Int. J. Human–Computer Studies 179 (2023) 103116

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Questionnaire experience of the pictorial usability inventory (PUI) – a comparison of pictorial and hybrid usability scales

Juergen Baumgartner [a,b,*], Andreas Sonderegger [a,c], Juergen Sauer [a]

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

ARTICLE INFO

Keywords:
Perceived usability
Pictorial scale
Hybrid scale
Consumer product evaluation
Questionnaire experience

ABSTRACT

In recent years, alternative types of usability questionnaires using graphical elements (pictorial scales) or a combination of graphical and verbal elements (hybrid scales) have been introduced. Previous research indicates that these questionnaires have advantages, such as increased respondent motivation, and drawbacks, such as extended questionnaire completion time. This study aimed to systematically investigate the psychometric properties and the respondents' experience of two versions of a recently developed questionnaire, the Pictorial Usability Inventory (PUI), consisting of a hybrid and pictorial version. Given that questionnaire length is a crucial factor for the usefulness of a scale, the study tested long and short versions (8 items vs 3 items) of both questionnaire types. The study involved an online usability test with 777 participants, who were asked to complete one of the four PUI versions and an established verbal usability scale after solving three tasks on a webpage. The results demonstrated high sensitivity, high convergent validity, and good internal consistency for all four PUI versions. While the long pictorial scale achieved the best psychometric properties overall, participants preferred the hybrid scales, particularly the short version. The study's findings are in line with previous research on pictorial and hybrid instruments and suggest that hybrid instruments, particularly short ones, may be superior to purely pictorial instruments in terms of respondent-centred aspects conceptualised in the term 'questionnaire experience'.

1. Introduction

1.1. Usability assessment

In the wake of the rapidly advancing technological development in work and leisure-related domains, usability assessment is gaining in importance across different industries. This is because, more than ever, it is crucial for the development of new technology to meet user needs by testing interactive products and services with representative users and to improve product design already in the early stages of development (ISO 9241–210; International Organization for Standardization, 2019).

The core usability principles are still the same today as in the 1990s. 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' (ISO 9241–210; International Organization for Standardization, 2019, p. 3). The definition of the usability concept is mainly focused on aspects of functionality and performance (effectiveness, efficiency) but covers with satisfaction also a subjective component. In contrast, the more recently coined concept of user experience (UX) adopts a broader focus on the entire spectrum of human experience (i.e. emotions and affect, aesthetic experience in addition to experiences of satisfaction and performance) when interacting with a technological artefact (International Organization for Standardization, 2019; Sauer et al., 2021). Although the UX concept is receiving more and more attention in practice and research, it is still essential to assess the usability component of a user interacting with a technological artefact (Sauer et al., 2021).

The field of usability evaluation offers a rich toolkit of methods and best practices. A cornerstone in usability assessment is the usability test, a method in which representative test users are observed while interacting with an artefact (Nielsen, 1994). However, a usability evaluation is often conducted using a combination of methods (Barnum, 2011). Typically, usability tests involve a quantitative subjective evaluation of

* Corresponding author at: Department of Psychology, University of Fribourg, Rue P.-A.-de-Faucigny 2, Fribourg CH-1700, Switzerland.
E-mail address: juergen.baumgartner@unifr.ch (J. Baumgartner).

https://doi.org/10.1016/j.ijhcs.2023.103116
Received 15 December 2022; Received in revised form 6 July 2023; Accepted 19 July 2023
Available online 20 July 2023

the artefact's usability by means of a questionnaire. Since the late 1980s, various verbal usability questionnaires have been published (Assila et al., 2016). Amongst them, the System Usability Scale (Brooke, 1996) is one of the most established and most often cited questionnaires in the usability domain (Lewis, 2018). Several reasons might have contributed to the popularity of the SUS, such as the availability of validated versions in various target languages being Arabic, Chinese, French, German, Hindi, Italian, Persian, Polish, Portuguese, Slovene, and Spanish (Gao et al., 2020; Lewis, 2018), the development of norms (Bangor et al., 2008, 2009; Lewis and Sauro, 2017; Sauro and Lewis, 2016), but also broad empirical evidence of a large number of validation studies (for a detailed overview see Lewis, 2018) and independent analyses of its factor structure and its relationship with other usability instruments (e.g. Borsci et al., 2009, 2015).

Verbal scales are a common tool used in usability evaluations. However, their usage can present challenges and drawbacks under certain conditions. While not all verbal questionnaires are long or require significant effort to answer, some can be strenuous, especially when presented in a battery of multiple questionnaires. Furthermore, answering similar questions repeatedly (Robins et al., 2001) or potential comprehension issues due to long or complex questions might lead to reduced motivation and response fatigue (Baumgartner et al., 2021). As a result, respondents may engage in undesirable answering behaviour, such as giving random answers, skipping questions, or even prematurely terminating the questionnaire (Herzog and Bachman, 1981; Robins et al., 2001). Such answering behaviour, in turn, may decrease the quality of the collected data (Herzog and Bachman, 1981).

1.2. The role of questionnaire experience (QX)

Recently, attempts have been made to extend the scope of traditional questionnaire characteristics (i.e. psychometric properties) by respondent-centred aspects, such as perceived questionnaire experience (QX; Baumgartner et al., 2021; Sauer et al., 2021). The term QX was mentioned first by Toepoel et al. (2019), referring to an overall experience measure for the response format representations in surveys (such as smileys and stars). The first definition of the term QX was put forward by Sauer et al. (2020), defining it as the entire experiential process a respondent goes through when completing a questionnaire or a test, subsuming several facets under its umbrella (e.g. respondent workload, respondent motivation, item comprehension). The goal of introducing the concept of QX is to provide a complementary perspective to the evaluation of questionnaires and to propose a framework of relevant measures that harbour valuable information for obtaining a more complete picture of an instrument. We believe that this approach of synthesising information from psychometric analysis and respondent-centred aspects is useful for evaluating existing and new questionnaires and is particularly valuable for evaluating newly developed pictorial or hybrid instruments (i.e. pictorial and verbal content). In this context, the concept of QX has gained some interest. It addresses the experiential consequences (e.g. feelings, emotions, attitudes, and beliefs) of a questionnaire respondent. Previous research has suggested that pictorial scales might be beneficial compared to verbal scales with regard to QX but also come with some potential disadvantages (e.g. increased item completion time; Baumgartner et al., 2020).

1.3. Pictorial scales in usability assessment

In contrast to verbal instruments, only a few pictorial instruments have been developed so far in the domain of human-machine interaction, which were mainly limited to the evaluation of product emotion (e.g. PREMO - Product Emotion Measurement Tool; Desmet, 2003; or the AniSAM – Animated Self-Assessment Manikin; Sonderegger et al., 2016). In recent years, efforts have been made to extend the toolbox of usability questionnaires by offering pictorial alternatives. Pictorial scales are promising for several reasons. Such scales offer practitioners and researchers a broader range of options when selecting a suitable instrument, including questionnaires that are not necessarily bound to language. Because pictorial scales are visual in nature, interpreting items is not limited to fully literate persons but is accessible to people with poor reading skills or non-native speakers (Ghiassi et al., 2011; Sauer et al., 2021). Furthermore, previous studies showed increased motivation in questionnaire completion using pictorial scales (Baumgartner et al., 2020, 2021; Baumgartner et al., 2019b). They gain users' attention and interest and prevent the effects of respondent fatigue or undesired response patterns (Haddad et al., 2012). Besides, purely pictorial questionnaires are not language-dependant (Betella and Verschure, 2016). Thus they do not need to be translated into different languages. Even if one raises questions concerning cultural differences in interpreting visualisations, pictorial scales can potentially be used across language borders. On the other hand, the development takes time and multiple iterations to create and validate a questionnaire are necessary (for a first draft of guidelines, see Sauer et al., 2021). Furthermore, comprehensibility issues and ambiguity increase with the complexity and the abstractness of the concept in question (see also Collaud et al., 2022). Therefore, the biggest challenge is to find concrete representations and visual metaphors that are easy to understand.

Currently, there are only a few pictorial usability scales available. One such scale is the PSIUS (Pictorial Single Item Usability Scale; Baumgartner et al., 2019a), which uses graphical elements like an avatar with different emotional expressions (satisfied vs frustrated) and hand gestures (thumbs up vs thumbs down) to measure usability. Two lab studies have shown that PSIUS has high convergent validity with the System Usability Scale ($r=0.881$, $r=0.696$; Baumgartner et al., 2019a). Another pictorial instrument is the P-SUS (Pictorial System Usability Scale; Baumgartner et al., 2019b), a multi-item scale based on the SUS. The P-SUS was developed using a user-centred approach, which involved conducting think-aloud protocols and comprehension checks to ensure that each item was accurately visualised (cf. ISO 9186–1; International Organization for Standardization, 2014). An online study showed significantly increased motivation compared to the SUS, measured with a short version of the Intrinsic Motivation Inventory (IMI; Wilde et al., 2009). Furthermore, high correlations with the SUS were obtained ($r=0.886$; Baumgartner et al., 2019b). However, data analysis on the item level showed that some P-SUS items had intermediate correlations with the corresponding SUS item ($r<0.500$) and extended answering times (3–4s longer per item), assuming comprehensibility issues due to ambiguous visualisations. A hybrid version of the P-SUS (i.e. H-SUS) was created to address these issues, combining pictorial and verbal content in one scale. In an online study (Baumgartner et al., 2021), H-SUS showed high correlations with SUS ($r=0.862$), and all items had strong correlations with the corresponding SUS items ($r>0.500$). Interestingly, 62.5% of participants preferred the hybrid version over the verbal one. Although there is room for improvement in pictorial scales through further design iterations, the development of P-SUS and H-SUS showed that converting an existing questionnaire to a pictorial one has limitations. Especially verbal items with abstract concepts narrow the possibilities of a concrete visualisation and increase ambiguity and misinterpretation. To work around this problem, a different approach was chosen to develop the first version of PUI (Pictorial Usability Inventory; Baumgartner et al., 2020). Instead of 'translating' one verbal source questionnaire into a pictorial version, suitable items of various verbal questionnaires were selected based on item quality (i.e. high correlation with the concept of usability) and feasibility for visualisation. A set of twelve pictorial items was tested in an online study (see Baumgartner et al., 2020, for a detailed description of the selection and design procedure). Increased motivation and high correlations with the SUS were observed ($r=0.852$). However, the completion time still took longer (about 3s longer per item), and 60% of participants preferred the verbal questionnaire over the pictorial one. Overall, previous attempts showed promising results in the form of increased motivation and high convergent validity. These advantages

are accompanied by drawbacks such as longer completion times and inconsistent preference findings. To tackle these drawbacks, this article deals with whether it was possible to shorten the PUI while maintaining high psychometric quality and whether a hybrid version would improve the psychometric and experiential qualities of the tool.

