the mental models of participants filling in the verbal SUS questionnaire. The procedure was inspired by the gesture elicitation study of Angelini et al. [1] but was used to gather associations with verbal items instead of gestures. Nine students and young professionals (sex: 5 male, 4 female; age: M = 32.00, SD = 14.29; occupation: 2 computer scientists, 3 students, 1 business analyst, 1 foreign language assistant, 1 commercial clerk, 1 homemaker) were first confronted with the verbal items of the SUS and asked to explain how they understand the items by making free associations. In a subsequent step, they were asked to sketch out briefly on paper what images come into their minds when thinking of the item. The drawing part was inspired by the design studio method, a collaborative approach for ideation and design critique with users [6]. Sketches from all nine participants were then analysed by three authors for recurring themes and used as the first basis for visual representations of the pictorial items. Interpretation issues were discussed in subsequent design meetings.

Design meeting. Regular design meetings were held for two reasons: to discuss ideas and results of think-aloud protocols and to generate solutions for interpretation issues of ambiguous items. The core team consisted of two undergraduate psychology students and a supervisor with research- and graphical skills (UX design). Occasionally, experts in design (graphic designer, art historian) were invited to collaboratively develop ideas for adequate representations. Ideas were sketched out explicitly with paper and pencil to assure a swift proceeding. Outcomes were concrete modifications such as replacements for misleading visual elements with more specific ones, recommendations for focus on meaningful elements, etc.

2.1.2 Methods Used for Realization

Implementation of changes. Ideas developed in the various design meetings were drawn as vector graphics to easily allow modifications for subsequent iterations. Depending on the suggested modifications, existing items were adapted or built from scratch based on core components of a previously developed style guide (for more details see 2.2).

Rapid prototyping. A rapid prototyping approach was used in the last phase to refine items and test them in rapid iteration cycles. This approach is used in development teams to test extensive changes or new designs on a software with a low-fidelity prototype (to e.g. identify usability flaws) before spending resources to implement the changes on the software [27]. The goal was to test the current version of the scale with a person in a think-aloud setting and implement obtained feedback directly after the session. The elaborated version was then tested with the next person, etc. Such rapid iteration cycles were applied five times.

2.1.3 Methods Used for Evaluation

Think-aloud protocols. Think-aloud protocols (e.g. [37]) were carried out in all iterations to evaluate comprehension for each item by asking users to verbalise their understanding of the pictorials. A facilitator wrote down the interpretations. They were then categorized by the same facilitator with regard to accuracy (high, medium, low). If an item was not sufficiently comprehensible, the facilitator asked the participant at the end of the session which visual elements were difficult to understand and how the item could be improved to match its original meaning. In the following design meeting, these inputs were shared and discussed within the whole team. They were used as a basis for brainstorming and further development of the items.

Expert survey. An expert survey was conducted online at the end of the second iteration to gather insights from experts in visual design and usability. Ten experts (usability experience in years, M = 9.90, SD = 5.82) were asked for their opinion with regard to each pictorial item. The procedure was as follows: (a) The items were shown one by one (exposure). (b) Participants had to write down their personal interpretation of the scale and describe which visual elements led to their interpretation. (c) After that, they were shown the verbal wording of the corresponding SUS item. They then had to give a subjective rating on a seven-point Likert scale of how comprehensible the pictorial scale was. (d) In the end, participants had to write down specific suggestions for improving the scales.

Validation study. An initial validation study was carried out with the goal to test several psychometric properties of the P-SUS in an experimental setting. Convergent validity, criterion-related validity, sensitivity, and internal consistency were assessed. A similar set of psychometric properties was already used by other researchers

conducting validation studies of newly developed instruments (e.g. [44,45].

2.2 Pictorial Items of the P-SUS

Ten pictorial items were developed for the P-SUS within three iterations (see figure 1). They are based on the verbal items of the SUS [10]. Items are bipolar and consist of two visual representations, one for each extreme point. A seven-point scale was used since research suggests that more than five answer options maximize the reliability of a scale (e.g. [48]). Furthermore, a seven-point version is more likely to provide integer interpolations than five-point scales [20].

Items were drawn using a professional design tool called Sketch (www.sketchapp.com) to obtain vector graphics. In the beginning, a provisional design style guide was developed that consists of a number of core components such as avatar representations, visual elements/objects (e.g. smartphone), symbols and icons. This style guide was continuously refined and specific details and components were adapted during the design process.

