Alternative Questionnaire Formats in Usability Evaluation: A Comparison of Pictorial and Hybrid Scales

Jürgen Johann Baumgartner
Oberriet-Montlingen (SG)

2023

Kumulative Dissertation zur Erlangung der Doktorwürde an der Philosophischen Fakultät der Universität Freiburg (Schweiz). Genehmigt von der Philosophischen Fakultät auf Antrag der Herren Professoren Jürgen Sauer (1. Gutachter), Andreas Sonderegger (2. Gutachter) und Kai-Christoph Hamborg (3. Gutachter). Freiburg, den 07.03.2024 (Datum der Thesenverteidigung). Prof. Dominik Schöbi, Dekan.

Abstract

This doctoral thesis investigated the use of image-based scales in usability assessment, examining their advantages and limitations. While previous research attributed favourable qualities to image-based scales, such as increased motivation, intuitive comprehension, reduced workload, and reduced completion times, these claims often lacked solid empirical support. This work aimed to systematically evaluate variations of image-based scales in relation to these respondent-centred aspects, which are subsumed under the concept of questionnaire experience (QX). Furthermore, traditional psychometric properties were assessed. Three studies were conducted in which image-based scales were used in an online usability test setting. Study one used a hybrid version (i.e. pictorial and verbal content) of an existing usability questionnaire. Study two compared hybrid and purely pictorial scales in short and long versions. Study three introduced an animated hybrid scale. The findings from the three studies suggest that pictorial, hybrid and animated scales demonstrated satisfactory psychometric properties, making them viable alternatives for assessing perceived usability. However, QX between pictorial and hybrid scales differed considerably. Hybrid instruments received higher ratings on respondent-centred aspects and were more preferred. Furthermore, the findings highlight that some of the favourable qualities of image-based scales appeared to be too optimistic (e.g. reduced completion times). Nevertheless, the notion of increased motivation when using image-based scales could be largely supported. Theoretical and practical implications are discussed.

https://doi.org/10.51363/unifr.lth.2024.030

Table of contents

1. Introduction

Long before the written word, images communicated ideas and experiences. The oldest preserved cave paintings are dated 40 ka (Aubert et al., 2018). They are exemplary evidence that long before the development of the first writing system (e.g. Sumerian, 3200 BCE, Steymans et al., 2012), concrete concepts were communicated with images. The desire to express and interpret visual content may be somewhere in our nature. Although the visualisation of ideas and concepts endured in some form (e.g. craftsmanship of any kind, picture books), the written word took over for the sake of communication efficiency.

Also profoundly rooted in human nature is the desire to learn and gain knowledge about how the world works. Especially empirical research is an important cornerstone of modern science, deriving knowledge from actual experiences and observations (e.g. Harari, 2014). A common and efficient way of gathering vast amounts of information about individuals' opinions and beliefs is using verbal questionnaires and surveys. The history of questionnaires is relatively young, beginning in the middle of the 19th century (Gault, 1907). Over time, quantitative questionnaires and screening methods became popular and were widely used in various disciplines (e.g. Army Alpha test for evaluation of recruits, Yerkes, 1921; Stanford-Binet Intelligence Scales, Terman & Merrill, 1960; Rosenberg Self-Esteem Scale, Rosenberg, 1965).

One domain of particular interest in this work is the field of usability evaluation. Usability has become important for many industries, such as software development and product design. It is based heavily on collecting opinions from individuals to identify problems of interactive goods. The usability domain has its roots in ergonomics and human factors, dating back to the first attempts to improve industrial efficiency (e.g. Taylor, 1911). It was formalised with the advent of commercially available personal computers in the early 1980s. Usability engineering has gradually become essential to product development, with user research at its core. Opinions and experiences of individuals are gathered to identify issues of interactive products, relying considerably on standardised questionnaires, among other methods.

Verbal questionnaires have their merits, such as being cost-effective and time-efficient. Nevertheless, there are circumstances in which verbal questionnaires also have their drawbacks. (1) When questionnaires are exceedingly long or several questionnaires of a similar kind are administered, monotony sets in, and respondents' motivation might suffer.