1.4. Development of pictorial and hybrid scales

The Pictorial Usability Inventory (PUI) is a usability questionnaire that uses image-based elements to convey the meaning of its items. The items consist of two pictures depicting the extreme poles of a specific usage situation where a person interacts with a device. Similar to a bipolar scale, the left picture shows the negative usage situation, and the right picture the positive one. Below the pictures, each item has a seven-point Likert scale anchored with numbers from left to right, ranging from −3 to 3. The pictures comprise an avatar (female or male) expressing some specific affective state, a device (desktop, tablet, or smartphone), and additional graphical representations of concrete or abstract concepts. The pictorial items were drawn with a vector graphics editor. Fig. 1 shows a PUI item referring to the concept of interface complexity.

Several design considerations were implemented to create the pictorial representations. Concrete visual elements or visual metaphors were used to make abstract concepts more tangible (e.g. target flag for goal, stopwatch for time spent, check marks to indicate completion/success and x marks to indicate error/failure). Furthermore, key elements were coloured in red and green to allow fast recognition between the negative and the positive usage situation (avatars' clothing, check marks and x marks, device frames for highlighting content).

The first version of the PUI consisted of 12 items and was tested in a pilot study (Baumgartner et al., 2020). While the results suggested good psychometric properties and high motivation in completing the questionnaire, 60% of participants preferred a verbal usability questionnaire over the pictorial one, and completion times were longer for the pictorial scale. Due to these results, we shortened the instrument by excluding redundant and less intuitive items. To identify these items, think-aloud protocols (TAP) were conducted with 14 participants (50% female; M=26.07 yrs, SD=10.01; occupation: 50% students, 50% employees). They were presented with all 12 items sequentially (half of the participants in regular order, half in reversed order) and were asked to verbalise the meaning of each item. After revealing the intended meaning, participants had to rate the comprehension of each item on a seven-point Likert scale ranging from 1 (not at all comprehensible) to 7 (very comprehensible). A facilitator took notes of the interpretations and the rating. The subsequent selection process was based on item comprehension (i.e. items had to have a rating of 5 or higher) and redundancy (i.e. in case of similar content, the one with the highest comprehension rating was retained). Six items from the original PUI were selected using this procedure. Since four out of six items were related to efficiency, we added one item each for effectiveness and satisfaction. The two items also originated from the original PUI but were modified based on ideas from the think-aloud sessions and the authors.

To pretest the final 8-item set of the PUI, eighteen participants (72.2% female; M=29.06 yrs, SD=12.43) recruited from a research seminar at the University of Fribourg were presented all eight items sequentially and were asked to indicate the meaning of the item. Two independent raters afterwards categorised the answers regarding their match with the intended meaning. As Table 1 indicates, comprehension rates are high for most of the items. Only PUI item 2 obtained a value lower than the minimal comprehension rate cut-off value of 67% required in ISO 3864 (see Hicks et al., 2003). Since these pictorial items are not used in a safety-critical environment, we considered these results satisfactory.

Based on this 8-item PUI, four versions were created for this study, varying on two characteristics: content type (pictorial vs hybrid) and the number of items (long vs short version). This resulted in the following versions: PUI-L (pictorial long version), PUI-S (pictorial short version), HUI-L (hybrid long version), and HUI-S (hybrid short version). The long version consists of eight items, whereas the short version comprises three items, referring to the three core components of usability (efficiency, effectiveness, and satisfaction). The authors chose the items for the short version based on what might represent each core component best. Content type distinguishes between pictorial and hybrid scales. The former only consists of non-verbal graphical elements, whereas the latter combines verbal and pictorial content (see Fig. 2).

In contrast to previous hybrid scales (e.g. H-SUS; Baumgartner et al., 2021), the verbal content was phrased as a question to match better the degree of agreement with the numerical answer options. Since the original wording did not always fit well in combination with the pictorial representation, two usability experts were asked to make suggestions for the wording of each item in order to obtain a suitable question for the HUI. In addition, an example item is shown to all participants to familiarise them with the questionnaire. The example consists of a short instruction on how to complete the questionnaire and what it means when '−3' is selected. Fig. 2 shows the example item and the complete set of items for the different versions.

1.5. The present study

This article aims to compare different versions of the Pictorial Usability Inventory (PUI) that were developed by crossing content type (pictorial vs hybrid) and questionnaire length (long vs short) in a 2×2 design. The goal of this study was hence to assess the strengths and weaknesses of four questionnaire versions, PUI-L (pictorial long version), PUI-S (pictorial short version), HUI-L (hybrid long version), and HUI-S (hybrid short version). The comparison is made by analysing psychometric properties and QX (i.e. respondent-centred measures). The System Usability Scale (SUS; Brooke, 1996) served as the main

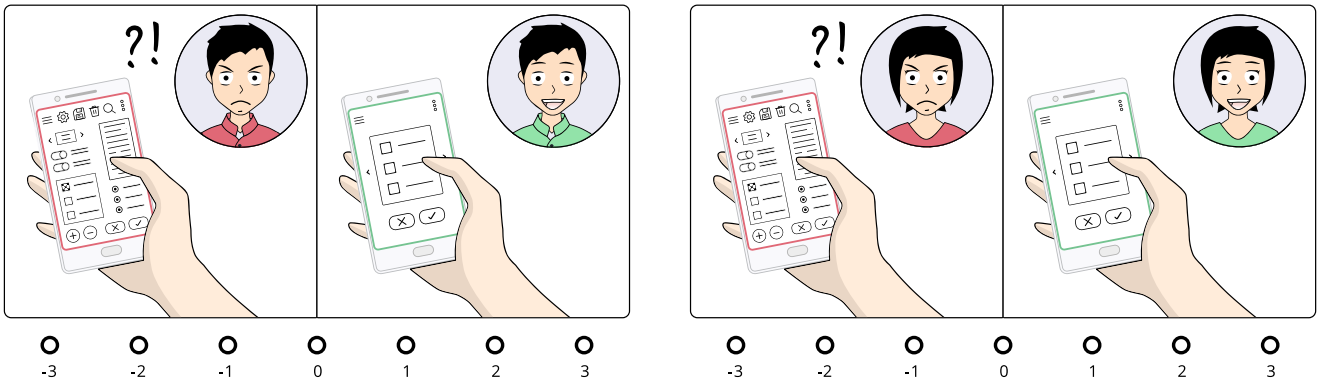

Fig. 1. Example of PUI item with male and female avatar referring to interface complexity of a smartphone.

Table 1
Comprehension rates in per cent for all 8 PUI items (N=18).

	PUI items							
	01	02	03	04	05	06	07	08
Comprehension rate (%)	100.00	61.11	94.44	72.22	88.89	100.00	94.44	72.22