The two visual extreme points used for each item depict a positive and a negative usage condition, consisting of an avatar interacting with a mobile device and further elements to convey and underline the specific meaning of the item. Several visual elements were used to increase comprehensibility, such as signal colours (red, green), meaningful symbols (check mark, x mark, light bulb, question mark), and elements from comic strips such as text bubbles [19] and onomatopoeia [33]. All items were designed in a similar fashion.

Two versions of the items were developed, a male and a female version with otherwise identical content. Since some items turned out to be difficult to understand, we added a few verbal elements in form of keywords and onomatopoeia for the items 01, 03, 04, 07 and 08 in order to increase comprehensibility and to minimize ambiguity. This was done because the above described iterative user-centred design-process indicated that some items turned out to be difficult to understand. A full list of the ten pictorial items and a full overview of their evolution through the development process can be found in the supplementary materials or under https://invis.io/FXNXGNBAVCT.

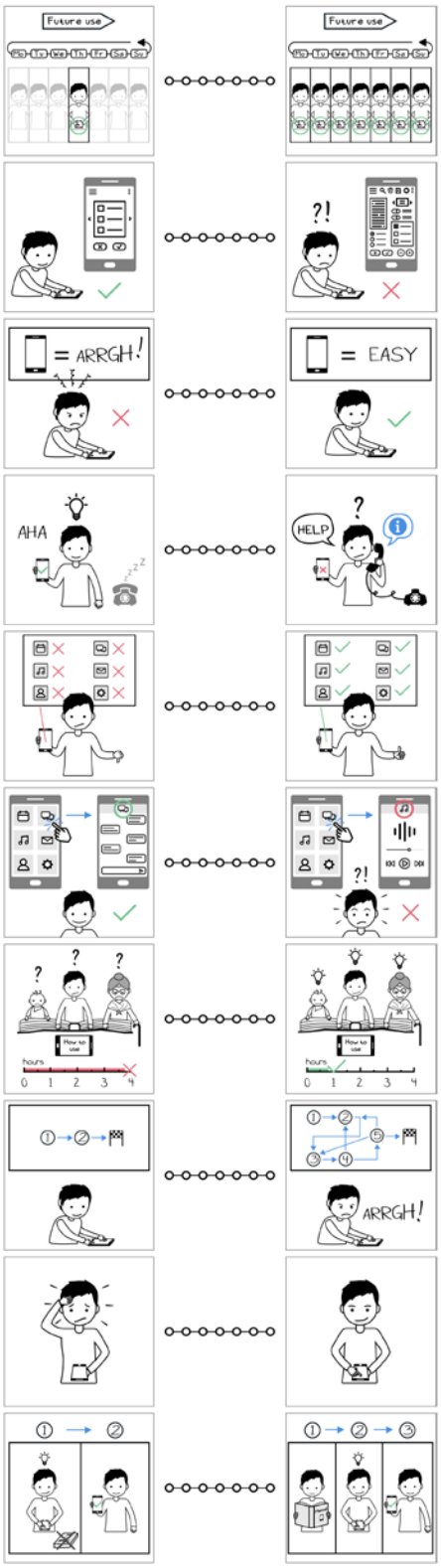

Figure 1. A pictorial version of the ten verbal SUS items with positive and negative extreme points and seven-point scale.

3 VALIDATION STUDY

3.1 Goal of the Study

The goal of the study was to collect indicators describing the validity and reliability of the developed P-SUS. Convergent validity, criterion-related validity, sensitivity, reliability, questionnaire completion time and motivation to complete the questionnaire were assessed in a lab-based experiment, in which participants first interacted with a smartphone prototype developed by Hamborg and colleagues [26], and second evaluated the prototype by means of usability questionnaires.

3.2 Method

3.2.1 Participants

Sixty participants took part in this study (65% female, 83.3% students). Participants were recruited from the Department of Psychology (33.3%) and other departments of the University of Fribourg and Berne, and did not have specific knowledge of HCI and usability. They were aged between 20 and 31 years (M = 22.88, SD = 1.56) and rated themselves rather highly in smartphone expertise (M = 5.73, SD = 0.78) on a seven-point Likert scale ranging from 1 (beginner) to 7 (expert). All participants possessed smartphones.

3.2.2 Experimental Design

A one-factorial between-subjects design was used with system usability as the independent factor. System usability was manipulated by providing the participant with a simple navigation structure of the prototype (high usability) vs. a complex one (low usability). The two prototypes differed exclusively with regard to their navigation structure; all visual and aesthetical elements were identical.