This can negatively impact the respondents' answering behaviour (e.g. inaccurate answers) and lead to poorer data quality (Herzog & Bachman, 1981). (2) Furthermore, some user groups might have problems processing verbal questionnaires, such as children, people with reading difficulties or reading disorders (e.g. some form of dyslexia), but also non-native speakers, people with low education levels or illiterate people (Ghiassi et al., 2011; Paunonen et al., 2001; Sonderegger et al., 2016). This language barrier can lead to comprehension issues and bias the results. (3) Frequently, the availability of standardised instruments in a specific target language (e.g. German) is problematic if they have only been validated in a different language (e.g. English). This shortcoming can tempt practitioners and researchers to translate questionnaires themselves, with unclear consequences for the quality of the translation if no appropriate translation procedure has been used.

This work aims to evaluate the suitability and usefulness of alternative image-based questionnaire types in the domain of usability evaluation, so-called pictorial scales. The rationale behind it is to offer all respondents an accessible approach to completing questionnaires while at the same time adequately capturing the underlying construct. Furthermore, such questionnaires were designed with the intent to stimulate engagement and to offer more variety in the otherwise text-dominated world while at the same time benefiting from our innate ability to process image-based material easily.

This work entails three empirical studies in which various image-based scales were developed and compared with a traditional verbal usability questionnaire. Psychometric properties and respondents' subjective experience were considered. Both concepts were used to determine the quality of the scales and to assess the strengths and weaknesses of pictorial scales. Besides, this work addresses claims made in the literature concerning the advantages and disadvantages of pictorial scales that have never been tested empirically.

2. Theoretical background

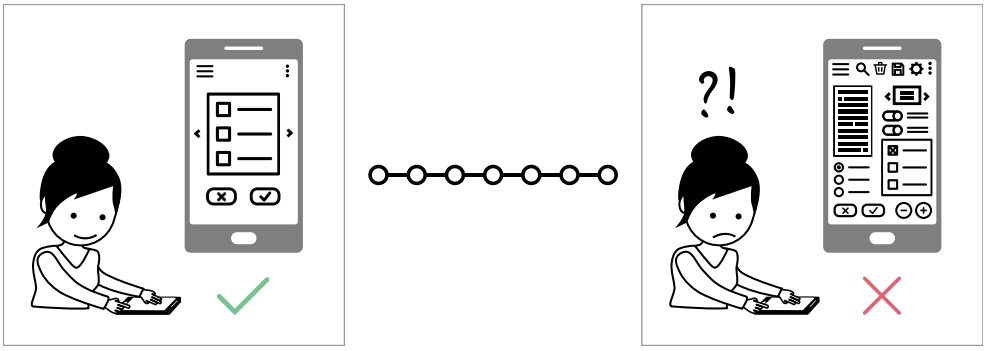
2.1. Pictorial scales in research and practice

Unlike verbal questionnaires, pictorial questionnaires have a short tradition. The first of its kind used faces with different expressions to measure job satisfaction (Kunin, 1955). Since the 1950s, over 60 pictorial instruments have been published in various research domains (Sauer et al., 2020). Although a large part of such instruments is related to evaluating

emotions, also more abstract concepts were considered suitable to assess using pictorial scales (e.g. presence, Wissmath et al., 2010).

Sauer and colleagues (2020) suggested a definition of pictorial scales, describing them as instruments that make ‘use of image-based elements to convey the meaning of its items‘ (p.1). Some instruments rely on image-based content only, but others use a combination of verbal and graphical elements, so-called hybrid scales (cf. Baumgartner et al., 2021). Furthermore, some instruments extend the static representation of image-based scales by integrating animations to convey more information (Laurans & Desmet, 2017). Figure 1 shows an item of a pictorial multi-item scale that was used to assess perceived usability (cf. Baumgartner et al., 2019b).

Figure 1. Item of the P-SUS (Pictorial System Usability Scale) depicting a female user interacting with a smartphone user interface (easy vs complex).