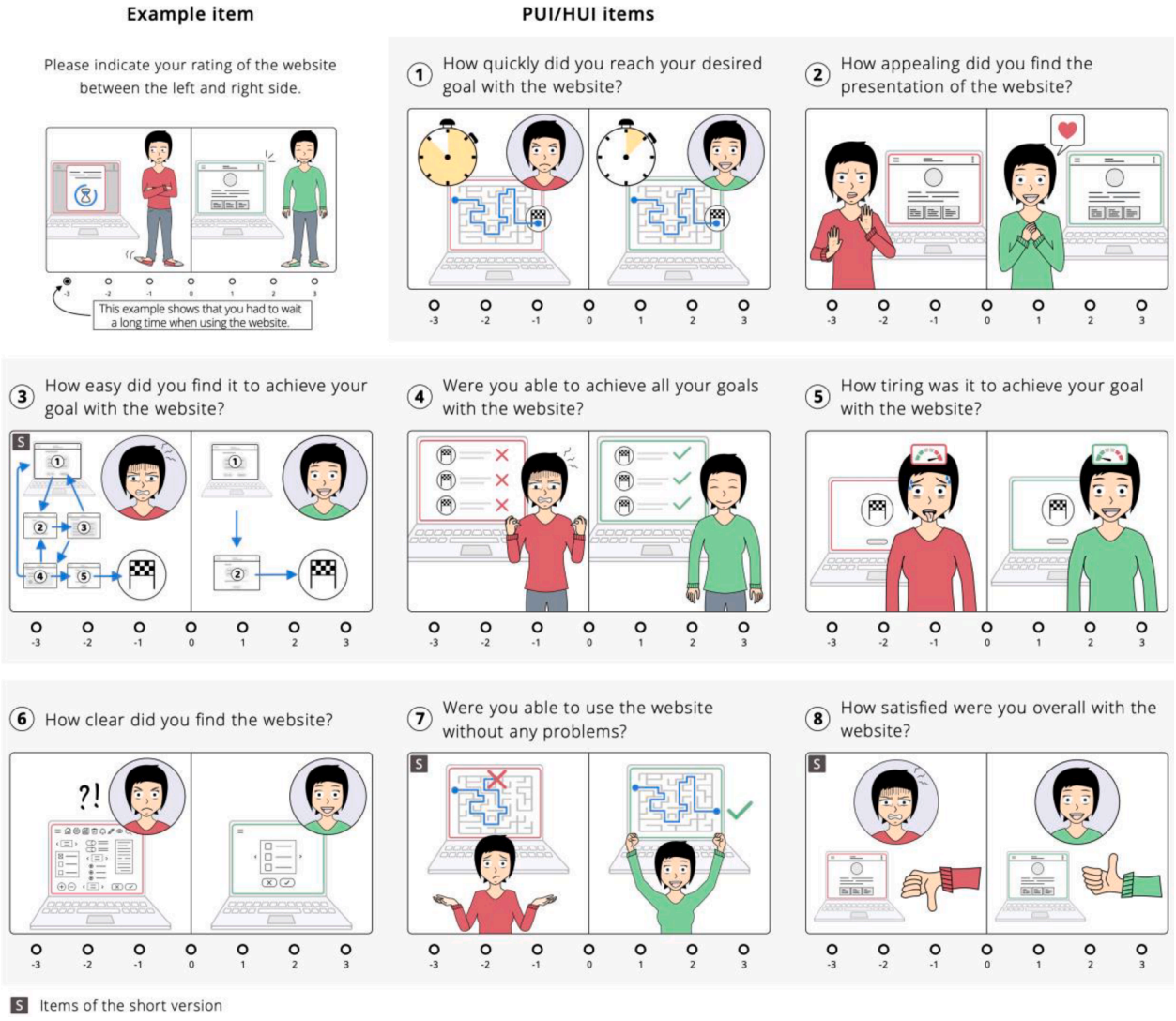

Fig. 2. Example item and complete set of items of the Pictorial Usability Inventory (PUI) in a female version. The verbal question was only shown for the hybrid version (HUI). The wording was translated from German to English.

instrument to assess convergent validity. An online study was conducted using a manipulated website prototype (low vs high usability). Participants solved three tasks on the website and subsequently completed several verbal questionnaires and one of the four PUI versions. Assuming a successful usability manipulation and considering the findings of previous studies, we generally predicted that the four PUI versions would be very similar in psychometric quality (i.e. high sensitivity, good convergent validity and good internal consistency). We expected the results to be comparable to those of an established usability questionnaire like the SUS. Moreover, we predicted that differences between PUI versions and verbal questionnaires would emerge rather on a subjective level (i.e. in respondent-centred aspects). For this reason, specific hypotheses were formulated regarding the effects of the manipulation of length (long vs short) and content type (pictorial only vs hybrid) on respondent-centred measures (between-subjects comparisons). Further hypotheses were made for respondent-centred measures of the four questionnaire versions in comparison with established verbal usability questionnaires (within-subjects comparisons).

1.5.1. *Hypotheses for manipulated factors (length and content type)*

We believe that questionnaire length influences several measurable aspects of the subjective experience when completing a questionnaire. Table 2 shows the respondent-centred aspects we assessed in this study and where we expected effects. In our first hypothesis (H1), we assumed that the length of the questionnaire influences motivation. There is evidence from research that longer questionnaires are associated with lower response rates (e.g. Galesic and Bosnjak, 2009; Heberlein and Baumgartner, 1978). Even if we do not assess response rates, we think that this effect can be transferred to our research question in the sense that the more items a questionnaire has, the lower the motivation is to

Table 2
Expected effects of the manipulation of independent variables questionnaire length and content type on respondent-centred measures of questionnaire experience.

Respondent-centred measures	H1: Questionnaire length	H2: Content type
Motivation	long < short	no effect
Comprehension	no effect	hybrid > pictorial
Workload	long > short	no effect
Satisfaction	long < short	no effect
Aesthetics	no effect	no effect

complete it. Consequently, we expect a long questionnaire to increase perceived workload and decrease satisfaction compared to a short one. We did not assume that the length of the questionnaire would influence item comprehension or whether a questionnaire is perceived as aesthetically pleasing.

In our second hypothesis (H2) concerning content type (pictorial vs hybrid), we expected that comprehension would be facilitated for hybrid questionnaires since they offer a pictorial and a verbal representation. There is evidence from research that the recognition of intended meaning is easier when using a hybrid scale (e.g. Ghiassi et al., 2011). For the other aspects, we did not expect any effects to occur.

1.5.2. Hypotheses for comparisons with verbal questionnaires

The next set of hypotheses (H3) is related to the comparison of respondent-centred aspects between the four PUI-versions and the verbal usability questionnaires. Table 3 shows an overview of the effect patterns we expected. Only hypotheses relative to the questionnaire type were formulated.

It is often argued that pictorial scales increase motivation and provide more pleasure than verbal scales (Desmet et al., 2001; Ghiassi et al., 2011; Haddad et al., 2012). This notion is backed by previous studies that indicated significant differences in motivation in favour of pictorial and hybrid scales (e.g. Baumgartner et al., 2020, 2021). Consequently, we expected all PUI versions to be rated significantly better for motivation than the verbal questionnaires. Concerning comprehension, we assume that the purely pictorial scales achieve similar comprehension ratings as the verbal ones since only the most comprehensible pictorial items were selected for this study (cf. selection process in the previous section). We expect the hybrid scales to be more comprehensible than the verbal questionnaire since they have the advantage of an additional pictorial component (in the sense of a redundancy gain, e.g. Backs and Walrath, 1995). Furthermore, we consider questionnaire workload as an antagonist to questionnaire motivation, representing aspects that prevent a positive experience from happening during questionnaire completion. It has been suggested in the literature that pictorial scales are less mentally demanding than verbal scales (e.g. Wissmath et al., 2010). We assume that the pictorial representations have a facilitating effect on questionnaire completion, providing more direct access to the intended meaning. Therefore, we expect all PUI versions to be rated lower for questionnaire workload than the verbal questionnaires.

Table 3
Hypotheses of expected differences between PUI-version and verbal usability questionnaires regarding respondent-centred measures of questionnaire experience.

	Pictorial scales		Hybrid scales	
Respondent-centred measures	Long	Short	Long	Short
Motivation	↑	↑	↑	↑
Comprehension	=	=	↑	↑
Workload	↓	↓	↓	↓
Satisfaction	↑	↑	↑	↑
Aesthetics	↑	↑	↑	↑
Preference	=	=	↑	↑
Item completion time	↑	↑	↑	↑

Concerning satisfaction and aesthetics, we believe all PUI versions to be rated significantly better than the verbal questionnaires due to the pictorial elements that are pleasant to see and the before mentioned advantages that may have a positive impact on perceived satisfaction.

In addition to the respondent-centred measures, we assessed questionnaire preference (verbal vs with pictures) and questionnaire completion time. We assumed that a majority of participants prefer hybrid scales. Our assumption is based on a previous study that pointed towards that direction (cf. Baumgartner et al., 2021). Regarding pictorial scales, we assumed lower preference ratings based on the pilot study (Baumgartner et al., 2020), in which most respondents preferred the verbal scales.

The last measure addressed in this study is completion time. Since the PUI versions and the verbal questionnaires differ substantially in questionnaire length, only hypotheses for the average item completion were put forward. Previous studies showed predominantly lower completion times for verbal questionnaires than pictorial and hybrid questionnaires. Therefore, we hypothesised for this study (using adult native speakers without impairments as participants) that verbal items are completed fastest, followed by pictorial and hybrid items.

2. Method

2.1. Participants

Participants were recruited by (1) sending an email to all bachelor and master students at the University of Fribourg, (2) advertising the study on the website of the Psychology Department of Fribourg, and (3) by sharing the study within the social networks of the experimenters. Ten vouchers worth 30 CHF each were raffled to increase participant motivation. The study was conducted in German and French language. In total, 777 participants (79.4% female, 19.2% male, 1.4% diverse) took part in the online study, with their ages ranging from 18 to 62 years (M=23.43 yrs, SD=4.82). There were 478 participants (61.5%) who completed the study in French and 299 (38.5%) in German language. The sample consisted of 714 students (91.9%), 53 employees (6.8%), and 10 participants (1.3%) who did not report their professional status. Six participants reported having some form of colour blindness. Participants rated their experience with websites between medium and high (M=4.73, SD=1.74) on a seven-point Likert scale ranging from 1 (very low) to 7 (very high). 554 participants (71.3%) completed the study on a laptop/desktop, 196 participants (25.2%) on a smartphone, and 27 participants (3.5%) on a tablet.

2.2. Website prototype, user tasks and pilot study

In order to evaluate the different questionnaire versions in a controlled and standardised environment, participants interacted with a website prototype of a fictitious leisure centre, which was created in German and French language for this study. The content of the website was adapted from a website that has been previously developed for research purposes (Schmutz et al., 2019). Furthermore, the website was adapted so that users could interact with the website using different device types (i.e. desktop, tablet, or smartphone). The website's usability was manipulated on two levels (low vs high). The low-usability version was created by violating usability heuristics (e.g. Nielsen and Molich, 1990) and best practices of interface design, resulting in (1) inappropriate interface patterns, (2) more complex information architecture, (3) deliberate delays when loading pages, (4) deliberate bugs in layout, (5) inadequate form design, and (6) placing information relevant to task completion in unexpected places on the webpage.

Participants were asked to solve three tasks on the website of the leisure centre: (1) finding out whether a specific sauna is open during winter, (2) buying an annual subscription for the centre, and (3) making a reservation for a bowling evening with friends. The study was set up so that participants could reread the task description at any time. If

participants could not solve a task within four minutes, they were instructed to move on to the next one.

