Questionnaire Experience and the Hybrid System Usability Scale: Using a Novel Concept to Evaluate a New Instrument

Juergen Baumgartner[1ab], Nicole Ruettgers[a], Annigna Hasler[a], Andreas Sonderegger[c], Juergen Sauer[a]

[a] Department of Psychology, University of Fribourg, Rue de Faucigny 2, 1700 Fribourg, Switzerland
[b] We Are Cube, Puzzle ITC Belpstrasse 37, 3007 Bern, Switzerland
[c] Bern University of Applied Sciences, Business School, Institute for New Work, Brückenstrasse 73, 3005 Bern, Switzerland

This article presents the concept of questionnaire experience (QX), with a view to adding a new element to the psychometric evaluation of questionnaires, which may eventually help increase the reliability and validity of instruments. The application of QX is demonstrated in the development of the Hybrid System Usability Scale (H-SUS), making use of items comprising pictorial and verbal elements to measure perceived usability. The H-SUS was modelled on the verbal version of the System Usability Scale (SUS). Since previous research showed advantages of pictorial scales over verbal scales (e.g., higher respondent motivation) but also disadvantages (e.g., longer completion times), we assumed that hybrid scales will combine the advantages of both scale types. The goal of this study was to compare the two instruments by assessing traditional psychometric criteria (convergent, divergent and criterion-related validity, reliability and sensitivity) and respondent-related aspects of QX (respondent workload, respondent motivation, questionnaire preference, and questionnaire completion time). An online experiment was carried out (N=152), in which participants interacted with a smartphone prototype and subsequently completed the verbal SUS together with the H-SUS. Results indicate good psychometric properties for the H-SUS. Compared to the SUS, the H-SUS showed similar workload levels for questionnaire completion, higher levels of respondent motivation, but longer questionnaire completion time. Overall, the H-SUS is considered a promising alternative for the evaluation of perceived usability. Finally, QX can be considered a useful concept for identifying potential problems of psychometric instruments in a respondent-centred way, which may help improve the quality of future scales.

Keywords: hybrid scale; questionnaire experience; consumer product evaluation; perceived usability; mobile device evaluation

Highlights

- The article presents an online study, in which a Hybrid System Usability Scale was compared with the original System Usability Scale.
- Besides traditional psychometric criteria (i.e. validity, reliability and sensitivity), measures of questionnaire experience (QX) were assessed (perceived respondent workload, respondent motivation, questionnaire preference, questionnaire completion time).
- The Hybrid System Usability Scale showed very similar psychometric properties as the verbal version, but respondents' questionnaire experience was more positive.

1 Introduction

The field of psychometrics has made great advancements over recent decades, resulting in the development of sound approaches to designing questionnaires (e.g. Coolican, 2017; Hinkin, 1995; Miller & Lovler, 2018). The focus was traditionally on achieving good scores on the standard coefficients used to determine the psychometric quality of a scale, such as reliability, validity, and, in

[1] Corresponding author. Phone: +41-26-3007663, Fax: +41-26-3009712. Rue Faucigny 2, CH-1700 Fribourg. Email address: juergen.baumgartner@unifr.ch

certain cases, sensitivity. There are other criteria, which are also important but have not received the same level of attention, though they may equally contribute to the improvement of the psychometric properties of questionnaires. This refers to the experience of the respondent during questionnaire completion, which may not always be positive (e.g., questionnaire is too long, some items are difficult to understand). We believe that a respondent's experience while answering questionnaires is important and hence suggest that by adopting a respondent-centred perspective in questionnaire design (similar to the user-centred approach in system design, e.g. Gould & Lewis, 1985; ISO 9241-210, International Organization for Standardization, 2019), a more positive experience can be achieved. We have coined the term 'questionnaire experience' (QX) to emphasise this approach. QX encompasses various factors that are relevant for creating a positive experience when respondents complete questionnaires. Such a positive experience is expected to have effects on several factors influencing respondents' behaviour and attitudes (e.g., conscientiousness of questionnaire completion, motivation to complete questionnaire again), which in turn could potentially affect the psychometric properties of the instrument.

In addition to the introduction of the concept of QX, we also examine whether hybrid scales as an alternative form of questionnaire design provide advantages over traditional verbal scales. Hybrid scales combine images with verbal elements to improve the comprehension of the scale (Sauer et al., in press). Due to their visual nature, hybrid scales are expected to influence QX in a positive way.

We believe that both principal issues dealt with in this article (i.e. hybrid scales, concept of QX) are relevant to a wide range of domains in which psychometric testing plays a role. In the present article, we focus on the usability domain because in this domain the use of hybrid scales and the application of the QX concept are expected to be of particular benefit.

1.1 Questionnaire Experience (QX)

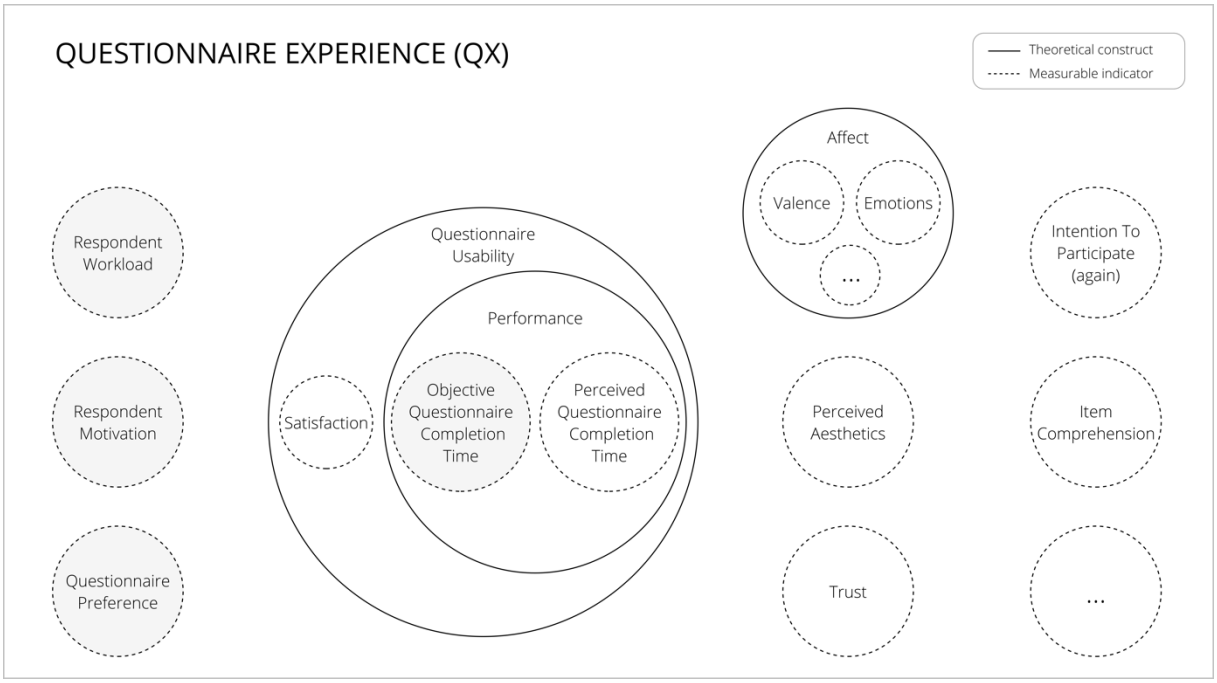
When developing questionnaires, many aspects are to be considered in order to create good instruments. The literature describes several steps to take for quality control prior to administering a questionnaire, such as using guidelines for formulation of good items (e.g. Thielsch et al., 2012), paying attention to questionnaire length (Galesic & Bosnjak, 2009), completing qualitative item analyses, carrying out expert reviews, and conducting a pilot test (Miller & Lovler, 2018). Before publishing a questionnaire, there are further steps to follow, such as assessing psychometric criteria (e.g., reliability and validity), having the questionnaire reviewed by test takers, and using expert panels to assess content validity (Miller & Lovler, 2018). All these steps are of importance because they help reduce measurement error, thus improving the validity and reliability of the instrument. However, an aspect that is rarely considered explicitly during questionnaire development concerns the experience of participants when completing a questionnaire. More precisely, it refers to the following questions: Is the workload of respondents too heavy because the items are difficult to understand? Is the questionnaire motivating or even fun to complete? Are questions (intuitively) comprehensible to all respondents? How do respondents experience the completion of several items, which seem to ask the same question (usually used to reduce measurement error)? If participants are not sufficiently motivated, the probability of undesirable response patterns increases, such as giving random responses or skipping questions (Robins et al., 2001). As a result, the outcomes of questionnaire application may be impaired. These points are rarely taken into consideration when questionnaires are developed. Therefore, it is advisable to pay attention to these points, especially when a battery of questionnaires is administered (e.g., after having completed an experimental task) or when the same questionnaire is administered repeatedly.