3.2.3 Measures and Instruments

To determine the quality of the pictorial questionnaire, four primary psychometric measures were assessed: (1) convergent validity, (2) criterion-related validity, (3) sensitivity and (4) internal consistency. Two secondary psychometric properties were obtained with regard to the questionnaire modality (pictorial vs. verbal), consisting of (5) questionnaire completion time and (6) motivation to complete the questionnaire. All measures and instruments are outlined below.

Convergent validity. Convergent validity is a subtype of construct validity and is presumed to correlate highly when two independent questionnaires measure the same construct (e.g. [43]). It was assessed by using the SUS [10] as the main convergent validity measure for perceived usability. The SUS consists of ten items and represents a well-established verbal questionnaire for the assessment of perceived usability. The original version of the SUS uses a five-point Likert scale (1 = strongly disagree, 5 = strongly agree), but for the purpose of this study, the response format was adapted to a seven-point scale using the original verbal anchors (for more details see 2.2). The SUS is a widely-used instrument with high internal consistency (Cronbach's $\alpha > .91$; [2]).

Criterion-related validity. Criterion-related validity describes the relationship between two independent measures of the same construct. An estimation of criterion-related validity was assessed by measuring two 'objective' performance aspects of the interaction with the smartphone prototype (task completion time and number of interactions) and correlating them with the 'subjective' score of the P-SUS and the SUS, respectively. We expect correlations of medium size between measures of subjective and objective usability (for a detailed discussion see [4]).

Sensitivity. Sensitivity describes the ability of a scale to distinguish between different levels of usability [39]. It was assessed by comparing group-means of the low usability condition with the high usability condition. A significant difference is to be expected when an instrument is said to be highly sensitive.

Internal consistency. Internal consistency 'refers to the homogeneity of the items in the measure' [26; p. 968] and is the most commonly accepted measure of reliability. It was assessed by calculating Cronbach's alpha separately for both pictorial and verbal scale. An internal consistency of $\alpha > .90$ is to be expected from a highly reliable questionnaire [49].

Questionnaire completion time. Questionnaire completion time refers to the time participants needed to fill in the questionnaire. It was measured in seconds and was automatically recorded by the online questionnaire.

Motivation to complete the questionnaire. Motivation to complete the questionnaire was assessed using the three-item subscale of the short version of the IMI (Intrinsic Motivation Inventory; [61]), originally developed by Ryan [52]. The items were used to determine the level of fun, joy, and interest experienced during an activity. The wording was adapted by specifying the kind of activity: 'Filling in the questionnaire was fun', 'I enjoyed filling in the questionnaire very much' and 'I would describe filling in the questionnaire as very interesting'. A seven-point

scale was used (1 = strongly disagree, 7 = strongly agree). McAuley and colleagues [42] reported an acceptable internal consistency for the subscale (α = .78).

3.2.4 Procedure

Participants were tested individually either in a quiet room at the University of Fribourg or at the participant's home. The facilitator provided a laptop with the smartphone prototype and the questionnaires. After filling in the form of informed consent and a demographical questionnaire (age, gender, field of study, expertise and possession of a smartphone), the participant carried out three tasks on a smartphone prototype that was displayed on the laptop. These consisted of (a) creating a new entry in the address book, (b) retrieving the phone bill, and (c) changing the ringtone of the smartphone. Half of the participants operated a prototype with a simple menu structure (high usability), whereas the other half operated one with a complex structure (low usability). After the completion of the tasks, participants filled in P-SUS and SUS. The sequence of these questionnaires was counterbalanced in order to avoid carry-over effects. After each questionnaire, motivation to complete the questionnaire was assessed using the short version of the IMI.

3.2.5 Data Analysis

Correlational analyses, comparisons of group means and reliability tests were used for data analysis. Correlations were computed to estimate convergent and criterion-related validity, comparisons of group means were used to determine if the instruments are capable to distinguish between different levels of usability, and reliability tests were carried out to assess internal consistency of the questionnaires. The interpretation of the effect size *r* was based on Cohen [15], who differentiates between small (*r* = .100), medium (*r* = .300) and large effects (*r* = .500). We applied non-parametric tests when requirements for normal distribution and homogeneity of variance of the data were not met.

4 RESULTS

4.1 Primary Psychometric Properties

Convergent validity. Analysis of the data revealed correlations of *r* > .500 for about two-thirds of the items. However, items 04, 06 and 10 were below the expected magnitude level (see table 2). Importantly, a high correlation was obtained between pictorial and verbal sum score of the SUS (*r* = .865).