A pilot study was carried out to test whether the usability manipulation of the website prototype was successful, employing a between-subjects design. Twenty-eight German-speaking participants (60.7% female; M=32.82 yrs, SD=14.99; Occupation: 9 students, 17 employees, 2 other) were asked to solve three tasks on the website using their personal devices. The assignment to the usability condition was counterbalanced (either low or high usability). Subsequently, participants reported how many tasks they could solve (none, one, two, or three) and completed the SUS. Table 4 shows the task completion rate, indicating that participants in the low-usability condition solved fewer tasks and spent more time on task completion than participants in the high-usability condition. The analysis of the SUS score (using a Mann-Whitney test) showed a significant difference between low and high-usability conditions (Mdn_{low}=43.75, Mdn_{high}=83.75, U=19.50, z=3.61, p=.000, r=0.682), suggesting a successful manipulation of usability.

2.3. Measures and instruments

Various measures and instruments were used to determine psychometric properties and subjective QX of the four versions of the Pictorial Usability Inventory. They are categorised into (1) measures of sensitivity, (2) measures of convergent validity, (3) objective measures of usability, (4) internal consistency and (5) respondent-centred measures. Whenever possible, validated instruments in German and French language were used. If none were available, they were translated with the help of a professional translator or a bilingual expert in the usability domain.

2.3.1. Sensitivity

Sensitivity is defined as the capability of an instrument to detect appropriate differences between different systems or between usability manipulations (Lewis, 2002; Sauro and Lewis, 2016), hence representing a vital quality of a usability questionnaire. Sensitivity was determined by comparing usability scores of low and high-usability conditions. Large effect sizes for comparing low and high usability webpage are good indicators of the scale's sensitivity.

2.3.2. Measures of convergent validity

SUS. The System Usability Scale (SUS; Brooke, 1996) was used as a primary measure for convergent validity. The SUS consists of 10 items to be rated on a five-point Likert scale ranging from 1 (totally disagree) to 5 (totally agree). After some mathematical transformation, an overall usability score ranging from 0 to 100 is obtained, which is often interpreted using the curved grading scale (grades ranging from 'A' to 'F'; Sauro and Lewis, 2016). The SUS is widely used in research and practice, considered a valid and reliable instrument for assessing perceived usability (e.g. Cronbach's α>0.910; Bangor et al., 2009). This study used the validated French and German versions by Gao et al. (2020).

UMUX LITE. The short version of the Usability Metric for User Experience (UMUX-LITE; Lewis et al., 2013) was used as an additional measure. The instrument consists of two items rated on a seven-point Likert scale ranging from 1 (totally disagree) to 7 (totally agree). The authors reported good reliability (α>0.820) and high concurrent validity with the SUS (r=0.810).

Single-item scales. Three self-created single-item scales were used to target the core components of usability: effectiveness ('I was able to successfully achieve my goals using the website.'), efficiency ('On the website, I found what I wanted very quickly.'), and satisfaction ('Overall, I was satisfied with this website.'). The items were rated on a seven-point Likert scale ranging from 1 (totally disagree) to 7 (totally agree).

NPS. The Net Promoter Score (NPS; Reichheld, 2003) was applied to assess the likelihood to recommend (LTR). It consists of a single item rated on an eleven-point Likert scale ranging from 0 (very unlikely) to 10 (totally likely). Previous studies reported strong correlations comparing LTR and SUS (r=0.623; Sauro and Lewis, 2016) and LTR and UMUX-LITE (r=0.730; Lewis et al., 2013).

2.3.3. Objective measures of usability

To evaluate objective usability measures, we recorded the performance of the interaction with the website using a browser script. The main performance indicators included the aggregated task completion time and the number of user interactions across all tasks. To obtain a measure of efficiency for participants with successful task completion, we calculated the optimal path deviation (OPD) by subtracting the minimal number of user interactions from the observed number of interactions. Finally, task completion rate was used as a measure of effectiveness.

2.3.4. Internal consistency

Internal consistency is a measure of reliability that describes the relationship between items and implies that related items are answered similarly (Coolican, 2017). Cronbach's alpha was calculated for all PUI versions and the SUS. High values of internal consistency are to be expected from highly reliable instruments (α>0.900; Nunnally and Bernstein, 1994).

2.3.5. Respondent-centred measures

Several respondent-centred aspects of completing a questionnaire were assessed using the Questionnaire Experience Questionnaire (QXQ). QXQ is a self-developed instrument consisting of three multi-item and two single-item scales that assess measurable indicators relevant to questionnaire experience. The multi-item scales for questionnaire motivation, comprehension and workload comprise three items each that use verbal statements to rate the experience of completing a questionnaire (e.g. 'the questionnaire was easy to fill in'). The single-item scales were added to assess the questionnaire's aesthetics and overall satisfaction. A seven-point Likert scale ranging from 1 (totally disagree) to 7 (totally agree) was used to measure the level of agreement with these statements. The aspect of 'questionnaire motivation' is based on a subscale of the Intrinsic Motivation Inventory (IMI; Ryan, 1982) already used in previous studies (e.g. Baumgartner et al., 2019b). Wilde et al. (2009) reported good reliability for the subscale (α=0.850 – 0.890). The other scales were developed for the purpose of this study. Data from the present study indicate acceptable to excellent reliability for the multi-item scales (α=0.738 – 0.903). Except for the questionnaire workload, all items were positively worded. QXQ was applied twice, once after completing the pictorial or hybrid questionnaire and once after the verbal usability questionnaire. Table 5 shows the specific wording of the QXQ items and the Cronbach alpha values for the multi-item scales.

Besides QXQ, questionnaire preference was assessed at the end of the study by asking participants which questionnaire they liked more (pictorial or verbal). Participants were asked with a bipolar seven-point Likert scale ranging from 1 (verbal questionnaire) to 7 (pictorial questionnaire). Previous studies have adopted a similar approach to measure

Table 4
Task completion rate, task completion time and SUS score as a function of usability level.

	Task completion rate	Task completion time (sec) M(SD)	SUS M(SD)
Low usability (N=13)	74.36%	813.38 (568.93)	48.57 (19.78)
High usability (N=14)	97.62%	367.57 (192.03)	81.96 (14.78)

Note: One participant was excluded from data analysis for taking long breaks during task completion.

Table 5
Items of the Questionnaire Experience Questionnaire (QXQ) and Cronbach alpha values for multi-item scales. The wording was translated from German to English.

Measurable indicator	Item	Cronbach's alpha
Questionnaire motivation	The questionnaire was fun. The questionnaire was entertaining. The questionnaire was interesting.	.903
Questionnaire comprehension	The questionnaire was comprehensible. The questions were clear. The questionnaire was easy to fill in.	.871
Questionnaire workload	The questionnaire was too long. The questionnaire was complicated. The questionnaire was tedious to fill in.	.738
Questionnaire satisfaction	Overall, I was satisfied with the questionnaire.	–
Questionnaire aesthetics	The questionnaire had an appealing design.	–

the acceptance of pictorial scales (cf. Baumgartner et al., 2020, 2021).

The final respondent-centred measure used was questionnaire completion time, automatically assessed by the survey platform. Completion time (in seconds) was calculated for the whole questionnaire and separately for each item. Since the items were all presented on one page, the average completion time was calculated by dividing the total amount of time by the number of items.

2.4. Experimental design

A 2×2 between-subjects design was used in this study. The following independent factors were manipulated, each on two levels: Type of pictorial questionnaire (pictorial vs hybrid) and questionnaire length (long vs short). Furthermore, system usability was manipulated (low vs high) to permit computation of sensitivity, and the order of questionnaire administration was counterbalanced to prevent any order effects (i.e. half of the participants completed the pictorial questionnaire first, the other half the verbal usability questionnaire first).

2.5. Procedure

The study was conducted using an online survey tool and a webpage prototype. By clicking on the link of the study invitation, participants were directed to an online survey, on which information about the study was provided (i.e. procedure, estimated time, raffle). After answering the informed consent form and responding to demographic questions, participants selected the gender they identified most with by clicking on a picture of an avatar (female or male). They were asked similarly to select the device they used to do the study (desktop, tablet, or smartphone). Afterwards, participants were randomly directed to the webpage prototype (i.e. either high or low usability condition). Participants had to solve three consecutive tasks. If the task was completed, they were automatically directed to the next task. If they could not solve the task, participants could skip it and go to the next one. After completing the last task, the tab with the webpage prototype was automatically closed, and participants could proceed with the online survey. Participants were asked to complete the NPS, followed by one of the pictorial usability questionnaires (PUI-L, PUI-S, HUI-L, or HUI-S, to which they were assigned randomly) and the verbal usability questionnaires (SUS, UMUX-Lite, and three single-item scales). The sequence of pictorial and verbal usability questionnaires was counterbalanced. QXQ was administered to assess the experience with the usability questionnaires. It was administered twice, once after completing the pictorial usability questionnaire and a second time after the verbal usability questionnaires. In the end, participants were asked which usability questionnaire they preferred and if they had completed the study seriously. Finally, they were informed about the raffle and thanked for participating.

2.6. Exclusion criteria and data treatment

The following criteria were used to exclude data sets from the analysis: (1) participants with incomplete data sets, (2) participants with multiple study participation, and (3) participants that responded 'no' to the question of whether they completed the study seriously. Out of 809 participants, 32 participants were excluded from data analysis according to these exclusion criteria. Concerning data treatment, non-parametric tests were used if requirements for normal distribution and homogeneity of variance were not met. The following analyses were carried out: Correlational analyses for convergent and objective measures (Spearman's rank correlation), comparisons of group means to determine the sensitivity and respondent-centred measures (Mann-Whitney U test, Wilcoxon signed-rank test), calculation of internal consistency (Cronbach's alpha), analysis of variance to evaluate the effects of the experimental manipulation (two-factorial analysis of variance), and frequency analyses for questionnaire preference (descriptive percentages). The level of significance was set to 5% for all analyses.

3. Results

3.1. Analysis of scales

3.1.1. Sensitivity

Mann-Whitney U-tests were carried out for all PUI versions and the SUS to assess the difference between low and high usability. As indicated in Table 6, the analysis showed significant differences for all PUI versions (PUI-L, HUI-L, PUI-S, HUI-S) and for the SUS. All usability instruments were highly sensitive to distinguish between low and high-usability conditions, with PUI versions having large effect sizes (all $r \approx .600$) and SUS having medium to large effect sizes (between $r=0.424$ and $r=0.594$).