Since the participants' point of view during questionnaire completion is a rather neglected topic in psychological research, we suggest the concept of questionnaire experience (QX) as a new term for the systematic evaluation of the subjective perception of completing a questionnaire. It is related to the concept 'user experience' (International Organization for Standardization, 2019), which is a well-

established term in the field of interactive product design (Kujala et al., 2011; Sauro & Lewis, 2016; Wright et al., 2003). Given that the methodological framework outlined by the concept of UX provided considerable benefit to the design of interactive consumer products, we believe that similar benefits can be reaped from using the concept of QX in the field of questionnaire design. QX is conceptualized as the entire set of a person's emotions, beliefs, preferences, perceptions, physical and psychological responses and behaviours that result from responding to a questionnaire. QX is considered an umbrella term (Hirsch & Levin, 1999) that brings together a set of indicators which altogether allow us to capture the experience of humans when completing a questionnaire. We believe that the use of umbrella terms can be useful under certain circumstances (c.f. Sauer et al., 2020; Sonderegger et al., 2019). Adopting a respondent-centred approach (by capturing in broader terms the experience of the respondent during questionnaire completion), we presume that QX has not only an influence on the willingness and motivation of respondents to participate in the study, but also influences the primary psychometric properties of the scale (i.e. validity, reliability).

Figure 1 shows how QX has been conceptualized. It is important to distinguish between elements in the conceptualisation of QX, which can be measured (e.g., by means of a questionnaire) and those that cannot. This distinction is visualised in figure 1 by using a solid line to designate theoretical constructs (i.e. not directly measurable) and a dotted line to designate measurable indicators.

Figure 1. The conceptualisation of the new term 'questionnaire experience' describing its constituting elements (grey circles denote indicators that were measured in empirical study).



In the present work, we employed some indicators with a view of gaining a better understanding to what extent respondents experience verbal questionnaires and hybrid questionnaires differently. The measurable indicators used in the present work included respondent workload, respondent motivation, questionnaire preference, and questionnaire completion time. The constructs and measurable indicators subsumed under the term QX go far beyond the elements that could be examined in the present work. They refer to various aspects of how the respondent interacts with the questionnaire, emotional reactions elicited by the questionnaire's presentation or content, aesthetic appeal of the questionnaire, level of trust, the willingness to complete the questionnaire again in the future, and the level of comprehensibility of specific items. This set of elements is not exhaustive and further constructs and dimensions may be added. This conceptualisation is considered a first attempt to capture the meaning of QX.

1.2 Hybrid Scales

Hybrid scales represent a combination between verbal and pictorial scales. In contrast to an exclusively pictorial or an exclusively verbal scale, a hybrid scale can be defined as an instrument that makes use of both image-based and verbal elements to convey the meaning of its items (Sauer et al., in press).

A substantial number of validated instruments in the research literature match this definition of hybrid scales. Out of 57 pictorial instruments analysed in an overview article by Sauer and colleagues (in press), 27 were of a hybrid nature. In sleep research for instance, the Pictorial Epworth Sleepiness Scale (Ghiassi et al., 2011) uses verbal statements and verbal anchors in combination with illustrations to visualize each response option of the scale. Other instruments such as the Levonn Scale (Richters et al., 1990) or the Cameron Complex Trauma Interview (CCTI, King et al., 2017) make also use of verbal and pictorial content but in a different way. For both instruments (which were developed for children), the verbal part is read out by the scale administrator and the pictorial part is used to illustrate the meaning of the item and/or the rating scale.

In the domain of human-computer interaction, no hybrid scales have been developed yet, though a relatively impressive number of pictorial scales exist. Most of the pictorial instruments available have been designed to assess emotions/affect when using interactive products (e.g. Bradley & Lang, 1994; Desmet, 2003; Sonderegger et al., 2016). With regard to the assessment of usability, only two instruments have been developed and tested so far: a pictorial single-item usability scale (PSIUS, Baumgartner et al., 2019a), and a pictorial version of the SUS (P-SUS, Baumgartner et al., 2019b), which is based on the established System Usability Scale (Brooke, 1996).

The use of hybrid scale offers several advantages because they satisfy the following three criteria: (a) facilitated recognition, (b) redundancy gain, and (c) individual preferences in information processing. (a) By using both verbal and visual information together, recognition of the intended meaning of the scale is easier (Ghiassi et al., 2011). It is essential that both cues provide congruent information. It follows a similar idea that is common practice in software design, which uses both a meaningful label and a well-chosen icon to facilitate recognition and comprehension of actions and controls (Harley, 2014; Wiedenbeck, 1999). (b) A further advantage lies in the representation of redundant information (be it in the verbal or in the visual part, following the principle of redundancy gain; e.g. Backs & Walrath, 1995). If one of the two parts has an unclear meaning, the other may help clarify the meaning, thus alleviating the negative effects of ambiguity. (c) When both, verbal and pictorial information is presented, respondents can choose how they would like to pay attention to the different modalities (i.e. verbal and pictorial). There is evidence from research that learning content (texts and images) that corresponds to the cognitive style of the participant (verbalizer vs. visualizer) is preferred by learners and better remembered (Koć-Januchta et al., 2017). Thus, an advantage of hybrid scales might be that they offer both a verbal and a pictorial access for both cognitive styles.

The use of hybrid scales might also be associated with two disadvantages. (a) Since content is presented in verbal and pictorial form, information processing might be slowed down. This might increase questionnaire completion time, due to the additional content that has to be decoded before a proper rating can be made. Completion time may depend on the length of the verbal item and the complexity of the pictorial item. (b) Since both pictorial and verbal content is presented, ambiguity might increase.