Table 2. Means of positively poled items (1-7) and sum score (0-100) of P-SUS and SUS, and correlation of verbal with pictorial items (*N*=60).

Item	verbal *M (SD)*	pictorial *M (SD)*	*r*
01	4.00 (1.66)	4.62 (1.69)	.624***
02	5.30 (1.91)	5.45 (1.81)	.673***
03	5.45 (1.59)	5.50 (1.67)	.801***
04	6.88 (0.37)	6.05 (1.13)	.378**
05	5.03 (1.56)	5.58 (1.44)	.548***
06	5.38 (1.53)	6.08 (1.21)	.211
07	5.50 (1.64)	5.52 (1.54)	.575***
08	5.35 (1.89)	5.08 (1.95)	.646***
09	5.32 (1.48)	5.65 (1.51)	.668***
10	6.22 (1.14)	5.83 (1.44)	.489***
Score	74.06 (19.21)	75.61 (19.39)	.865***

* $p < .05$; ** $p < .01$; *** $p < .001$

Criterion-related validity. The sample size was reduced to *N*=57 due to technical issues of recording the data of three participants. Moderately negative correlations were obtained between task completion time (*M* = 75.20, *SD* = 33.24) and sum scores of both the pictorial SUS (*r* = −.342**) and the verbal SUS (*r* = −.419**), respectively. The magnitude of correlations between user interactions (*M* = 31.44, *SD* = 15.61) and sum scores for both instruments were identical (*r* = −.495***). Overall, the criterion-related validity for both pictorial and verbal scales revealed a highly similar pattern.

Sensitivity. A Mann-Whitney test was conducted to assess if the sum scores of both scales are able to distinguish between low and high usability condition. The P-SUS obtained significantly higher scores in the high-usability condition than in the low-usability condition (Mdn_{high} = 42.52, Mdn_{low} = 19.26, *U* = 101.00, *z* = −5.16, *p* = .000, *r* = −.666). A highly similar pattern was observed for the SUS (Mdn_{high} = 42.76, Mdn_{low} = 19.03, *U* = 94.00, *z* = −5.26, *p* = .000, *r* = −.680). The results suggest that P-SUS and SUS distinguish very well between devices of high and low usability.

Internal consistency. The reliability analysis revealed high internal consistency for both scales, the P-SUS (α = .912) and the verbal SUS (α = .914). Both values were calculated using all ten items.

4.2 Secondary Psychometric Properties

Questionnaire completion time. T-tests were carried out to assess if there was a difference in the amount of time needed to fill in the P-SUS and SUS, respectively. Participants needed significantly more time filling in the pictorial questionnaire (M = 113.33, SD = 44.42) than the verbal one (M = 78.12, SD = 29.37, t (59) = 6.520, p = .000, r = .424).

Motivation to complete the questionnaire. T-tests were carried out to assess if there were significant differences with regard to the motivation of filling in the pictorial and verbal questionnaire. Results showed on all three items, motivation was rated higher when filling in the pictorial questionnaire: fun ($M_{pictorial}$ = 5.03, SD = 1.63; M_{verbal} = 4.22, SD = 1.45, t (59) = 3.673, p = .001, r = 0.254), joy ($M_{pictorial}$ = 5.33, SD = 1.59; M_{verbal} = 3.85, SD = 1.46, t (59) = 6.710, p = .000, r = .436), and interest ($M_{pictorial}$ = 4.95, SD = 1.41; M_{verbal} = 3.92, SD = 1.38, t (59) = 4.689, p = .000, r = .346). Comparison of sum scores revealed a significant difference in motivation between pictorial and verbal questionnaire ($M_{pictorial}$ = 5.11, SD = 1.44; M_{verbal} = 3.99, SD = 1.26, t (59) = 5.732, p = .000, r = .382).

5 DISCUSSION

The findings of the validation study revealed high convergent validity for seven out of ten pictorial items, with correlations of r > .500. Three items showed a correlation below .500 (item 04: 'I think that I would need the support of a technical person to be able to use this system'; item 06: 'I thought there was too much inconsistency in this system'; item 10: 'I needed to learn a lot of things before I could get going with this system'). One possible reason for the low correlations might be that the items were still too ambiguous and open to multiple interpretations, despite the efforts that have been made during the iterative design phase. Interestingly, the global scores of P-SUS and SUS were highly correlated (r = .866), which demonstrates that a very similar overall result can be obtained and that both instruments measure the same construct.