Various advantages of pictorial scales are mentioned in the literature. The most important advantages are: (1) they provide pleasure and increase respondents’ engagement (Desmet, 2003; Ghiassi et al., 2011), (2) they are intuitively comprehensible (Bradley & Lang, 1994; Kunin, 1955) and therefore less mentally demanding than verbal scales (Wissmath et al., 2010), and (3) they are language-independent, and therefore eliminate the need to translate them (Betella & Verschure, 2016). Furthermore, it has been argued that pictorial scales are well suited when participants have insufficient competence in a target language, such as non-native speakers, children or people with poor language skills or limited reading ability (Ghiassi et al., 2011; Paunonen et al., 2001; Sonderegger et al., 2016). Notably, claims such as increased motivation, reduced mental workload, and improved comprehensibility have been widely accepted without empirical testing.

There are also disadvantages mentioned in the literature, for instance (1) that pictorial content offers the potential for misinterpretation when participants assign a different meaning to the

visualisation than the intended one (e.g. Betella & Verschure, 2016). (2) Additionally, cultural differences are mentioned to impact how pictorial content such as gestures are interpreted (e.g. Sauer et al., 2020). (3) Furthermore, the creation and validation of pictorial scales are resource-consuming, especially when they need to be customised for a specific research context (e.g. Desmet et al., 2016).

Although there may be some drawbacks, the advantages of such instruments appear to outweigh the potential disadvantages under certain circumstances, resulting in the development of more instruments across various research fields, including usability evaluation.

2.2. Usability evaluation

In the early 1980s, the advent of personal computers led to the rapid development and growth of the computer technology segment. As computers became more widely used in work and later in leisure contexts, the need to design and evaluate computer systems increased considerably (Lewis, 2018). This trend has continued over the past 40 years, with the proliferation of digital technologies in many aspects of our lives. User-centred design is applied to create interactive systems that meet users' needs, considering human factors, ergonomics, and usability techniques (cf. ISO 9241-210, 2019).

The International Organisation for Standardisation (2019) defines usability as the degree to which specified users can use an interactive system to accomplish specified objectives within a specified usage context while maintaining effectiveness, efficiency, and user satisfaction (ISO 9241-210). The first two components describe performance aspects, whereas the latter is focused on subjective user perception. Since the quality of interacting with a technical system has many facets that can be described as non-instrumental or hedonic (for instance, beauty, pleasure, and emotions), the concept of usability was deemed too narrow, and that effectiveness, efficiency and satisfaction do not fully cover the individual's experience (Norman, 2004; Robert & Lesage, 2017). For this reason, usability was integrated into the umbrella concept of user experience (UX), where it is considered a crucial part of a positive overall experience (such as whether an interactive product is perceived as intuitive or easy to use).

One highly effective method for gathering data on how users perceive the usability of a product is through questionnaires. Over the past 30 years, more than 20 standardised instruments have been developed for this purpose (see Assila & Ezzedine, 2016, for a comprehensive overview). They differ vastly in the number of items (e.g. UMUX LITE, Lewis et al., 2013, 2 items vs PUTQ, Lin et al., 1997, 100 items) or the target product to assess (e.g. websites, mobile devices). However, they all have in common that they use verbal items. Most of these questionnaires were initially developed in the English language. Only a few instruments have been validated in other target languages, such as the System Usability Scale (SUS, Brooke, 1996), of which versions in various languages exist (Gao et al., 2020). All these questionnaires have in common that they are verbal. Pictorial scales have been introduced as alternative methods to avoid relying solely on the comprehensibility of verbal scales, facilitating an inclusive approach to evaluating users' attitudes.

Measuring perceived usability with pictorial scales is a relatively recent approach. In work prior to this doctoral thesis, we created several pictorial scales to assess the perceived usability of an interactive product. Besides the three instruments outlined in this work, three other image-based scales have already been created and tested in iterative cycles (Baumgartner et al., 2020, 2019b, 2019a). Methods such as think-aloud protocols (Lewis & Mack, 1982) or comprehension tests (ISO 9186-1; International Organization for Standardization, 2014) were used to determine respondents' interpretation and gradually improve the scales. Pilot studies were conducted using lab or online test settings to determine their psychometric properties (i.e. validity, reliability, sensitivity). Table 1 gives an overview of the pictorial usability scales. The results indicate generally high convergent validity and mostly substantial effect sizes for sensitivity.

Table 1. Overview of pictorial usability scales and their psychometric characteristics.