1.3 Aim of the Research and Hypotheses

In this study, a hybrid usability questionnaire is used to assess perceived usability. Usability is specified in the ISO norm 9241-11, describing that a user should achieve a specific goal in a specific context in an effective, efficient and satisfied way (International Organization for Standardization, 2016). A considerable number of validated verbal instruments is available for the measurement of

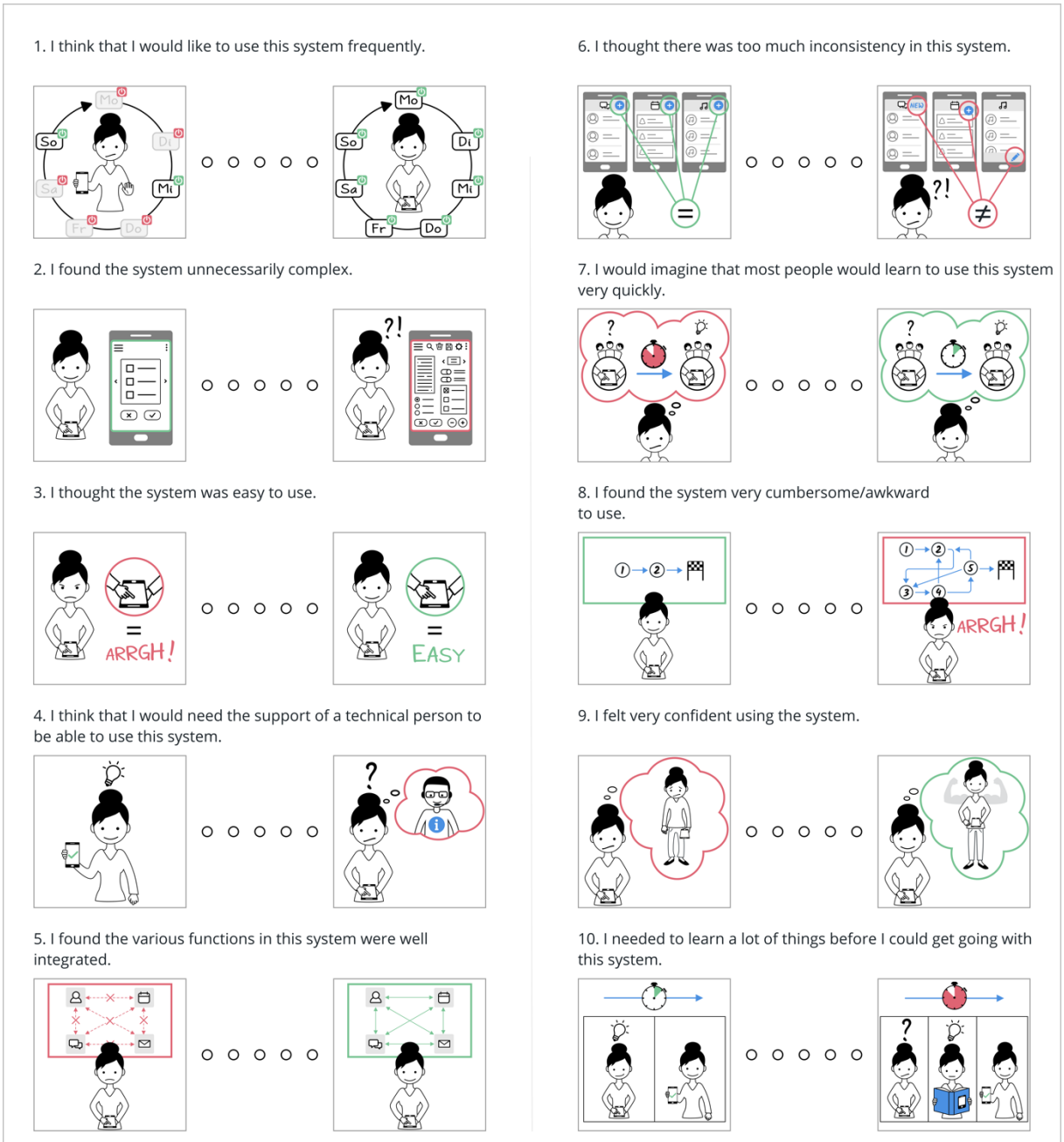
perceived usability, with each having its merits and drawbacks (for a recent overview see: Assila & Ezzedine, 2016). One of the most widely used instruments is the System Usability Scale (SUS, Brooke, 1996), which provides a general usability estimate based on ten items. Since there is little empirical work about the advantages and disadvantages of hybrid scales, the aim of this article is to evaluate a hybrid version of the SUS and to compare it to its verbal origin. As part of this comparative evaluation, we rely not only on classic psychometric criteria (reliability, validity, sensitivity, etc.), but also assess criteria that are not typically considered in scale development, such as perceived questionnaire workload, respondent motivation, questionnaire preference, and questionnaire completion time, which we subsume under the term of QX.

We hypothesized that a hybrid scale would have similar psychometric properties (i.e. convergent, divergent and criterion-related validity) compared to the verbal version. Furthermore, we assumed that using a hybrid scale would result in higher scores in measures of QX.

2 Hybrid System Usability Scale (H-SUS)

The items of the Hybrid System Usability Scale (H-SUS) combine pictorial and verbal information in the same scale (see figure 2).

Figure 2. H-SUS items (female version) with verbal content and the five-point rating scale using pictorial representations for the positive and negative end points.



The pictorial information consists of two visual representations, which depict the extreme points of a bipolar scale. An avatar is presented interacting with a mobile device in a specific usage situation (negative vs. positive experience). In between, a five-point Likert scale is provided for the ratings to be given. The verbal content is placed above the pictorial scale, containing the exact wording of the specific SUS item. The pictorial content of the H-SUS was based on the Pictorial System Usability Scale (P-SUS, Baumgartner et al., 2019b). The scale was designed to match as closely as possible the verbal content of corresponding SUS item. A male and a female version of the avatar were developed with identical content to increase respondent identification with the scale.

3 Online Validation Study

3.1 Goal of the Validation Study

The first goal of the validation study was to determine the psychometric properties of H-SUS with a view to comparing it to the well-established verbal SUS. The psychometric properties assessed included convergent validity, divergent validity, criterion-related validity, sensitivity, and reliability in the form of internal consistency. The second goal was to apply the concept of QX in scale design by comparing the two instruments with regard to measures of QX. The concept was assessed by subjective ratings (i.e. respondent workload, respondent motivation, questionnaire preference) but also by objective measures such as questionnaire completion time. In order to be able to assess these concepts, participants took part in an online usability test, in which they interacted with a smartphone prototype. Subsequently, they completed several questionnaires needed to meet the two goals of the study.

3.2 Method

3.2.1 Participants

Participants were recruited in the following ways: (a) an email was sent to all bachelor and master students of the University of Fribourg, (b) an advertisement was placed on the website of the German-language magazine 'Psychologie Heute', (c) a link was sent to a school teacher of a class in computer science, whose school classes took part in the study, and (d) the link was shared within the social networks of the experimenters. In addition, participants were asked at the end of the study to forward the link to their friends. To increase participant motivation, five vouchers worth €50 each were raffled.

A total of 152 participants (73% female) took part in the online study, with their ages ranging from 16 to 78 years (M = 28.11 yrs., SD = 13.90). The sample consisted of 95 students (62.5%), 29 employees (19.1%), 19 pupils (12.5%), and 9 participants choosing the option 'other' as their professional status (5.9%). Two participants (1.3%) reported to have some form of colour blindness.

Participants rated the frequency of using a smartphone as high (M = 4.51, SD = 0.87) on a five-point Likert scale ranging from 1 (very rarely) to 5 (very often). They rated their experience in using smartphones similarly high (M = 4.22, SD = 0.79) on a five-point Likert scale ranging from 1 (very low) to 5 (very high).

3.2.2 Measures and Instruments

Several measures were used in this study. This comprised measures for the assessment of psychometric properties, such as (1) convergent validity, (2) divergent and (3) criterion-related validity, (4) reliability and (5) sensitivity. Furthermore, it consisted of measures of QX, such as (6) respondent workload, (7) respondent motivation to complete the questionnaire, (8) questionnaire preference, and (9) questionnaire completion time.