3.1.2. Convergent validity

Correlations were computed to analyse convergent measures (see Table 7). The analysis showed a strong correlation of $r=0.857$ between PUI-L and SUS. The other versions (HUI-L, PUI-S, HUI-S) correlated slightly lower with SUS in a narrow range of $r=0.773$ and $r=0.784$. A similar trend emerged for correlations with the other convergent measures. PUI-L obtained correlations of $r>0.800$ with UMUX-LITE and the two single items for efficiency and satisfaction. In contrast, the other versions had slightly lower correlations ($r>0.700$). Only the correlations with NPS and the single-item scale for effectiveness were generally lower for all pictorial questionnaires in the range between $r=0.553$ and $r=0.664$, compared to the correlation with the SUS.

3.1.3. Objective measures of usability

The analysis of objective usability measures showed for all PUI versions a negative relationship with the two performance measures (i.e. the number of interactions and completion time, cf. Table 8). Moderate effect sizes for the number of interactions ($r \approx .350$) and completion time ($r \approx .300$) were observed. Overall, effect sizes between PUI versions and performance measures were more pronounced and showed stronger effects than those between SUS and performance measures. Furthermore, the PUI versions showed medium effect sizes with the optimal path deviation ($r \approx .450$). Again, the relationship between SUS and optimal path deviation was generally of lower magnitude. With regard to task completion rate, small to medium-sized effects were observed with pictorial and hybrid versions ($r \approx .200$), whereas nonsignificant to small-sized effects were obtained with the SUS ($r \approx .100$).

3.1.4. Internal consistency

The analysis of internal consistency was conducted for all pictorial

Table 6
Scale sensitivity of PUI versions and SUS as a function of usability levels, including mean scores, grades, and statistical parameters of Mann-Whitney U test.

	Low usability M (SD), grade	High usability M (SD), grade	U	z	p	r
PUI-L (N=191)	63.85 (22.05), C−	89.97 (9.67), A+	1210.00	8.78	.000***	0.635
SUS (N=191)	65.21 (20.83), C	89.30 (9.01), A+	1429.50	8.21	.000***	0.594
PUI-S (N=196)	62.77 (22.52), C−	86.60 (14.74), A+	1709.00	7.82	.000***	0.559
SUS (N=196)	69.71 (19.47), C	85.66 (12.53), A+	2442.00	5.94	.000***	0.424
HUI-L (N=197)	65.65 (22.93), C	90.57 (9.21), A+	1457.00	8.50	.000***	0.605
SUS (N=197)	67.30 (23.04), C	86.88 (10.03), A+	2234.50	6.55	.000***	0.467
HUI-S (N=193)	63.83 (24.08), C−	89.29 (13.90), A+	1348.00	8.59	.000***	0.618
SUS (N=193)	68.75 (20.58), C	86.24 (13.08), A+	2117.00	6.56	.000***	0.472

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

Table 7
Correlations between PUI versions and SUS with convergent measures.

	SUS	UMUX-LITE	NPS	Effectiveness (single item)	Efficiency (single item)	Satisfaction (single item)
PUI-L (N=191)	.857***	.813***	.649***	.614***	.828***	.809***
SUS (N=191)	–	.898***	.723***	.621***	.806***	.855***
PUI-S (N=196)	.784***	.722***	.592***	.553***	.699***	.766***
SUS (N=196)	–	.888***	.686***	.613***	.756***	.856***
HUI-L (N=197)	.773***	.727***	.636***	.573***	.763***	.743***
SUS (N=197)	–	.814***	.655***	.561***	.733***	.755***
HUI-S (N=193)	.774***	.734***	.664***	.629***	.741***	.771***
SUS (N=193)	–	.818***	.671***	.646***	.711***	.798***

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

Table 8
Correlations between PUI versions and SUS with objective measures of usability.

	Number of interactions	Completion time	OPD interactions	Task completion rate
PUI-L (N=179)	−.315***	−.332***	−.536*** (N=114)	.238**
SUS (N=179)	−.302***	−.285***	−.450*** (N=114)	.201**
PUI-S (N=181)	−.376***	−.347***	−.360*** (N=121)	.129*
SUS (N=181)	−.313***	−.281***	−.284** (N=121)	.087
HUI-L (N=190)	−.317***	−.283***	−.471*** (N=116)	.255***
SUS (N=190)	−.142*	−.132*	−.279** (N=116)	.208**
HUI-S (N=181)	−.380***	−.343***	−.486*** (N=103)	.191**
SUS (N=181)	−.301***	−.264***	−.380*** (N=103)	.167*

Notes: Performance data of N=46 participants (5.92% of the overall sample) was not included in the analysis because it was not correctly recorded in the database; OPD=Optimal Path Deviation; OPD was only computed for participants with successful task completion.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

and hybrid versions and the SUS using all items. Results showed excellent Cronbach alpha values for both pictorial long versions ($\alpha_{PUI\text{-}L}$=0.944, $\alpha_{HUI\text{-}L}$=0.932) and good alpha values for the short versions ($\alpha_{PUI\text{-}S}$=0.875, $\alpha_{HUI\text{-}S}$=0.896). Excellent internal consistency was also achieved for the SUS (α=0.912).

3.2. Analysis of manipulated factors

A two-factorial analysis of variance was conducted with respondent-centred measures as dependant variables to assess the effects of questionnaire length and content type. Tables 9 and 10 summarise the data of the analysis.

Results showed that the variable questionnaire length is strongly related to comprehension and workload. The other indicators showed no effect (all $F<1$).

Concerning the variable content type, results showed a strong relationship with the indicators comprehension, workload, satisfaction and aesthetics. No interaction between the two variables of questionnaire length and content type was found (all $p>.05$).

3.3. Comparisons with verbal questionnaires

3.3.1. QXQ

For the analysis of the QXQ, Wilcoxon tests were conducted to detect whether there are significant differences between pictorial and verbal instruments on these dimensions (see Fig. 3).

The results of the dimension questionnaire motivation showed significant differences for the HUI-L, PUI-S and HUI-S. Only the PUI-L achieved no significant difference, although the mean value was in tendency higher than for the verbal questionnaires. With regard to questionnaire comprehension, the hybrid versions were rated similarly high as the verbal questionnaires, showing no significant difference for the HUI-L and the HUI-S. On the other side, comprehension for the nonverbal versions was rated significantly lower, with the lowest scores for the PUI-L, followed by PUI-S. On the workload dimension, the results showed the lowest workload for the HUI-S, with a significant difference from the verbal questionnaires. No significant differences were obtained for HUI-L and PUI-S. The highest workload resulted for PUI-L, rated

Table 9
Indicators of QX as a function of questionnaire length, including statistical parameters of factor analysis.

QX indicator	Questionnaire length	M (SD)	df	F	p	$\eta^2_{partial}$
Questionnaire motivation	Short	5.53 (1.36)	1, 773	0.00	.995	<0.001
	Long	5.54 (1.35)				
Questionnaire comprehension	Short	6.07 (1.14)	1, 773	7.76	.005**	.010
	Long	5.85 (1.32)				
Questionnaire workload	Short	1.79 (1.05)	1, 773	13.53	<0.001***	.017
	Long	2.08 (1.15)				
Questionnaire satisfaction	Short	5.93 (1.32)	1, 773	.71	.400	.001
	Long	5.86 (1.35)				
Questionnaire aesthetics	Short	6.06 (1.18)	1, 773	0.00	.997	<0.001
	Long	6.07 (1.17)				

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

Table 10
Indicators of QX as a function of content type, including statistical parameters of analysis of variance.

QX indicator	Content type	M (SD)	df	F	p	$\eta^2_{partial}$
Questionnaire motivation	Pictorial	5.46 (1.37)	1, 773	2.51	.113	.003
	Hybrid	5.61 (1.33)				
Questionnaire comprehension	Pictorial	5.54 (1.40)	1, 773	101.89	<0.001***	.116
	Hybrid	6.37 (0.87)				
Questionnaire workload	Pictorial	2.09 (1.16)	1, 773	17.02	<0.001***	.022
	Hybrid	1.77 (1.04)				
Questionnaire satisfaction	Pictorial	5.63 (1.42)	1, 773	31.64	<0.001***	.039
	Hybrid	6.16 (1.18)				
Questionnaire aesthetics	Pictorial	5.96 (1.20)	1, 773	5.83	.016*	.007
	Hybrid	6.17 (1.14)				