The results with regard to criterion-related validity showed correlations of medium size for task completion time and medium to large effects for the number of interactions with the prototype. Research shows generally mixed evidence for the relationship between subjective and objective usability, with meta-analyses indicating medium to large effects (e.g. [47,55]), whereas others indicate rather small effects (e.g. [29]). Nevertheless, since both instruments showed a similar pattern of effects, we consider these results to be acceptable.

The analysis of sensitivity demonstrated that the P-SUS is able to detect changes in usability, with effect sizes indicating a large effect. A highly similar pattern could be demonstrated for the verbal SUS.

Reliability analysis revealed a Cronbach's Alpha from above .900 for both scales, which is considered as an excellent internal consistency [49]. With regard to the verbal SUS, indicators of this study are very similar to outcomes of reliability analyses from other authors [2].

Questionnaire completion time was significantly lower for the verbal questionnaire compared to the pictorial one. Participants spent in the mean three to four seconds longer per item when filling in the P-SUS compared to the verbal version. One interpretation might be that participants are more used to fill in verbal questionnaires and less to fill in pictorial ones, that is they process verbal content easier and quicker. Another interpretation could be that the participants needed more time to decipher the true meaning of the scale, since some items consisted of various visual elements. The second interpretation is in stark contrast with the initial hypothesis of this study suggesting that pictorial scales are more intuitively understandable and thus processed quicker (see also [8]).

However, questionnaire motivation was significantly higher for the P-SUS compared to the verbal questionnaire. The largest difference was obtained for the joy item, indicating that participants enjoyed more filling in the pictorial questionnaire than the verbal one. This goes in line with the opinion of other authors (e.g. [58]) that pictorial scales are more pleasant, which is likely due to the visual nature and the amusing images of such questionnaires. Some participants also commented and verbalised between the different parts of the experiment that they liked the pictorial scale very much. Similar behaviour has also been reported by other researchers (e.g. [16]).

With regard to these findings, however, some limitations need to be addressed: (a) The sample consisted of only sixty participants and was rather homogeneous with regard to participants' occupation, education level, experience, and age. In order to generalize results, further studies need to be conducted using a larger and more heterogeneous sample. (b) Another potential limitation that should be considered is the possibility that the results might be biased to a certain degree by expert effects. It can be assumed that participants with an academic

background (as the ones in this study) are more often faced with verbal questionnaires than those without. They perceive such questionnaires as routine, the shorter completion time for the SUS might be interpreted as an indicator for this assumption. Thus, such participants encounter fewer problems filling in verbal questionnaires compared to pictorial ones, which, by contrast, pose a completely new challenge. Therefore, it would be interesting to compare the two questionnaire versions without such a bias by, for example, training participants to get used to pictorial scales. (c) Another limitation addressed the interpretation of the pictorials across cultures. Even if pictorial instruments are mainly nonverbal, they may contain visual elements that are more prevalent in certain cultures [21]. Therefore, such questionnaires need to be evaluated across cultures in order to assess to which degree consistent interpretations are obtained. (d) A further limitation is linked to the nonverbal nature of pictorial scales. Not all pictorial items used in the P-SUS are exclusively nonverbal. Half of the items used verbal keywords in order to increase their comprehensibility and reduce ambiguity (see section 2.2). Therefore, the P-SUS can be described as a hybrid scale, consisting of pictorial and verbal elements such as onomatopoeia and short keywords. We assume that user groups with special needs such as dyslexic users and users with lower levels of education would benefit from hybrid scales, since pictures combined with simple keywords are presented instead of exclusively verbal statements. Future studies should include users with special needs to address this assumption. (e) Since the majority of our participants were students or people with a high level of education, we used a 7-point scale to give them more options when responding to the questions (see e.g. [60]). Since we used a 7-point scale for both the pictorial and standard SUS in the present study, it is not likely that the extended response format has affected the results (an appropriate adjustment was made to the multiplier of the usability score). However, practitioners who want to use the normative data available for the SUS might wish to continue using a 5-point scale. (f) The last limitation relates to the three items (04, 06 and 10) that obtained low convergent validity. Two possible solutions for future development could be: (1) to refine the pictorial items in order to get acceptable correlations (i.e. $r > .500$), or (2) to remove the three items and use a seven-item version of the P-SUS for further validation. Further research needs to address the question whether the low correlations might also be due to issues in understanding the verbal items.