Convergent Validity. It is considered a part of construct validity, describing the relationship between two different measures that aim to capture the same construct (Messick, 1979). Since they measure the same construct, high correlations between convergent measures are to be expected. As a measure of convergent validity, the verbal SUS was used. It consists of ten items, on which usability is rated on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). A usability score is calculated by aggregating the ratings (for the detailed computing procedure see Brooke, 1996). Good psychometric properties were reported in several studies (Cronbach's $\alpha > .90$, e.g. Bangor et al., 2009; Brooke, 2013). Since the study was conducted in the German-speaking part of Switzerland and in Germany, a German version of the SUS was used (Rummel, 2015).

Divergent Validity. It refers to the idea that there should not be a relationship between measures that are not conceptually related (Messick, 1979). As a result, rather low correlations between divergent measures are to be expected. To compute divergent validity, the concepts of affect and visual aesthetics were assessed. Affect was measured using the AniSAM (Sonderegger et al., 2016), which is a nonverbal instrument based on the Self-Assessment Manikin (SAM; Bradley & Lang, 1994). It consists of two pictorial items assessing valence and arousal. The item for valence depicts a manikin with a facial expression that ranges from frowning to smiling on five levels. The item for arousal depicts the selected level of valence and adds an animated heart as indicator for physiological arousal. The intensity of arousal is indicated by the frequency by which the heart bumps. For the assessment of visual aesthetics, the short version of the Visual Aesthetics of Websites Inventory (VisAWI-S) was used (Moshagen & Thielsch, 2013). This instrument measures the four underlying facets of visual aesthetics with one item each: simplicity, diversity, colourfulness and craftsmanship. The wording of the items was slightly modified, replacing the term 'website' with the name of the device tested (i.e. 'smartphone'). Being evaluated in three studies with large samples (N=764, N=305, N=604), the psychometric properties of the VisAWI-S are considered to be good (Cronbach's α = .81).

Criterion-related Validity. It refers to the relationship between a measure in question and an external objective measure, such as a performance measure (Coolican, 2017). Previous research showed that medium-sized correlations are to be expected when comparing subjective usability with objective performance measures (Baumgartner et al., 2019a; Baumgartner et al., 2019b). In this study, task completion time (in seconds) and the number of user interactions with the prototype interface were used as external criteria.

Reliability. As a measure of reliability, internal consistency was computed. It describes how the items of a questionnaire relate to each other (Coolican, 2017). It was calculated for H-SUS and SUS using Cronbach's Alpha (Hinkin, 1995).

Sensitivity. Sensitivity is considered the extent to which differences can be detected by an instrument when an independent variable (such as usability) is manipulated (Lewis, 2002, 2018). An instrument that measures the underlying construct should be sensitive to these differences and consequently reflect them in the scores obtained. Sensitivity was assessed in this study for H-SUS and SUS by comparing group means of the high-usability condition with the low-usability condition. The sensitivity of SUS has already been demonstrated in previous studies (Bangor et al., 2008; Kortum & Bangor, 2013).

Respondent Workload. The workload for questionnaire completion was assessed using a single-item scale ('It was exhausting for me to respond to the questions.'), which was presented after completion of the H-SUS and the SUS. A single item was used to reduce questionnaire length and because it is capable of assessing the main concept it intends to measure (Wanous et al., 1997). Participants rated on a five-point Likert scale ranging from 1 (totally disagree) to 5 (totally agree).

Respondent Motivation. To assess the motivation of questionnaire completion, the short version of the Intrinsic Motivation Inventory (IMI) was used (Wilde et al., 2009). The short version of IMI captures four different types of intrinsic motivation: Interest/pleasure, perceived competence, perceived freedom of choice, and pressure/tension. According to Deci and Ryan (2003), interest/pleasure is regarded as self-experience value for intrinsic motivation. For this reason, only this three-item subscale was used in this study. The three items (fun, joy, and interest in completing a questionnaire) make use of a five-point Likert scale ranging from 1 (totally disagree) to 5 (totally agree). Wilde and colleagues (2009) reported good internal consistency for this subscale (α = .85).

Questionnaire Preference. Participants were asked at the end of the survey, which questionnaire type they preferred. A bipolar single-item five-point Likert scale with three adjective anchor points (1: verbal questionnaire; 3: both; 5: picture questionnaire) was presented to assess participants' preference.

Questionnaire Completion Time. Completion time for each item was automatically recorded by the online questionnaire. Completion times for all items were aggregated for the H-SUS and SUS separately.

3.2.3 Prototype, User Tasks and Pilot Study

Prototype. A web-based smartphone prototype was developed to allow participants to interact online. It was based on the prototype developed by Hamborg et al. (2014), but the design was changed to a more modern appearance, offering a contemporary technical specification that ensures its compatibility with current browsers. Two versions of the prototype were provided for this study: a high-usability and a low-usability one. The two versions differed regarding navigation structure (simple vs. complicated), whereas all visual and aesthetical elements were identical.

User Tasks. Participants were asked to perform three tasks on the smartphone prototype: (a) creating a new entry in the address book, (b) retrieving the last phone bill, and (c) changing the ringtone of the smartphone. Two performance measures (task completion time and number of user interactions) were recorded automatically during task completion.

Pilot study. A pilot study was carried out prior to the online validation study to test whether the manipulation of usability succeeded. Twenty participants (Age: M = 31.20 yrs., SD = 14.74; 70% female, Occupation: 10 students, 6 employees, 4 others) interacted either with the high-usability prototype or the low-usability one, and subsequently rated its usability using the SUS. The assignment of participants to the high or low-usability condition was counterbalanced. Interpreting the SUS scores using the grades of the curve grading scale (CGS) proposed by Lewis and Sauro (2017), low usability corresponded to a 'C grade' (M_{low} = 65.33, SD = 23.78), whereas high usability corresponded to a 'A+ grade' (M_{high} = 92.50, SD = 4.18). The Mann-Whitney test showed a significant difference between low and high-usability conditions (Mdn_{high} = 14.10, Mdn_{low} = 6.90, U = 14.00, z = −2.734, p = .005, r = −0.611), confirming that the experimental manipulation of usability was successful.

3.2.4 Experimental Design

A one-factorial between-subjects design was implemented, with system usability as the independent factor being varied at two levels: low vs. high. Furthermore, the order of administering the questionnaires was counterbalanced (i.e. half of participants completed H-SUS first, the other half SUS first).

3.2.5 Procedure

The study was conducted using an online questionnaire platform. It typically took participants between 10 to 15 minutes to carry out the tasks and to complete the online questionnaire. On the first page, participants were presented an image of a male and a female avatar. By clicking, they selected the gender with which they most likely identified themselves. After receiving instructions and providing their informed consent, participants were explained how to interact with the smartphone prototype. The prototype was displayed in a separate browser window together with the three tasks to be completed. Before and after the interaction with the prototype, participants were asked to rate their level of arousal and valence with the AniSam. Before participants could continue with the questionnaire, they were asked whether they had completed all three tasks with the prototype. Then, the visual aesthetics of the prototype was assessed by using the short version of VisAWI. This was followed by participants completing the SUS and the H-SUS. In order to avoid carry-over effects, the sequence of these two questionnaires was counterbalanced. Before each questionnaire, the instruction was given that the following questions refer to the interaction with the prototype. Prior to processing the H-SUS, participants were presented an example item to give them an idea of the new questionnaire type (i.e. they were shown a verbal question and the pictographic representation). Furthermore, they were explained how to give their response on the scale between the two images. After each questionnaire, participants responded to an item assessing workload and the three items of the IMI

(fun, joy and interest). Finally, questions were asked about the preference for the hybrid-based or verbal-based questionnaire. In a comment field, participants could enter suggestions or improvements for the study. If they were interested in participating in a follow-up study, they could enter their email address in another field. On the last page, the participants were thanked, given information about the raffle and asked to forward the email to other interested persons.