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

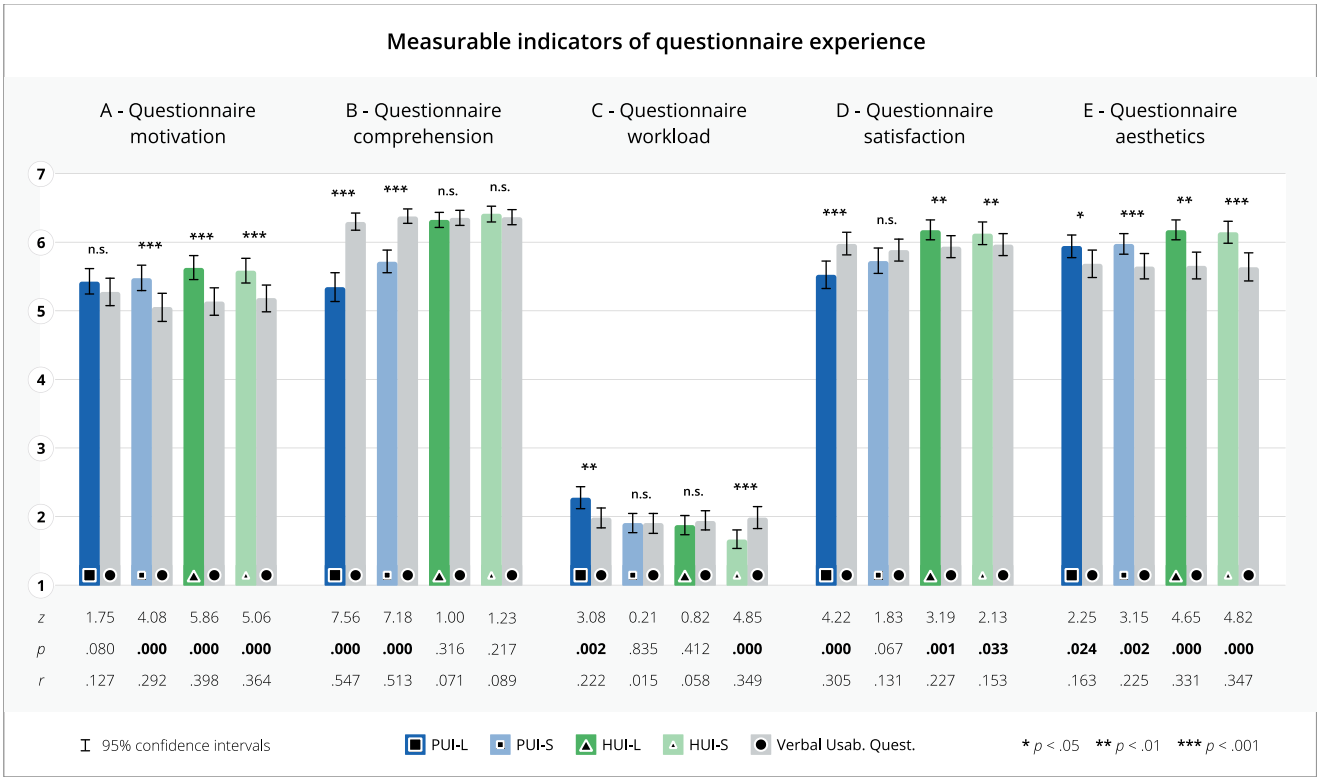

Fig. 3. Overview of QXQ indicators, including statistical parameters of Wilcoxon test between PUI-L, PUI-S, HUI-L, HUI-S and verbal usability questionnaires. Verbal Usability questionnaires comprised SUS, UMUX-LITE and three single-item scales (effectiveness, efficiency, and satisfaction).

significantly higher than the verbal questionnaires. The analysis of questionnaire satisfaction revealed higher scores for the hybrid versions compared to the verbal questionnaires. Significant differences were observed for HUI-L and HUI-S. However, PUI-L was rated significantly lower than the verbal questionnaires. No significant difference to the verbal version was detected for the HUI-S. Finally, regarding questionnaire aesthetics, all pictorial and hybrid versions obtained significantly higher ratings than the verbal questionnaires. The biggest difference was detected for the hybrid versions HUI-L and HUI-S, followed by PUI-S and PUI-L.

3.3.2. Questionnaire preference

The data for questionnaire preference are presented in Fig. 4. Both hybrid versions achieved higher preference ratings than the verbal versions, with HUI-S having the highest preference (63.7%), followed by HUI-L (56.9%). The nonverbal scales PUI-L (30.4%) and PUI-S (41.9%) received preference ratings below 50%.

3.3.3. Questionnaire completion time

The analysis of questionnaire completion time showed that the short versions (HUI-S, PUI-S) were completed the fastest, ranging from 21.33 – 23.20 s, followed by the long versions (HUI-L, PUI-L) ranging from 49.66 – 49.69 s, and at last the verbal scales ranging from 64.84 – 68.82 s (cf. Fig. 5). Since the pictorial and hybrid versions (3 items/8 items) and the verbal usability scales (15 items) vary fairly in the number of items, no further comparisons of group means were conducted.

Concerning item completion time, verbal items were completed the fastest, within 4.33 – 4.59 s. Both long versions (PUI-L and HUI-L) have an average completion time of 6.21 s, followed by HUI-S with 7.11 s and the PUI-S with 7.73 s. Wilcoxon tests were conducted between each pictorial and hybrid version and verbal questionnaires, showing highly significant differences (all $p<.001$).

4. Discussion

This study aimed to compare four versions of the Pictorial Usability Inventory with regard to their psychometric properties and respondent-centred aspects (i.e. questionnaire experience). Considering psychometric measures, the long version of the PUI (PUI-L) showed (with a slight advantage) the best psychometric properties in this study, indicated by the strongest effect sizes for sensitivity, the highest correlation with SUS, similar effect sizes to objective measures of usability, and excellent internal consistency. The other PUI versions are still satisfactory, not lagging much behind in psychometric quality. Concerning respondent-centred measures, the analysis of the two independent variables (i.e. questionnaire length and content type) was in favour of the short version and the hybrid mode in general. In this regard, the hybrid short version (HUI-S) achieved overall the best results, with the highest scores on almost all QXQ dimensions, best preference ratings (roughly two-thirds of participants) and shortest questionnaire completion time (Ø 21s).

Regarding the psychometric properties, the sensitivity analysis indicated a tendency that pictorial and hybrid versions generally have more extreme mean scores than the SUS (i.e. lower means in low-usability and higher means in high-usability condition), which is an indicator of high sensitivity. Consequently, larger effects for all pictorial and hybrid versions were obtained (all $r>0.559$) than for the SUS (all $r>0.424$). Using the curved grading scale (Lewis and Sauro, 2017) – as a helpful approach for interpreting SUS scores using letter grades – grades were the same for all instruments in the high-usability condition (all A+). In the low-usability condition, they were slightly more severe for PUI-L, PUI-S and HUI-S (all C−) than for HUI-L and SUS (both C). While some minor differences may exist, we do not consider them significant enough to suggest a radically different experience. Taken together, the results suggest that all pictorial and hybrid versions can adequately distinguish between low and high-usability conditions. This result is also in line with previous findings of the PUI pilot study (Baumgartner et al., 2020).

With regard to measures of convergent validity, PUI-L showed a very high correlation of $r=0.857$ with the main convergent measure SUS. HUI-L, PUI-S and HUI-S have slightly lower correlations with the SUS in the range of $r=0.773$ and $r=0.784$. Correlations with other convergent measures (UMUX-LITE, NPS, single-item scales for effectiveness, efficiency and satisfaction) tend to be higher for the SUS. However, they are still reasonably high for the PUI versions to describe them as robust. Overall, results on convergent validity imply that all pictorial and hybrid versions measure what they are supposed to measure.

The analysis of performance measures indicated a medium-sized negative relationship for all pictorial and hybrid versions between their usability score and the number of interactions/completion time. They showed medium effect sizes for the optimal path deviation and small to medium effect sizes for the task completion rate. Overall, correlations were stronger for pictorial and hybrid versions than for the SUS. We assume that stronger correlations refer to the fact that some of the PUI items specifically target effectiveness and efficiency and consequently better operationalise aspects related to performance.

Finally, the analysis of internal consistency revealed excellent alpha values for the pictorial long versions (PUI-L and HUI-L, both $\alpha>0.930$) and good alpha values for the short versions (PUI-S and HUI-S, both $\alpha>0.870$). Results for the PUI-L are similar to the findings of the pilot study, where excellent internal consistency was found as well ($\alpha=0.961$, Baumgartner et al., 2020). Furthermore, results are consistent with the

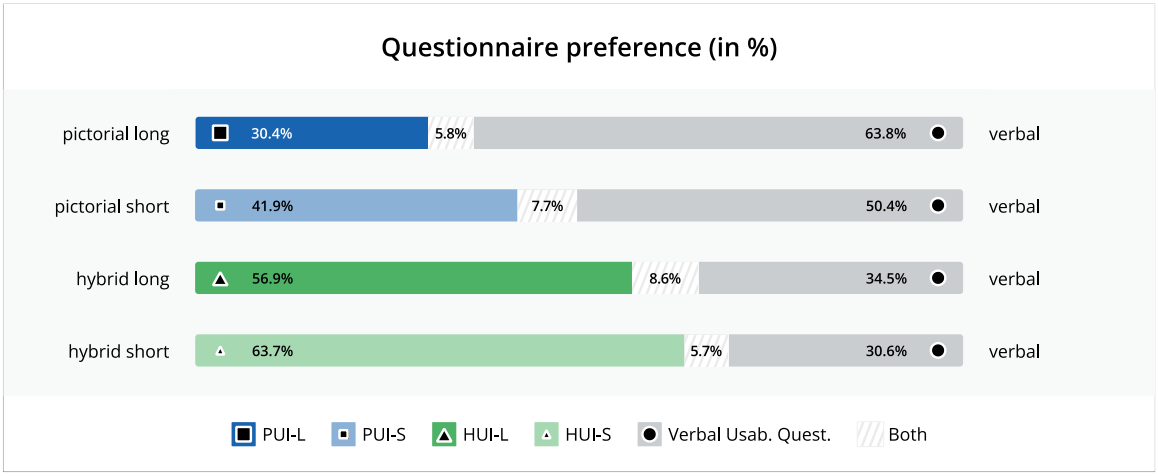

Fig. 4. Overview of questionnaire preference for all PUI versions and the verbal usability questionnaires.