6 CONCLUSION

Overall, the psychometric properties of the P-SUS and the additional variables evaluated in this study can be considered as satisfactory given its close conformity with the results obtained using the verbal scale. These first results with regard to the psychometric properties of a pictorial usability questionnaire are encouraging. They demonstrate that similar results can be obtained with a pictorial scale compared to a verbal one, with the additional advantage of increased motivation, but also with the drawback of a longer completion time. Furthermore, the findings show that the application of a mix of various design-methods supports the scale development, considering inputs of users and experts as well. Nevertheless, more work on the scale and further validation studies with a larger and more heterogeneous sample are needed in order to permit a valid and reliable pictorial assessment across cultural and linguistic borders.

ACKNOWLEDGEMENTS

The research was funded by a grant (No 100014_140359) from the Swiss National Science Foundation (SNSF). Their support is gratefully acknowledged. We are grateful to Veronica Solombrino for the numerous design reviews, to Franziska Asenbauer for her help in conception and to the UX team of We Are Cube/Puzzle ITC for the valuable feedback and the support during the development process.

REFERENCES

[1] Leonardo Angelini, Francesco Carrino, Stefano Carrino, Maurizio Caon, Omar Abou Khaled, Jürgen Baumgartner, Andreas Sonderegger, Denis Lalanne, and Elena Mugellini. 2014. Gesturing on the Steering Wheel: a User-elicited taxonomy. In Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 1–8.

[2] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of usability studies 4, 3: 114–123.

[3] Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An empirical evaluation of the system usability scale. Intl. Journal of Human–Computer Interaction 24, 6: 574–594.

[4] Juergen Baumgartner, Andreas Sonderegger, and Juergen Sauer. 2019. No need to read: Developing a pictorial single-item scale for measuring perceived usability. International Journal of Human-Computer Studies 122: 78–89. https://doi.org/10.1016/j.ijhcs.2018.08.008

[5] Alberto Betella and Paul F. M. J. Verschure. 2016. The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. PLOS ONE 11, 2: e0148037. https://doi.org/10.1371/journal.pone.0148037

[6] Eli Blevis, Youn-kyung Lim, Erik Stolterman, Tracee Vetting Wolf, and Keichi Sato. 2007. Supporting design studio culture in HCI. In CHI '07 extended abstracts on Human factors in computing systems - CHI '07, 2821. https://doi.org/10.1145/1240866.1241086

[7] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry 25, 1: 49–59.

[8] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry 25, 1: 49–59.

[9] Joost Broekens and Willem-Paul Brinkman. 2013. AffectButton: A method for reliable and valid affective self-report. International Journal of Human-Computer Studies 71, 6: 641–667. https://doi.org/10.1016/j.ijhcs.2013.02.003

[10] John Brooke. 1996. SUS-A quick and dirty usability scale. Usability evaluation in industry 189, 194: 4–7.

[11] John Brooke. 2013. SUS: A Retrospective. J. Usability Studies 8, 2: 29–40.

[12] Heather Buchanan and N. Niven. 2002. Validation of a Facial Image Scale to assess child dental anxiety. International journal of paediatric dentistry 12, 1: 47–52.

[13] Kevin Capota, Marco van Hout, and Thea van der Geest. 2007. Measuring the emotional impact of websites: a study on combining a dimensional and discrete emotion approach in measuring visual appeal of university websites. In Proceedings of the 2007 conference on Designing pleasurable products and interfaces, 135–147. Retrieved April 27, 2017 from http://dl.acm.org/citation.cfm?id=1314173

[14] Timo Christophersen and Udo Konradt. 2011. Reliability, validity, and sensitivity of a single-item measure of online store usability. International Journal of Human-Computer Studies 69, 4: 269–280. https://doi.org/10.1016/j.ijhcs.2010.10.005

[15] J. Cohen. 1988. The effect size. Statistical power analysis for the behavioral sciences: 77–83.

[16] Pieter Desmet. 2003. Measuring Emotion: Development and Application of an Instrument to Measure Emotional Responses to Products. In Funology, Mark A. Blythe, Kees Overbeeke, Andrew F. Monk and Peter C. Wright (eds.). Springer Netherlands, 111–123. https://doi.org/10.1007/1-4020-2967-5_12

[17] Pieter Desmet, Kees Overbeeke, and Stefan Tax. 2001. Designing products with added emotional value: Development and appllcation of an approach for research through design. The design journal 4, 1: 32–47.

[18] Pieter Desmet, Martijn H. Vastenburg, and Natalia Romero. 2016. Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. Journal of Design Research 14, 3: 241–279.