3.2.6 Exclusion Criteria

Prior to data analysis, the following set of criteria was defined, which specified under what circumstances datasets of participants are to be excluded: (1) Participants providing incomplete datasets were excluded. (2) Participants having completed the online study more than once were excluded. (3) Participants who responded ’no’ to at least one of the two control items (‘Did you do the three tasks with the prototype?’ and ‘Did you complete the questionnaires seriously?’) were excluded. (4) Participants who took more than 40 minutes to complete the study were excluded. A total of 11 participants were excluded according to the criteria just described.

3.2.7 Data Treatment

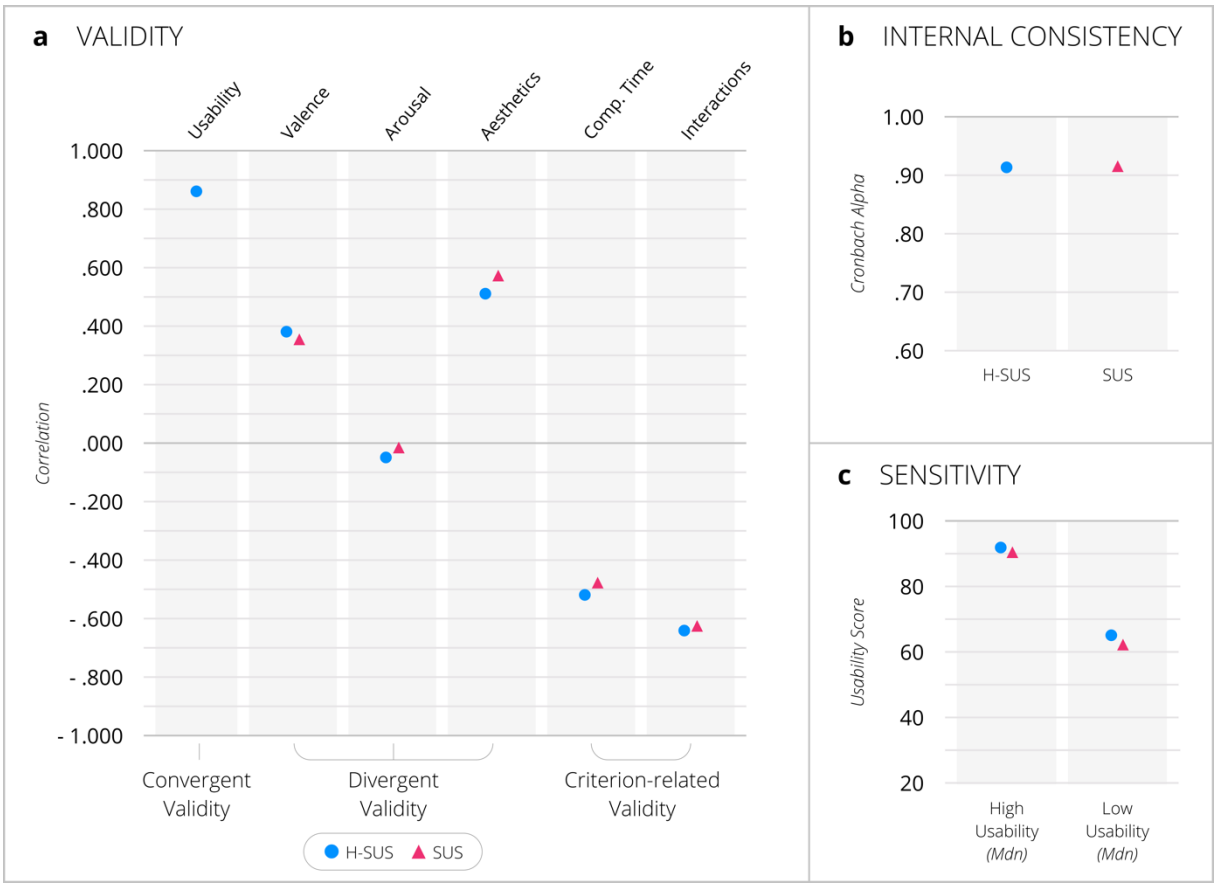
Whenever requirements for normal distribution and homogeneity of variance were violated, non-parametric tests were used. Correlational analyses were used for the calculation of convergent, divergent and criterion-related validity by using Spearman’s rank correlation coefficient. Comparisons of group means were carried out to determine sensitivity by using Mann-Whitney *U*-test. Reliability in the form of internal consistency was determined by calculating Cronbach's alpha. Finally, frequency analyses were used to determine questionnaire preference in the form of descriptive percentages. We set the level of significance for all analyses to 5 %.

3.3 Results

3.3.1 Psychometric Criteria

Psychometric criteria of the tested instruments are described in the following paragraphs. Figure 3 summarizes the main results of the analyses of the psychometric criteria.

Figure 3. Comparative analysis of psychometric criteria of H-SUS and SUS: (a) Correlations of usability scores with scores of convergent, divergent and criterion-related validity, (b) internal consistency score and (c) sensitivity score.



3.3.1.1 *Convergent Validity*

In figure 3a, the scores for convergent validity of H-SUS with SUS are presented. The detailed item-based analyses are presented in table 1, together with the usability score. The results show largely high correlation coefficients. Nine out of ten items showed correlations of $r > .600$ and the overall usability score reached an even higher correlation ($r = .862$).

Table 1. Spearman correlation coefficients between H-SUS and SUS on item level and overall usability score (N=152).

	Item-based correlations between H-SUS and SUS (N=152)										
	01	02	03	04	05	06	07	08	09	10	Overall score
r	.660***	.729***	.759***	.544***	.762***	.751***	.803***	.745***	.694***	.644***	.862***

*Note: * p < .05; ** p < .01; *** p < .001*

3.3.1.2 *Divergent Validity*

Correlational analyses for the evaluation of divergent validity were conducted (see table 2). The results for valence showed significant small to medium-sized correlations of around $r < .400$. Concerning arousal, non-significant correlations were obtained.

With regard to aesthetics, significant correlations of around $r = .500$ were observed. As expected, measures of divergent validity tended to have a smaller score than measures of convergent validity.

Table 2. Correlations of aesthetics (VisAWI) and affect (AniSAM) with H-SUS and SUS (N=152).

	Valence (AniSAM)	Arousal (AniSAM)	Aesthetics (VisAWI)
	r	*r*	*r*
H-SUS	.378***	−.050	.507***
SUS	.348***	−.025	.569***

*Note: * p < .05; ** p < .01; *** p < .001*

3.3.1.3 Criterion-related Validity

For the assessment of criterion-related validity, correlations of performance measures (task completion time and number of interactions) with both H-SUS versions and SUS were analysed (see table 3). We found significant negative correlations with task completion time of around $r = -.500$ for H-SUS. Similar results were obtained for the SUS evaluation. With regard to number of interactions, correlations were similar for the H-SUS and SUS, at around $r = -.600$.

Table 3. Correlations of performance (task completion time and number of interactions with the prototype) with H-SUS and SUS (N=152).

	Task Completion Time	Number of Interactions
	r	*r*
H-SUS	−.521***	−.639***
SUS	−.484***	−.632***

*Note: * p < .05; ** p < .01; *** p < .001*

3.3.1.4 Internal Consistency

Figure 3b shows Cronbach Alpha values for all instruments, which were calculated using all items. Analysis of reliability revealed high internal consistency for the H-SUS ($\alpha = .91$). Similarly, a high internal consistency score was found for the SUS ($\alpha = .91$).