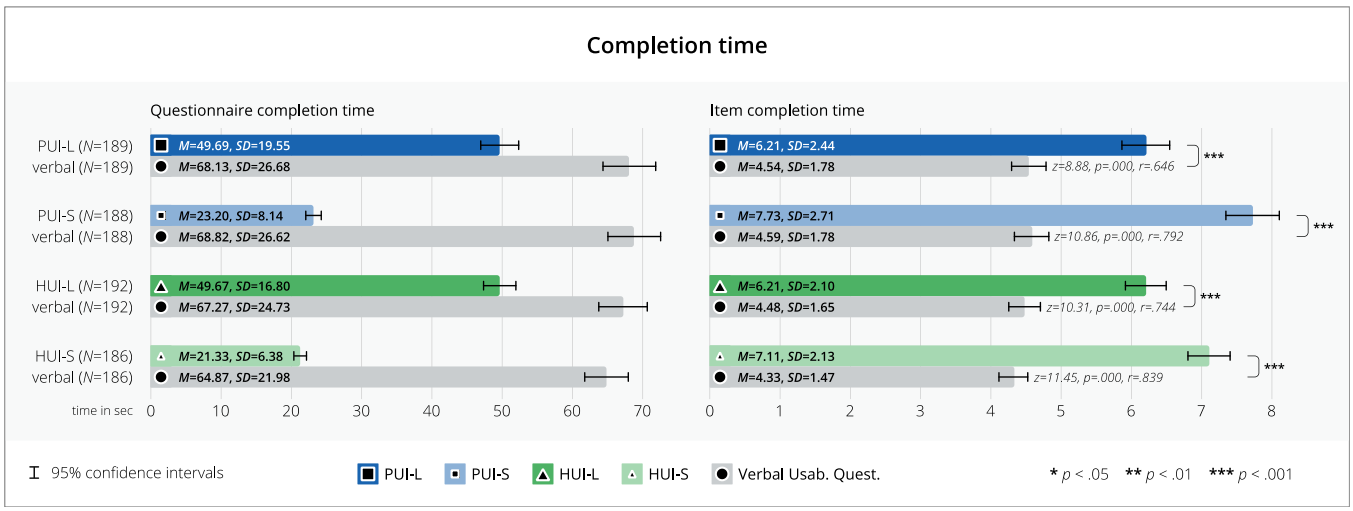

Fig. 5. Overview of the questionnaire and item completion time for all PUI versions and the verbal usability questionnaires.
Notes: Data of N=22 participants (2.83% of the overall sample) were excluded from data analysis since it was identified as outliers (i.e. completion time per item >16s)

idea that alpha values increase with an increasing number of items (e.g. Tavakol and Dennick, 2011). In general, internal consistency is acceptable for all pictorial and hybrid questionnaires, implying that their items relate well to each other.

The next part is dedicated to the results addressing questionnaire length and content type. Our first hypothesis (H1) stated that questionnaire length would influence motivation, workload, and satisfaction, favouring the short version. The analysis showed a large effect on workload, but no effects on motivation and satisfaction were found. Instead, a medium effect emerged for comprehension. According to the data, the short versions were perceived as more comprehensible and less demanding than the long ones. In this study, motivation and satisfaction are not directly linked to questionnaire length, or the difference in the number of items between short and long questionnaires was not big enough to provoke meaningful effects. Herzog and Bachmann (1981) argue that questionnaire length is one factor amongst others affecting motivation. An alternative explanation might be that the pictorial character of the scales counteracted potential negative effects related to length, as some researchers argue that they increase motivation and interest (e.g. Haddad et al., 2012). Following our second hypothesis (H2), the manipulation of content type had a large effect on comprehension in favour of the hybrid modality, but contrary to the hypothesis also had large effects on workload and satisfaction and a medium effect on aesthetics. The effects on the first three aspects could be explained by the advantage of the hybrid instrument having a verbal component, thus facilitating the recognition of the intended meaning (Ghiassi et al., 2011) and other aspects related to questionnaire completion (such as workload and satisfaction). The last effect seems at first sight counter-intuitive since the same visualisations were used for pictorial and hybrid scales. We assume that there might have been some kind of an irradiation effect at work, in the sense of 'what is comprehensible is beautiful', based on stereotypes found in social psychology (Dion et al., 1972) and also in the domain of usability and aesthetics research (e.g. Kurosu and Kashimura, 1995; Sauer and Sonderegger, 2009). Taken together, the effect pattern discovered in this study demonstrates that the length of the questionnaire affects perceived comprehension and workload. Furthermore, pictorial and hybrid questionnaires differed on most QX indicators except for motivation, with the hybrid version performing better than the pictorial version.

Concerning the within-subjects comparisons, no significant difference was found in motivation between PUI-L and the verbal questionnaires. This result partially contradicts the assumptions made in H3 and the findings of previous studies, in which pictorial scales were always perceived as more motivating than verbal ones. The other pictorial and hybrid versions were rated significantly better regarding motivation than the verbal questionnaires. One reason might be that some aspects related to questionnaire completion (e.g. increased workload, lowered comprehension) negatively affected the overall experience, thus lowering the rating of motivation.

With regard to questionnaire comprehension, results were also different than assumed in H3. Comprehension of hybrid instruments (HUI-L, HUI-S) was on the same level as the verbal scales. However, it was rated significantly lower for the purely pictorial instruments (PUI-L, PUI-S). One reason might be that there is still too much ambiguity in the meaning of the pictorial items, leading to decreased perceived comprehension. It could also have to do with the sample composition consisting mainly of students, who are more used to interpreting verbal than pictorial content. Ratings of questionnaire workload were highest for PUI-L, in a similar range for PUI-S and HUI-L and the verbal scales, and lowest for the HUI-S. This finding does not support H3 and indicates a different pattern at play. It seems that pictorial content as the only source of information for interpretation, and the greater number of items in long versions generally increases the perceived workload. Results of questionnaire satisfaction showed that participants were more satisfied with both hybrid questionnaires than verbal ones, which confirms assumptions made in H3. Against expected effect patterns in H3, PUI-S was perceived as equally satisfying as the verbal scale, and PUI-L was rated even less satisfying. The last dimension of the QXQ, questionnaire aesthetics, revealed that all pictorial and hybrid versions were perceived as more aesthetically pleasing than the verbal questionnaire, suggesting that pictorial content is prettier to look at than only verbal content. This finding follows the expected effect patterns in H3 and confirms the findings of previous studies.

The results of the QXQ are complemented by the preference rating, which shows that in direct comparison with the verbal scales, nonverbal pictorial scales (PUI-L, PUI-S) are less preferred than hybrid scales (i.e. HUI-L, HUI-S). HUI-S was rated the preferred instrument, with almost two-thirds of participants preferring the pictorial scales to the verbal ones. In contrast, the PUI-L was rated as the least preferred. These findings do not support H3, where we expected equal preference ratings for pictorial and verbal scales but can be considered an additional indicator for the assumption that redundant information in the form of a combination of pictorial and verbal content is superior to only verbal or only pictorial content. One reason might be that both facets of conveying information complement each other, making an abstract concept more tangible than if only one facet of information was presented.

Regarding the respondent-centred measure task completion time, the lowest average completion times were recorded for the PUI-S versions, followed by the PUI-L versions and the verbal usability questionnaires, which took the most time to answer. This difference is no surprise and is owed to the fact that instruments vary in the number of items. Worth noting is that the average item completion time was generally shorter for the verbal scales (Ø 4.49s) than for the PUI version (Ø>6.21s), which follows expected effects in H3 and is consistent with completion times reported in a previous study (Baumgartner et al., 2020).

The analysis of respondent-centred measures suggests a superiority of hybrid instruments and an inferiority of nonverbal instruments, with the short version being more advantageous than the long one. We assume that the main reason for this response pattern in favour of hybrid scales lies in an increased comprehension due to redundant verbal information that frames the decoding of pictorial information and hence facilitates interpretation. In contrast, the nonverbal instruments might be more prone to comprehensibility problems since the pictorial elements are the only source of information for interpretation. Furthermore, the shortness of the scale is another advantage that positively influences most respondent-centred measures.

The present study has some limitations. A large part of the sample consisted of female participants (79.4%). As analysis of this rather large data set did not reveal systematic effects of gender on the various usability and QX ratings, we believe this imbalance should not impinge on the interpretability of our findings. In addition, the sample consisted mainly of student participants (91.9%), representing a rather young and well-educated part of the population. This well-educated sample may have resulted in a better score for the verbal scales since the sample was very literate. Considering this limitation, it must be noted that future studies need to evaluate these instruments with samples with special needs, such as young or illiterate persons or persons of age or foreign language. In this context, validating these instruments in other cultural and ethnic backgrounds might be of interest for future use in research and practice worldwide. Another limitation relates to the online test setting, especially with regard to the interaction with the website prototype, which could only be controlled to a certain degree. However, there is a considerable amount of research in the domain of UX and usability evaluation (Sauer et al., 2019) as well as in research in general (Dandurand et al., 2008; Prissé and Jorrat, 2022; Schidelko et al., 2021), supporting the validity and reliability of findings obtained in online experiments. Finally, respondent-centred measures (QX) for the verbal usability instruments were assessed collectively (i.e. SUS, UMUX-LITE, and three single-item scales). This approach was chosen to simplify the process and reduce the cognitive load on respondents due to questionnaire completion. Consequently, we cannot rule out that results could have differed had we measured QX for each verbal instrument individually. Taking all limitations into consideration, it can be concluded that the findings mentioned above apply to young and well-educated test participants from the western culture, while validation studies with participants of a broader variability regarding needs and requirements as well as cultural background need to be conducted in future research.

Based on the findings of this study, we would like to propose suggestions for future development and research. Results indicated that, on the one hand, longer instruments have better psychometric properties. On the other hand, respondents prefer the short versions over the long ones. A viable compromise for future development could be an instrument with less than eight and more than three items to find a balance between respondents' acceptance and psychometric quality. Another improvement for future versions of PUI could rely on simplifying visual elements, such as a generic interface instead of three device-dependent depictions and a gender-neutral or gender-fluid avatar instead of gender binary representations. These improvements would have a positive impact on the complexity of implementing pictorial or hybrid scales in an online questionnaire. They would also be preferable from a gender point of view. Furthermore, future studies should focus on developing the QXQ, such as refining and extending relevant aspects or providing normative data for interpreting scores. Overall, we believe that the analysis of respondent-centred measures is a valuable extension to the traditional psychometric approach that sheds light on potential benefits and issues in questionnaire assessment.