Instrument	Test setting	Item number	*N*	Convergent validity (*r*)	Sensitivity (*r*)	Cronbach Alpha (α)	Reference
PSIUS (Pictorial Single Item Usability Scale)	Lab	1	60 38	.881 .696	.360 .550	-	Baumgartner et al., 2019a
P-SUS (Pictorial System Usability Scale)	Lab	10	60	.865	.666	.912	Baumgartner et al., 2019b
PUI (Pictorial Usability Inventory) – v1	Online	12	64	.852	.465	.961	Baumgartner et al., 2020

Notes: SUS was used in all studies to measure convergent validity.

While psychometric properties are important for assessing the quality of newly developed scales, they cannot fully capture the unique advantages of pictorial scales. Therefore, we found it necessary to complement the psychometric assessment with additional experiential measures to get a complete picture of the usefulness and effectiveness of pictorial scales.

2.3. Beyond psychometrics – the evaluation of questionnaire experience

The assessment of psychometric quality is regarded as the central pillar of evaluating the quality of a questionnaire (Miller & Lovler, 2018). Reliability and validity are crucial to obtaining sound measurements (Hinkin, 1995). Therefore, rigorous procedures for scale development were proposed to provide evidence that a specific test measures what it is supposed to measure and that results are consistent and dependable (e.g. DeVellis, 2016; Hinkin, 1995). Besides systematic procedures to determine psychometric properties, the literature suggests approaches and guidelines on item creation, selection, and formulation. For the latter, concrete recommendations exist, for instance, avoiding jargon, not using both positive and negative wording, avoiding ambiguity and double-barreled questions, and keeping items concise and not overly lengthy (DeVellis, 2016; Streiner et al., 2015).

Although these recommendations are valuable for improving aspects such as the comprehensibility and simplicity of items, there is no obligation to consider them or to test items with actual users of the target population. Furthermore, such approaches are only conducted during item creation. After achieving acceptable psychometric quality in a validation study, considerations regarding the questionnaire's ease of comprehension, optimal length, and perceived workload are often neglected. Furthermore, recommendations of proper formulation can only be implemented when the items are of a verbal nature. When evaluating pictorial scales, alternative ways of assessing aspects such as comprehensibility, motivation, or perceived workload are needed.

We developed the concept of questionnaire experience (QX) to fill the gap and evaluate important non-psychometric aspects. The assessment of QX is a respondent-centred approach with the goal of evaluating a questionnaire from the respondents' point of view. QX consists of aspects that are relevant during questionnaire completion. The term was first mentioned by Toepoel and colleagues (2019), who used the term as a designation for a global measure assessing the experience after having answered pictorial Likert scales (e.g. rating with smileys, stars, or hearts). Inspired by the definition of user experience, we defined QX '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' (Baumgartner et al., 2020, p. 2).

The rationale behind QX is to improve questionnaires by systematically evaluating respondents' perceptions. This approach helps identify weaknesses of an instrument regarding important respondent-centred aspects (e.g. scale comprehensibility). By subsequently addressing these aspects (e.g. refining scale wording), a more positive experience can be fostered, and negative outcomes such as boredom or fatigue can be reduced. The aspects we considered part of QX developed from one study to the next and were primarily used as measures of comparison to identify differences between verbal and pictorial scales or between variations of pictorial scales (e.g. pictorial vs hybrid). Study one considered aspects such as motivation, workload, and preference. For study two, comprehension, satisfaction, and aesthetic appeal were added. Finally, study three built upon the same aspects as study two while incorporating subjective time perception. These aspects are not final but an encouraging start for obtaining insights into respondents' experiences. Finally, including a respondent-centred view in scale development is not about replacing psychometrics properties but complementing them. The objective is to facilitate the creation of instruments that reliably measure what they are supposed to measure while also ensuring a positive questionnaire experience, thus keeping respondents motivated throughout the process.

3. The present work

3.1. Overview of studies

This work contains three online studies that evaluate the appropriateness and efficacy of pictorial, hybrid, and animated scales in the context of usability evaluation. Besides psychometric properties, respondent-centred measures were assessed to determine the quality of the instruments. The System Usability Scale was used in all studies as a principal measure of comparison, and usability was manipulated to determine sensitivity. Table 2 provides the key facts of the studies, followed by a summary and the full-text versions of the journal publications.