[19] Will Eisner. 1985. Theory of Comics and Sequential Art. F.: Poorhouse press.

[20] Kraig Finstad. 2010. Response interpolation and scale sensitivity: Evidence against 5-point scales. Journal of Usability Studies 5, 3: 104–110.

[21] Dawn P. Flanagan, Samuel O. Ortiz, and Vincent C. Alfonso. 2013. Essentials of cross-battery assessment. John Wiley & Sons.

[22] Mirta Galesic and Michael Bosnjak. 2009. Effects of questionnaire length on participation and indicators of response quality in a web survey. Public opinion quarterly 73, 2: 349–360.

[23] Günther Gediga, Kai-Christoph Hamborg, and Ivo Düntsch. 1999. The IsoMetrics usability inventory: An operationalization of ISO 9241-10 supporting summative and formative evaluation of software systems. Behaviour & Information Technology 18, 3: 151–164. https://doi.org/10.1080/014492999119057

[24] Ramesh Ghiassi, Kevin Murphy, Andrew R. Cummin, and Martyn R. Partridge. 2011. Developing a pictorial Epworth Sleepiness Scale. Thorax 66, 2: 97–100. https://doi.org/10.1136/thx.2010.136879

[25] Shamila Haddad, Steve King, Paul Osmond, and Shahin Heidari. 2012. Questionnaire Design to Determine Children's Thermal Sensation, Preference and Acceptability in the Classroom. 7.

[26] Kai-Christoph Hamborg, Julia Hülsmann, and Kai Kaspar. 2014. The Interplay between Usability and Aesthetics: More Evidence for the "What Is Usable Is Beautiful" Notion. Advances in Human-Computer Interaction 2014: 1–13. https://doi.org/10.1155/2014/946239

[27] Rex Hartson and Pardha S. Pyla. 2012. The UX Book: Process and guidelines for ensuring a quality user experience. Elsevier.

[28] Timothy R. Hinkin. 1995. A review of scale development practices in the study of organizations. Journal of Management 21, 5: 967–988. https://doi.org/10.1016/0149-2063(95)90050-0

[29] Kasper Hornbæk and Effie Lai-Chong Law. 2007. Meta-analysis of Correlations Among Usability Measures. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07), 617–626. https://doi.org/10.1145/1240624.1240722

[30] International Organization for Standardization. 2010. Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems (Standard No. 9241-210). Retrieved from https://www.iso.org/standard/52075.html

[31] International Organization for Standardization. 2016. Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts (Standard No. 9241-11.2). Retrieved from https://www.iso.org/standard/63500.html

[32] Alice M. Isen and Johnmarshall Reeve. 2005. The influence of positive affect on intrinsic and extrinsic motivation: Facilitating enjoyment of play, responsible work behavior, and self-control. Motivation and emotion 29, 4: 295–323.

[33] Heike Jüngst. 2010. Information Comics–Knowledge Transfer in a Popular Format.

[34] Philip Kortum and S. Camille Peres. 2014. The relationship between system effectiveness and subjective usability scores using the System Usability Scale. International Journal of Human-Computer Interaction 30, 7: 575–584.

[35] Philip T. Kortum and Aaron Bangor. 2013. Usability ratings for everyday products measured with the System Usability Scale. International Journal of Human-Computer Interaction 29, 2: 67–76.

[36] Theodore Kunin. 1955. The Construction of a New Type of Attitude Measure. Personnel Psychology 8, 1: 65–77. https://doi.org/10.1111/j.1744-6570.1955.tb01189.x

[37] Clayton Lewis and Robert Mack. 1982. Learning to Use a Text Processing System: Evidence from "Thinking Aloud" Protocols. In Proceedings of the 1982 Conference on Human Factors in Computing Systems (CHI '82), 387–392. https://doi.org/10.1145/800049.801817

[38] James (Jim) R. Lewis and Jeff Sauro. 2017. Revisiting the Factor Structure of the System Usability Scale. J. Usability Studies 12, 4: 183–192.