5. Conclusion

This study is the first that systematically compared pictorial, hybrid, and verbal usability scales concerning psychometric properties and respondent-centred aspects. In conclusion, since the results of this study indicate that all tested pictorial and hybrid versions achieved good psychometric properties, they may all be suitable to be used by researchers and practitioners alike. Taking respondent-centred aspects into consideration, the results of this study suggest advantages of hybrid instruments over pictorial and verbal ones and advantages of short instruments over long ones. Considering the cost-benefit ratio and the respondents' acceptance, the short hybrid version (HUI-S) may be considered the best choice, especially from a practitioner's point of view, when testing time is limited and costly.

CRediT authorship contribution statement

Juergen Baumgartner: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Andreas Sonderegger:** Conceptualization, Writing – review & editing. **Juergen Sauer:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The research was funded by a grant (No 100019_188808) from the Swiss National Science Foundation (SNSF). Their support is gratefully acknowledged. We want to thank 'We Are Cube' and 'Puzzle ITC' for their support in design and technical matters. In particular, we are grateful to Julian Infanger for the code reviews, Mona Nagel and Tania Leitão Carvas for their help in scale development and data collection, Veronica Solombrino for the numerous design reviews, and Dr Alain Chavaillaz and Dr Carli Ochs for their support with the French translations.

References

Assila, A., De Oliveira, K., Ezzedine, H., 2016. Standardized usability questionnaires: features and quality focus. J. Comput. Sci. Inf. Technol. 6 (1), 15–31.
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. J. Usability Stud. 4 (3), 114–123.
Bangor, A., Kortum, P.T., Miller, J.T., 2008. An empirical evaluation of the system usability scale. Int. J. Hum. Comput. Interact. 24 (6), 574–594.
Barnum, C.M., 2011. Usability Testing Essentials: Ready, Set–Test/Carol Barnum. Morgan Kaufmann Publishers, Burlington, MA.
Baumgartner, J., Frei, N., Kleinke, M., Sauer, J., Sonderegger, A., 2019b. Pictorial system usability scale (P-SUS) developing an instrument for measuring perceived usability. In: Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems, pp. 1–11.

Baumgartner, J., Ruettgers, N., Hasler, A., Sonderegger, A., Sauer, J., 2021. Questionnaire experience and the hybrid system usability scale: using a novel concept to evaluate a new instrument. Int. J. Hum. Comput. Stud. 147, 102575.
Baumgartner, J., Sauer, J., Sonderegger, A., 2020. Pictorial usability inventory (PUI) a pilot study. In: Proceedings of the Conference on Mensch Und Computer, pp. 43–52.
Baumgartner, J., Sonderegger, A., Sauer, J., 2019a. No need to read: developing a pictorial single-item scale for measuring perceived usability. Int. J. Hum. Comput. Stud. 122, 78–89.
Betella, A., Verschure, P.F., 2016. The affective slider: a digital self-assessment scale for the measurement of human emotions. PLOS One 11 (2), e0148037.
Borsci, S., Federici, S., Bacci, S., Gnaldi, M., Bartolucci, F., 2015. Assessing user satisfaction in the era of user experience: comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. Int. J. Hum. Comput. Interact 31 (8), 484–495.
Borsci, S., Federici, S., Lauriola, M., 2009. On the dimensionality of the System Usability Scale: a test of alternative measurement models. Cogn. Process 10, 193–197.
Brooke, J., 1996. SUS-A quick and dirty usability scale. In: Jorden, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (Eds.), Usability Evaluation in Industry. Taylor and Francis, pp. 189–194.
Collaud, R., Reppa, I., Défayes, L., McDougall, S., Henchoz, N., Sonderegger, A., 2022. Design standards for icons: the independent role of aesthetics, visual complexity and concreteness in icon design and icon understanding. Displays 74, 102290.
Coolican, H., 2017. Research Methods and Statistics in Psychology. Psychology press.
Dandurand, F., Shultz, T.R., Onishi, K.H., 2008. Comparing online and lab methods in a problem-solving experiment. Behav. Res. Methods 40 (2), 428–434.
Desmet, P., 2003. Measuring emotion: development and application of an instrument to measure emotional responses to products. Funology. Springer, pp. 111–123.
Desmet, P., Overbeeke, K., Tax, S., 2001. Designing products with added emotional value: development and appllcation of an approach for research through design. Des. J. 4 (1), 32–47.
Dion, K., Berscheid, E., Walster, E., 1972. What is beautiful is good. J. Pers. Soc. Psychol. 24 (3), 285–290. https://doi.org/10.1037/h0033731.
Galesic, M., Bosnjak, M., 2009. Effects of questionnaire length on participation and indicators of response quality in a web survey. Public Opin. Q. 73 (2), 349–360.
Gao, M., Kortum, P., Oswald, F.L., 2020. Multi-language toolkit for the system usability scale. Int. J. Hum. Comput. Interact. 36 (20), 1883–1901.
Ghiassi, R., Murphy, K., Cummin, A.R., Partridge, M.R., 2011. Developing a pictorial Epworth sleepiness scale. Thorax 66 (2), 97–100.
Haddad, S., King, S., Osmond, P., Heidari, S., 2012. Questionnaire design to determine children's thermal sensation, preference and acceptability in the classroom. In: Proceedings of the 28th International PLEA Conference on Sustainable Architecture + Urban Design: Opportunities, Limits and Needs-towards an Environmentally Responsible Architecture.
Heberlein, T.A., Baumgartner, R., 1978. Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature. Am. Sociol. Rev. 43 (4), 447–462.
Herzog, A.R., Bachman, J.G., 1981. Effects of questionnaire length on response quality. Public Opin. Q. 45 (4), 549–559.
Hicks, K.E., Bell, J.L., Wogalter, M.S., 2003. On the prediction of pictorial comprehension. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 47, pp. 1735–1739.
International Organization for Standardization, 2014. ISO 9186-1:2014. ISO. https://www.iso.org/standard/59226.html.
International Organization for Standardization, 2019. ISO 9241-210:2019. ISO. https://www.iso.org/standard/77520.html.
Kurosu, M., Kashimura, K., 1995. Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability. In: Proceedings of the Conference Companion on Human Factors in Computing Systems, pp. 292–293. http://dl.acm.org/citation.cfm?id=223680.
Lewis, J., 2018. The system usability scale: past, present, and future. Int. J. Hum. Comput. Interact. 34 (7), 577–590.
Lewis, J.R., 2002. Psychometric evaluation of the PSSUQ using data from five years of usability studies. Int. J. Hum. Comput. Interact. 14 (3–4), 463–488. https://doi.org/10.1080/10447318.2002.9669130.
Lewis, J.R., Utesch, B.S., Maher, D.E., 2013. UMUX-LITE: when there's no time for the SUS. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2099–2102. https://doi.org/10.1145/2470654.2481287.
Lewis, J., Sauro, J., 2017. Revisiting the factor structure of the system usability scale. J. Usability Stud. 12 (4).
Nielsen, J., 1994. Usability Engineering. Elsevier.
Nielsen, J., Molich, R., 1990. Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 249–256.
Nunnally, J.C., Bernstein, I.H., 1994. Psychometric Theory (McGraw-Hill Series in Psychology, 3. McGraw-Hill, New York.
Prissé, B., Jorrat, D., 2022. Lab vs online experiments: no differences. J. Behav. Exp. Econ. 100, 101910.
Reichheld, F.F., 2003. The one number you need to grow. Harv. Bus. Rev. 81 (12), 46–55.
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. Personal. Soc. Psychol. Bull. 27 (2), 151–161.
Ryan, R.M., 1982. Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. J. Pers. Soc. Psychol. 43 (3), 450.
Sauer, J., Baumgartner, J., Frei, N., Sonderegger, A., 2021. Pictorial scales in research and practice. Eur. Psychol. 26 (2), 112–130.
Sauer, J., Sonderegger, A., 2009. The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion. Appl. Ergon. 40 (4), 670–677.
Sauer, J., Sonderegger, A., Heyden, K., Biller, J., Klotz, J., Uebelbacher, A., 2019. Extra-laboratorial usability tests: an empirical comparison of remote and classical field testing with lab testing. Appl. Ergon. 74, 85–96.
Sauer, J., Sonderegger, A., Schmutz, S., 2020. Usability, user experience and accessibility: towards an integrative model. Ergonomics 63 (10), 1207–1220.
Sauro, J., Lewis, J.R., 2016. Quantifying the User Experience: Practical Statistics For User Research. Morgan Kaufmann.
Schidelko, L.P., Schünemann, B., Rakoczy, H., Proft, M., 2021. Online testing yields the same results as lab testing: a validation study with the false belief task. Front. Psychol. 4573.
Schmutz, S., Sonderegger, A., Sauer, J., 2019. Easy-to-read language in disability-friendly web sites: effects on nondisabled users. Appl. Ergon. 74, 97–106.
Sonderegger, A., Heyden, K., Chavaillaz, A., Sauer, J., 2016. AniSAM & AniAvatar: animated visualizations of affective states. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 4828–4837.
Tavakol, M., Dennick, R., 2011. Making sense of Cronbach's alpha. Int. J. Med. Educ. 2, 53.
Toepoel, V., Vermeeren, B., Metin, B., 2019. Smileys, stars, hearts, buttons, tiles or grids: influence of response format on substantive response, questionnaire experience and response time. Bull. Sociol. Methodol./Bull. Méthodol. Sociol. 142 (1), 57–74.
Wilde, M., Bätz, K., Kovaleva, A., Urhahne, D., 2009. Überprüfung einer Kurzskala intrinsischer Motivation (KIM). Z. Didakt. Naturwiss. 15, 31–45.
Wissmath, B., Weibel, D., Mast, F.W., 2010. Measuring presence with verbal versus pictorial scales: a comparison between online- and ex post-ratings. Virtual Real 14 (1), 43–53. https://doi.org/10.1007/s10055-009-0127-0.