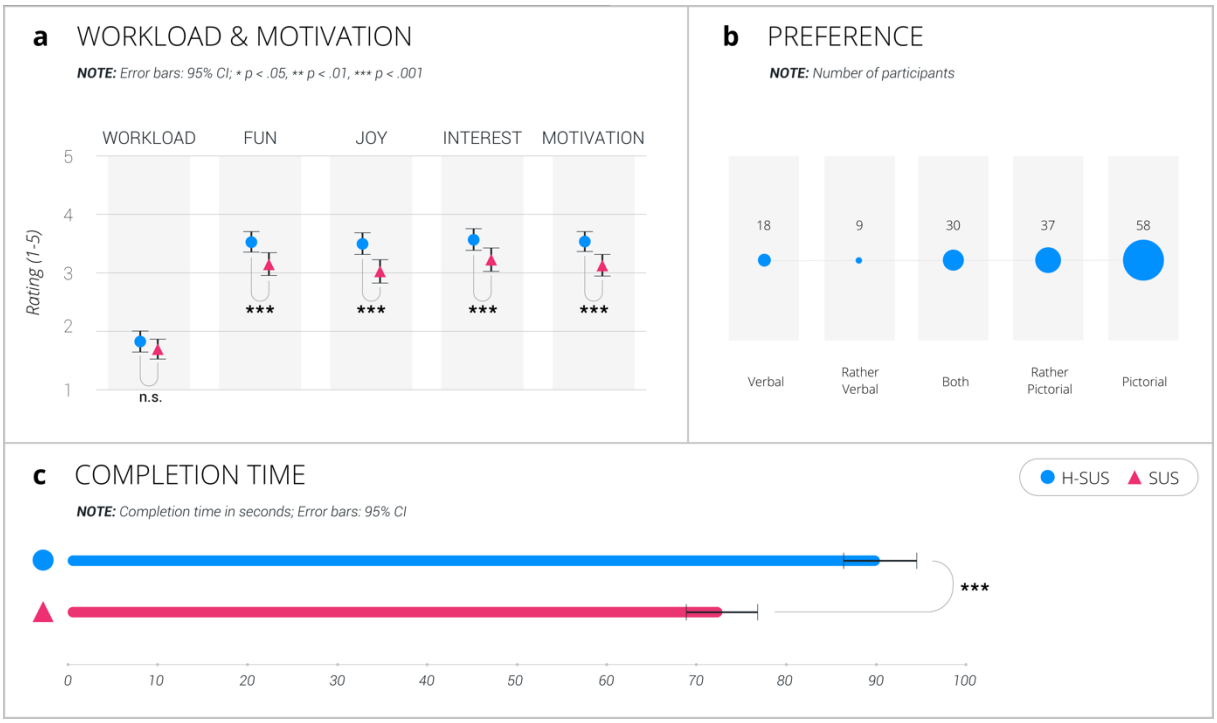
3.3.1.5 Sensitivity

In Figure 3c, usability scores in low and high-usability conditions are presented for H-SUS and SUS. To assess whether there is a difference between low and high usability, a Mann-Whitney test was carried out. Analysis showed highly significant differences for H-SUS ($Mdn_{high} = 92.50$, $Mdn_{low} = 65.00$, $U = 798.50$, $z = -7.71$, $p = .000$, $r = -0.626$), as well as for SUS ($Mdn_{high} = 90.00$, $Mdn_{low} = 62.50$, $U = 792.00$, $z = -7.74$, $p = .000$, $r = -0.628$). H-SUS and SUS were both sufficiently sensitive to distinguish between levels of low and high usability.

3.3.2 Questionnaire Experience

The analysis of QX is described in the following paragraphs. Figure 4 summarizes the main results of analysing the different QX measures.

Figure 4. Comparative analysis of different dimensions of questionnaire experience for H-SUS and SUS: (a) Respondent workload and motivation, (b) preference and (c) questionnaire completion time.



3.3.2.1 *Respondent Workload and Motivation*

Figure 4a summarizes the descriptive data for respondent workload and motivation. For the analysis of perceived respondent workload and motivation, Wilcoxon signed-rank tests were carried out.

The results showed no significant difference for respondent workload of H-SUS compared to the one of SUS ($Mdn_{H\text{-}SUS}$ = 1.00, Mdn_{SUS} = 1.00, z = −1.367, p = .171, r = −0.115). However, there were large effects for motivation. All three items obtained higher scores for the H-SUS than for the SUS, which resulted in a significant difference on the IMI overall score ($Mdn_{H\text{-}SUS}$ = 3.67, Mdn_{SUS} = 3.00, z = −4.858, p = .000, r = −0.408). The ratings of workload and motivation are shown in table 4.

Table 4. Means, standard deviations and p-values for workload and motivation of H-SUS and SUS (N=152); IMI: Intrinsic Motivation Inventory.

	H-SUS *M (SD)*	SUS *M (SD)*	*p*
Workload (1-5)	1.82 (1.05)	1.68 (1.01)	.171
IMI item 1 - fun (1-5)	3.53 (1.05)	3.15 (1.17)	.000***
IMI item 2 - joy (1-5)	3.50 (1.09)	3.03 (1.18)	.000***
IMI item 3 - interest (1-5)	3.57 (1.08)	3.22 (1.18)	.000***
IMI Overall Score	3.53 (0.98)	3.13 (1.09)	.000***

*Note: * p < .05; ** p < .01; *** p < .001*

3.3.2.2 *Questionnaire Preference*

The results of the questionnaire preference rating (see figure 4b) showed that about two thirds of the participants (62.5%) preferred the H-SUS, whereas 17.8% of participants favoured the SUS. 19.7% of participants liked both questionnaires.

3.3.2.3 Questionnaire Completion Time

Completion time was recorded for each item and aggregated to questionnaire completion time. In order to control for the unwanted effect of participant interruption, participants were excluded from the analysis when they spent more than 60 seconds on an item. As a result, 10 participants were excluded from the analysis. Item completion time and total questionnaire completion time are shown in table 5.

Table 5. Means, standard deviations and p-values for completion time (in seconds) for H-SUS and SUS (*N*=142).

	H-SUS *M (SD)*	SUS *M (SD)*	*p*
Item 01	14.42 (5.88)	8.35 (4.58)	.000***
Item 02	9.88 (5.11)	7.32 (5.21)	.000***
Item 03	6.77 (3.42)	5.92 (2.51)	.002**
Item 04	9.99 (4.62)	7.71 (3.52)	.000***
Item 05	8.29 (3.91)	8.11 (4.71)	.635
Item 06	10.61 (5.06)	7.97 (5.39)	.000***
Item 07	8.64 (4.36)	7.75 (4.33)	.001**
Item 08	7.32 (3.35)	6.46 (3.86)	.002**
Item 09	6.46 (2.98)	6.32 (2.91)	.514
Item 10	8.01 (3.41)	6.98 (4.36)	.000***
Total	90.40 (24.39)	72.91 (23.78)	.000***

*Note: * p < .05; ** p < .01; *** p < .001*

The results indicated that participants needed about 20 seconds longer to complete the H-SUS than the SUS. Wilcoxon signed-rank tests indicated that this difference was statistically significant (see table 5 for p-values). Furthermore, Wilcoxon signed-rank tests revealed that out of the ten items only items 5 and 9 of the H-SUS did not differ significantly from the SUS (both $p > .500$).

4 Discussion

The aim of this study was to compare H-SUS to the verbal SUS with regard to their psychometric properties. In addition to the classic measures of psychometric quality, this comparative test also included various measures of QX. When examining the indicators that allow making a comparison between the two scales (i.e. divergent validity, criterion-related validity, reliability in the form of internal consistency, and sensitivity), it showed that the psychometric properties of H-SUS were overall of a similar quality than the ones of the established verbal SUS scale, which served as a kind of benchmark. With regard to the indicators of QX, the findings showed overall that the H-SUS had better scores than the SUS for most subjective ratings, whereas the SUS emerged as the better alternative when considering objective QX measures (e.g., questionnaire completion time).