Table 2. Overview of studies, including research objective, sample size, study design, independent and dependent variables, and main findings.

	Study one	Study two	Study three
Research objective	Comparison of a hybrid SUS (H-SUS) with the original SUS	Comparison of PUI with HUI (long and short versions)	Comparison of static HUI with animated HUI (AniHUI)
Sample size	*N*=152	*N*=777	*N*=192
Study design	1-factorial	2x2	1-factorial
Independent variables	Questionnaire type (verbal vs hybrid)	Questionnaire type (pictorial vs hybrid); questionnaire length (short vs long)	Questionnaire type (static vs animated)
Dependent variables	Psychometric properties, respondent-centred measures	Psychometric properties, respondent-centred measures	Psychometric properties, respondent-centred measures
Main findings	SUS and H-SUS had very similar psychometric properties. H-SUS was preferred and obtained better motivation ratings. SUS had shorter completion times.	The long version of PUI had the best psychometric properties. The short hybrid version enjoyed the best questionnaire experience.	Static and animated scales had very similar psychometric properties. AniHUI did not considerably differ in questionnaire experience.

Notes: SUS=System Usability Scale; H-SUS=Hybrid System Usability Scale; PUI=Pictorial Usability Inventory; HUI=Hybrid Usability Inventory; AniHUI=Animated Hybrid Usability Inventory

3.2. Study One (H-SUS) – Summary

Study one aimed to develop and test a hybrid version of an established usability questionnaire that measures perceived usability. For this reason, all ten items of the System Usability Scale (Brooke, 1996) were visualised and complemented with the original wording of the scale. The purpose of the study was to investigate potential advantages and disadvantages associated with the use of a hybrid System Usability Scale (H-SUS). Respondent-centred aspects of QX were assessed, consisting of motivation, workload, preference, and completion time. We

hypothesised that H-SUS would have similar psychometric properties as the SUS but better QX. An online study was conducted in which 152 participants interacted with an app prototype (low vs high usability) and subsequently completed SUS and H-SUS. The results of the study showed that H-SUS and SUS were very similar in psychometric quality. They distinguished equally well between usability levels, had high internal consistency, correlated strongly with each other, and showed similar correlational patterns with divergent and criterion-related measures. Major differences were found in respondent-centred measures. Most participants preferred H-SUS, which was also evaluated as more motivating to complete than the SUS. However, SUS obtained shorter completion times than the H-SUS. Although results indicated that H-SUS has good psychometrics and QX scores, some items still appeared to be ambiguous, raising concerns about their suitability. Furthermore, H-SUS uses verbal information and cannot be considered a nonverbal instrument. Therefore, we explored further options and considered alternative pictorial scale development procedures for the following study.

3.3. Study Two (PUI/HUI) – Summary

Study two compared pictorial and hybrid scales regarding respondent-centred measures and traditional psychometric properties. Besides, pictorial and hybrid scales were presented in a long and a short version (8 vs 3 items). In contrast to study one, the scales were not based on a specific verbal usability questionnaire but on items from different usability questionnaires (see Baumgartner et al., 2020 for more details). The main interest of this study was to systematically assess differences between the pictorial and hybrid scales and their long and short versions regarding QX and psychometrics. The same set of respondent-centred measures was used as in study one but complemented with a few additional measures (i.e. comprehension, satisfaction, and aesthetic appeal). Furthermore, we wanted to test a more compact item visualisation that also works on small screens. An online experiment was conducted in which 777 participants interacted with a website prototype (low vs high usability) and subsequently completed verbal usability instruments (i.e. the SUS, UMUX-LITE, and single-item measures) and one of the four pictorial scales. The results showed that all pictorial and hybrid versions obtained good psychometric quality, but the hybrid short version was rated best on respondent-centred measures. Furthermore, pictorial scales had lower scores than the hybrid scales in almost all respondent-centred measures. For the subsequent study, we wanted to build upon the hybrid scale development and explore whether other means of representation would improve respondent-centred measures even more.