[39] James R. Lewis. 2002. Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. International Journal of Human-Computer Interaction 14, 3–4: 463–488. https://doi.org/10.1080/10447318.2002.9669130

[40] James R. Lewis. 2018. The System Usability Scale: Past, Present, and Future. International Journal of Human–Computer Interaction 34, 7: 577–590. https://doi.org/10.1080/10447318.2018.1455307

[41] Bernd Marcus, Michael Bosnjak, Steffen Lindner, Stanislav Pilischenko, and Astrid Schütz. 2007. Compensating for Low Topic Interest and Long Surveys: A Field Experiment on Nonresponse in Web Surveys. Social Science Computer Review 25, 3: 372–383. https://doi.org/10.1177/0894439307297606

[42] E. McAuley, T. Duncan, and V. V. Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: a confirmatory factor analysis. Research Quarterly for Exercise and Sport 60, 1: 48–58. https://doi.org/10.1080/02701367.1989.10607413

[43] Samuel Messick. 1979. Test Validity and the Ethics of Assessment. ETS Research Report Series 1979, 1: i–43. https://doi.org/10.1002/j.2333-8504.1979.tb01178.x

[44] 44. Michael Minge and Laura Riedel. 2013. meCUE-Ein modularer Fragebogen zur Erfassung des Nutzungserlebens. In Mensch & Computer, 89–98.

[45] Morten Moshagen and Meinald T. Thielsch. 2010. Facets of visual aesthetics. International Journal of Human-Computer Studies 68, 10: 689–709. https://doi.org/10.1016/j.ijhcs.2010.05.006

[46] Clive Nancarrow and Trixie Cartwright. 2007. Online access panels and tracking research: the conditioning issue. International Journal

of Market Research 49, 5: 573–594. https://doi.org/10.1177/147078530704900505

[47] Jakob Nielsen and Jonathan Levy. 1994. Measuring Usability: Preference vs. Performance. Commun. ACM 37, 4: 66–75. https://doi.org/10.1145/175276.175282

[48] J. C. Nunnally. 1978. Psychometric Theory, Second. New York: McGrawHill.

[49] Jum C. Nunnally and I. H. Bernstein. 1994. Psychometric Theory (McGraw-Hill Series in Psychology). McGraw-Hill New York.

[50] Sampo V. Paunonen, Michael C. Ashton, and Douglas N. Jackson. 2001. Nonverbal assessment of the Big Five personality factors. European Journal of Personality 15, 1: 3–18. https://doi.org/10.1002/per.385

[51] Laura Reynolds-Keefer, Robert Johnson, and S. Carolina. 2011. Is a picture is worth a thousand words? Creating effective questionnaires with pictures. Practical Assessment, Research & Evaluation 16, 8: 1–7.

[52] Richard M. Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. Journal of personality and social psychology 43, 3: 450.

[53] Young Sam Ryu and Tonya L. Smith-Jackson. 2006. Reliability and Validity of the Mobile Phone Usability Questionnaire (MPUQ). J. Usability Studies 2, 1: 39–53.

[54] Jeff Sauro. 2011. A practical guide to the system usability scale: Background, benchmarks & best practices. Measuring Usability LLC Denver, CO.

[55] Jeff Sauro and James R. Lewis. 2009. Correlations among prototypical usability metrics: evidence for the construct of usability. In Proceedings of the SIGCHI conference on human factors in computing systems, 1609–1618.

[56] Jeff Sauro and James R. Lewis. 2016. Quantifying the user experience: Practical statistics for user research. Morgan Kaufmann.

[57] Andreas Sonderegger, Klaus Heyden, Alain Chavaillaz, and Juergen Sauer. 2016. AniSAM & AniAvatar: Animated Visualizations of Affective States. 4828–4837. https://doi.org/10.1145/2858036.2858365

[58] R. A. Stern, J. E. Arruda, C. R. Hooper, G. D. Wolfner, and C. E. Morey. 1997. Visual analogue mood scales to measure internal mood state in neurologically impaired patients: Description and initial validity evidence. Aphasiology 11, 1: 59–71.

[59] Noam Tractinsky. 2018. The Usability Construct: A Dead End? Human–Computer Interaction 33, 2: 131–177. https://doi.org/10.1080/07370024.2017.1298038

[60] Bert Weijters, Elke Cabooter, and Niels Schillewaert. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. International Journal of Research in Marketing 27, 3: 236–247.

[61] Matthias Wilde, Katrin Bätz, Anastassiya Kovaleva, and Detlef Urhahne. 2009. Überprüfung einer Kurzskala intrinsischer Motivation (KIM). Zeitschrift für Didaktik der Naturwissenschaften 15.

[62] Rüdiger Zarnekow and Walter Brenner. 2004. Integriertes Informationsmanagement: Vom Plan, Build, Run zum Source, Make, Deliver. Informationsmanagement: Konzepte und Strategien für die Praxis. dpunkt, Heidelberg: 3–24.