With regard to convergent validity, we recorded a very high correlation between overall scores of H-SUS and SUS ($r > .800$). Furthermore, an analysis at the item level revealed that for nine out of ten items, correlations between SUS and H-SUS were larger than $r > .600$. Overall, the items of the H-SUS showed very high correlations, which may be considered a large effect (based on the recommendations of Cohen, 1988). This may be less surprising given that the H-SUS shares many elements with the SUS. However, the high convergent validity score may suggest that the one of

concerns raised in the literature review about hybrid scales (i.e. increased ambiguity if verbal and pictorial content does not match) may be unfounded in the case of the H-SUS.

With regard to divergent validity, the results for H-SUS and SUS showed very similar correlation coefficients for all three measures of divergent validity, suggesting that both instruments showed similar psychometric qualities with regard to this type of validity. Furthermore, the analysis of the three validity scores showed overall that correlations were lower than for the measures of convergent validity. This result is partially in line with the principles underlying the notion of divergent validity, which presumes that there should be no association between measures that are not conceptually related (Messick, 1979). For both H-SUS and SUS, the correlations coefficients were higher for aesthetics ($r \approx .55$) than for valence ($r \approx .35$) and arousal ($r \approx -.05$). The reason why aesthetics as a measure of divergent validity had a rather high score may be explained by particularities of this concept. Empirical evidence from research on aesthetics suggests a close relationship between user ratings of usability and of the aesthetic appeal of a device (e.g., Hamborg et al., 2014; Tuch et al., 2012). This relationship is often described as the "what is beautiful is good"-effect (Tractinsky et al., 2000). Owing to this close relationship, a higher score for aesthetics than for the other two measures of divergent validity may not come as a surprise. Nevertheless, we believe the results to be sufficient for divergent validity, although the choice of aesthetics as a measure of divergent validity should be reconsidered for future research in the usability domain.

With regard to criterion-related validity, the correlation coefficients for the H-SUS and SUS were very similar, suggesting again that the psychometric properties of both instruments were of similar quality. Furthermore, the results for the H-SUS revealed highly significant correlation coefficients (between $r = -.500$ and $r = -.600$) for both performance measures (i.e. task completion time and number of interactions). The correlation coefficients are generally slightly lower for criterion-related validity than for convergent validity. In addition to this general difference between the two types of validity, there are domain-specific aspects to be considered. In the usability domain, evidence from meta-analyses suggests a substantial relationship between perceived usability and objective performance measures, ranging from $r = .35$ to $r = .60$ (Nielsen & Levy, 1994; Sauro & Lewis, 2009). Few validation studies of scales assessing perceived usability have included criterion-related validity as an indicator of their psychometric quality. The validation studies of two pictorial usability scales revealed much smaller coefficients of criterion-related validity in one study (Baumgartner et al., 2019a) and similar coefficients in the other (Baumgartner et al., 2019b), compared to the present work. There is a need for future research to investigate in more detail the effect patterns, and the circumstances under which lower or higher effect sizes are to be expected. Considering the available findings of the two meta-analyses and the two studies cited, we regard the criterion-related validity of the H-SUS to be satisfactory.

There has been convergent evidence from the three validity coefficients (i.e. convergent, divergent and criterion-related) that the H-SUS has very similar psychometric properties than the SUS as the established scale being used a benchmark. This converging evidence is also supported by the results for internal consistency and sensitivity. Concerning internal consistency, both instruments achieved Alpha values in the same range (all $\alpha > .90$), which indicates excellent internal consistency (DeVellis, 2016). With regard to sensitivity, we found for both instruments highly significant differences between low and high-usability condition. Therefore, both instruments are considered sufficiently sensitive of distinguishing between low and high levels of usability.

Having examined indicators traditionally used for evaluating the psychometric properties of scales, we will now discuss the results obtained from indicators summarised under the conceptual umbrella of QX, which are not very often considered when determining the quality of a scale. The analysis of respondent workload indicated no significant difference between H-SUS and SUS, which suggests that concerns that a hybrid scale might lead to a considerably higher information load may have been unfounded. With regard to respondent motivation, the H-SUS obtained significantly higher scores than

the SUS, which indicates that participants appreciated completing the H-SUS more than the SUS. In line with the results for motivation, preference ratings also showed that a clear majority of respondents preferred the H-SUS to the SUS. However, completion time was significantly longer for the H-SUS compared to the SUS by about 20 seconds, which may be interpreted as respondents requiring more time to scan both verbal and pictorial information. Interestingly, the analysis at the item level revealed that the biggest difference was found for the first item, which may have due to the fact that this type of questionnaire was new to most participants (even if a sample item had been given for practice prior to it). Overall, the analysis of the QX measures revealed considerable evidence at the subjective level for the H-SUS being the better alternative, though at the expense of increasing questionnaire completion time.

The present work has a number of limitations. The first limitation refers to the test setting. Since the H-SUS was tested in an online study, it was not possible to standardize the testing procedure to the same extent, as it would have been possible in a lab-based study. For example, test participants may have used different devices (e.g., laptop, tablet, smartphone) and the environmental conditions may have varied (visual and auditory distractions, short interruptions, etc.). All these factors may have contributed to a higher variance of test scores. A second limitation refers to the assessment of convergent validity, which relied on the SUS as the only measure. Using a further scale assessing perceived usability (e.g., PSSUQ; Lewis, 2002) could have strengthened confidence in the results on convergent validity. However, the very high correlation between the two scales suggests that the H-SUS is very similar to the SUS with regard to this form of validity, which is expected to be mainly due to the two scales sharing the verbal content of item formulation.

Based on the experience gained in the development of this hybrid questionnaire, we would like to make a number of suggestions for future work making use of pictorial content in scale development. (a) When developing a scale with pictorial content, it should be considered visualizing only some items of a standardised verbal questionnaire rather than all items (as it was the case in this study). We would recommend selecting those items that are less ambiguous and easier for participants to understand. Lewis (2017) already demonstrated for the verbal SUS that it would be possible to obtain comparable results even if one of the items was removed. Alternatively, suitable items could be taken from different usability questionnaires to create a new pictorial usability scale based on the best fitting items of all verbal instruments. (b) A different approach could also be used for the validation procedure of pictorial scales. For example, rather than having to rely entirely on the convergent validity coefficient to assess the quality of a pictorial item, the validity could be evaluated, in addition, by means of extensive comprehension tests with heterogeneous samples. (c) Future studies should consider elaborating the concept of QX, notably by identifying further suitable measures that would fit under this umbrella. This may result in the development of a standardised instrument, which would provide questionnaire developers with a tool to measure QX. This tool could be employed to capture QX for established instruments but also when developing new ones. For this purpose, benchmarks and cut-off values for QX would be highly valuable. (d) Finally, there is a need for future studies that involve cross-cultural testing. This is because the visual elements are not always understood in the same way across different countries and cultures. Often, the comprehension of visual elements depends strongly on whether the symbol is used in one's own culture or not (Chu, 2003; Knight et al., 2009).

6 Conclusion

This study is the first that examined the psychometric properties of hybrid scales compared to traditional verbal scales by making use of an additional set of quality indicators (integrated under the umbrella of QX) that go beyond the indicators traditionally used for that purpose (e.g., convergent and divergent validity, criterion-related validity, and sensitivity). Using a large and heterogeneous sample (comprising students, professionals and pensioners), the methodological approach also considered the

identification of the respondent with the gender of the avatar by allowing them to choose between different options. Considering the findings of the present work, we can overall conclude that a hybrid version of a scale can obtain good psychometric properties being comparable in quality to a verbal scale. At the same time, the subjective components of QX have improved for the hybrid version, which may result in higher commitment and motivation when completing questionnaires. The only drawback of the hybrid version was that questionnaire completion time has increased by an average of two seconds per item. Nevertheless, the H-SUS represents a viable alternative to the well-established verbal version of the SUS. With regard to QX, its assessment offers some potential for the development of future questionnaires, be it a verbal one, a hybrid one, or a pictorial one. The list of components of QX assessed in this study is not exhaustive. It should rather be seen as a starting point for developing the concept further. We believe that the assessment of QX will help us identify better how the psychometric properties of an instrument can be improved. We assume that improvements based on QX in turn affect the traditional psychometric properties in a positive way and help to gain more confidence when choosing an appropriate instrument.

7 Acknowledgements

The research was funded by a grant (No 100019_188808) from the Swiss National Science Foundation (SNSF) and was also supported by We Are Cube / Puzzle ITC. Their support is gratefully acknowledged. We are very grateful to Veronica Solombrino and Mayra Overney-Falconí for the numerous design reviews and the valuable feedback during the development process, and to Quentin Meteier for the implementation of the smartphone prototype.

8 References

Assila, A., & Ezzedine, H. (2016). Standardized usability questionnaires: Features and quality focus. *Electronic Journal of Computer Science and Information Technology: EJCIST*, *6*(1).

Backs, R. W., & Walrath, L. C. (1995). Ocular measures of redundancy gain during visual search of colour symbolic displays. *Ergonomics*, *38*(9), 1831–1840.

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.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, *24*(6), 574–594.

Baumgartner, J., Frei, N., Kleinke, M., Sauer, J., & Sonderegger, A. (2019b). Pictorial System Usability Scale (P-SUS): Developing an Instrument for Measuring Perceived Usability. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 69.

Baumgartner, J., Sonderegger, A., & Sauer, J. (2019a). 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

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7.

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

Chu, S. (2003). Cross-cultural comparison of the perception of symbols. *Journal of Visual Literacy*,

23(1), 69–80.

Cohen, J. (1988). The effect size. *Statistical Power Analysis for the Behavioral Sciences*, 77–83.

Coolican, H. (2017). *Research methods and statistics in psychology*. Psychology Press.

Deci, E. L., & Ryan, R. M. (2003). Intrinsic motivation inventory. *Self-Determination Theory*, *267*.

Desmet, P. (2003). Measuring emotion: Development and application of an instrument to measure emotional responses to products. In *Funology* (pp. 111–123). Springer.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage publications.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360.

Ghiassi, R., Murphy, K., Cummin, A. R., & Partridge, M. R. (2011). Developing a pictorial Epworth sleepiness scale. *Thorax*, *66*(2), 97–100.

Gould, J. D., & Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM*, *28*(3), 300–311.

Hamborg, K.-C., Hülsmann, J., & Kaspar, K. (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

Harley, A. (2014, July 27). *Icon Usability*. Nielsen Norman Group. https://www.nngroup.com/articles/icon-usability/

Hinkin, T. R. (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

Hirsch, P. M., & Levin, D. Z. (1999). Umbrella advocates versus validity police: A life-cycle model. *Organization Science*, *10*(2), 199–212.

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

International Organization for Standardization. (2019). *Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems (Standard No. 9241-210)*. https://www.iso.org/standard/77520.html

King, J. A., Solomon, P., & Ford, J. D. (2017). The Cameron Complex Trauma Interview (CCTI): Development, psychometric properties, and clinical utility. *Psychological Trauma: Theory, Research, Practice and Policy*, *9*(1), 18–22. https://doi.org/10.1037/tra0000138

Knight, E., Gunawardena, C. N., & Aydin, C. H. (2009). Cultural interpretations of the visual meaning of icons and images used in North American web design. *Educational Media International*, *46*(1), 17–35.

Koć-Januchta, M., Höffler, T., Thoma, G.-B., Prechtl, H., & Leutner, D. (2017). Visualizers versus verbalizers: Effects of cognitive style on learning with texts and pictures–An eye-tracking study. *Computers in Human Behavior*, *68*, 170–179.

Kortum, P. T., & Bangor, A. (2013). Usability ratings for everyday products measured with the System Usability Scale. *International Journal of Human-Computer Interaction*, *29*(2), 67–76.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., & Sinnelä, A. (2011). UX Curve: A

method for evaluating long-term user experience. *Interacting with Computers*, *23*(5), 473–483.

Lewis, J. R. (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

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

Lewis, J. R., & Sauro, J. (2017). Can i leave this one out?: The effect of dropping an item from the sus. *Journal of Usability Studies*, *13*(1), 38–46.

Messick, S. (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

Miller, L. A., & Lovler, R. L. (2018). *Foundations of psychological testing: A practical approach*. Sage Publications.

Moshagen, M., & Thielsch, M. (2013). A short version of the visual aesthetics of websites inventory. *Behaviour & Information Technology*, *32*(12), 1305–1311.

Nielsen, J., & Levy, J. (1994). Measuring Usability: Preference vs. Performance. *Commun. ACM*, *37*(4), 66–75. https://doi.org/10.1145/175276.175282

Richters, J. E., Martinez, P., & Valla, J. P. (1990). Levonn: A cartoon-based structured interview for assessing young children's distress symptoms. *Bethesda, MD: National Institute of Mental Health*.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*(2), 151–161.

Rummel, B. (2015, January 12). *System Usability Scale – jetzt auch auf Deutsch.* SAP User Experience Community. https://experience.sap.com/skillup/system-usability-scale-jetzt-auch-auf-deutsch/

Sauer, J., Baumgartner, J., Frei, N., & Sonderegger, A. (in press). Pictorial scales in research and practice: A review. *European Psychologist*.

Sauer, J., Sonderegger, A., & Schmutz, S. (2020). Usability, user experience and accessibility: Towards an integrative model. *Ergonomics*, *just-accepted*, 1–23.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1609–1618.

Sonderegger, A., Heyden, K., Chavaillaz, A., & Sauer, J. (2016). AniSAM & AniAvatar: Animated visualizations of affective states. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4828–4837.

Sonderegger, A., Uebelbacher, A., & Sauer, J. (2019). The UX Construct–Does the Usage Context Influence the Outcome of User Experience Evaluations? *IFIP Conference on Human-Computer Interaction*, 140–157.

Thielsch, M. T., Lenzner, T., & Melles, T. (2012). Wie gestalte ich gute Items und Interviewfragen. *Praxis Der Wirtschaftspsychologie II: Themen Und Fallbeispiele Für Studium Und Praxis*, 221–240.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127–145.

Tuch, A. N., Roth, S. P., Hornbæk, K., Opwis, K., & Bargas-Avila, J. A. (2012). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior*, *28*(5), 1596–1607.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247–252. https://doi.org/10.1037/0021-9010.82.2.247

Wiedenbeck, S. (1999). The use of icons and labels in an end user application program: An empirical study of learning and retention. *Behaviour & Information Technology*, *18*(2), 68–82.

Wilde, M., Bätz, K., Kovaleva, A., & Urhahne, D. (2009). Überprüfung einer Kurzskala intrinsischer Motivation (KIM). *Zeitschrift Für Didaktik Der Naturwissenschaften*, *15*.

Wright, P. C., McCarthy, J. M., & Meekison, L. (2003). A framework for analysing user experience. *Funology: From Usability to User Enjoyment. Kluwer, Dordrecht*.