3.4. Study Three (HUI/AniHUI) – Summary

In the third study, we aimed to investigate further ways of improving image-based scales. For this reason, we considered extending the static representation of a hybrid usability scale by adding animations. The main interest of this study was to compare the static hybrid scale (HUI) with an animated one (AniHUI). The verbal SUS served as a yardstick for assessing convergent validity. We hypothesised that an animated scale would influence the respondents' engagement and perceived experience with the questionnaire more positively than the hybrid or verbal questionnaire. The same respondent-centred measures used in study two were employed to assess QX, with the addition of perceived questionnaire completion time. Besides, a new genderfluid avatar was used as the main character for static and animated scales. An online study was conducted with 192 participants who interacted with the same website as in study two. The inherent usability of the website was manipulated to create two conditions: low usability and high usability. After the interaction, participants completed the SUS and either the HUI or the AniHUI. This study showed no striking difference between static and animated scales in terms of psychometric quality and QX. Contrary to our assumption, participants were not more engaged with the animated scale than with the static one. However, both scales were rated better regarding questionnaire motivation, aesthetic appeal, and perceived completion time. These results reinforce findings from studies I and II that found similar effects. While this study did not find any additional advantages of animated scales compared to static scales (e.g. increased motivation), it is important to acknowledge that different outcomes may emerge if the scales or the interaction with the scales were designed differently.

4. Study One – Questionnaire experience and the hybrid System Usability Scale: Using a novel concept to evaluate a new instrument

Int. J. Human–Computer Studies 147 (2021) 102575

International Journal of Human - Computer Studies

Questionnaire experience and the hybrid System Usability Scale: Using a novel concept to evaluate a new instrument

Juergen Baumgartner [a,b,*], Nicole Ruettgers [a], Annigna Hasler [a], Andreas Sonderegger [c,a], Juergen Sauer [a]

[a] Department of Psychology, University of Fribourg, Rue P.-A.-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

ARTICLE INFO

Keywords:
Hybrid scale
Questionnaire experience
Consumer product evaluation
Perceived usability
Mobile device evaluation

ABSTRACT

This article presents the concept of questionnaire experience (QX), intending to add a new element to the psychometric evaluation of questionnaires, which may eventually help increase the validity and reliability 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 would 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 of 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.

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 and 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 validity, reliability, and, in certain cases, sensitivity. There are other criteria, which are also essential but have not received the same level of attention, though they may equally contribute to the improvement of the psychometric properties of questionnaires. These criteria refer to the experience of the respondent during questionnaire completion, which may not always be positive (e.g., the 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 and 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., the conscientiousness of questionnaire completion, the motivation to complete questionnaire again), which in turn could possibly affect the psychometric properties of the instrument.

In addition to the introduction of the concept of QX, we also examine

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

https://doi.org/10.1016/j.ijhcs.2020.102575
Received 2 July 2020; Received in revised form 6 November 2020; Accepted 2 December 2020
Available online 5 December 2020

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., 2020). Due to their visual nature, hybrid scales are expected to influence QX positively.

We believe that both principal issues dealt with in this article (i.e. hybrid scales, the 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 the formulation of good items (e.g. Thielsch et al., 2012), paying attention to questionnaire length (Galesic and Bosnjak, 2009), completing qualitative item analyses, carrying out expert reviews, and conducting a pilot test (Miller and Lovler, 2018). Before publishing a questionnaire, there are further steps to follow, such as assessing psychometric criteria (e.g., validity and reliability), having the questionnaire reviewed by test takers, and using expert panels to assess content validity (Miller and 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 and 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 conceptualised 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 and 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).

Fig. 1 shows how QX has been conceptualised. 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 Fig. 1 by using a solid line to designate theoretical constructs (i.e. not directly measurable) and a dotted line to designate measurable indicators.

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, the 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 of 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., 2020).

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 hybrid. 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. Since both instruments were developed for children, the verbal part is read out by the scale administrator while the pictorial part is used to illustrate the meaning of the item 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 and Lang, 1994; Desmet, 2003; Sonderegger et al., 2016). Concerning 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). The latter is based on the established System Usability Scale (Brooke, 1996).

The use of a 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). Both cues should 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 and Walrath, 1995). If one of

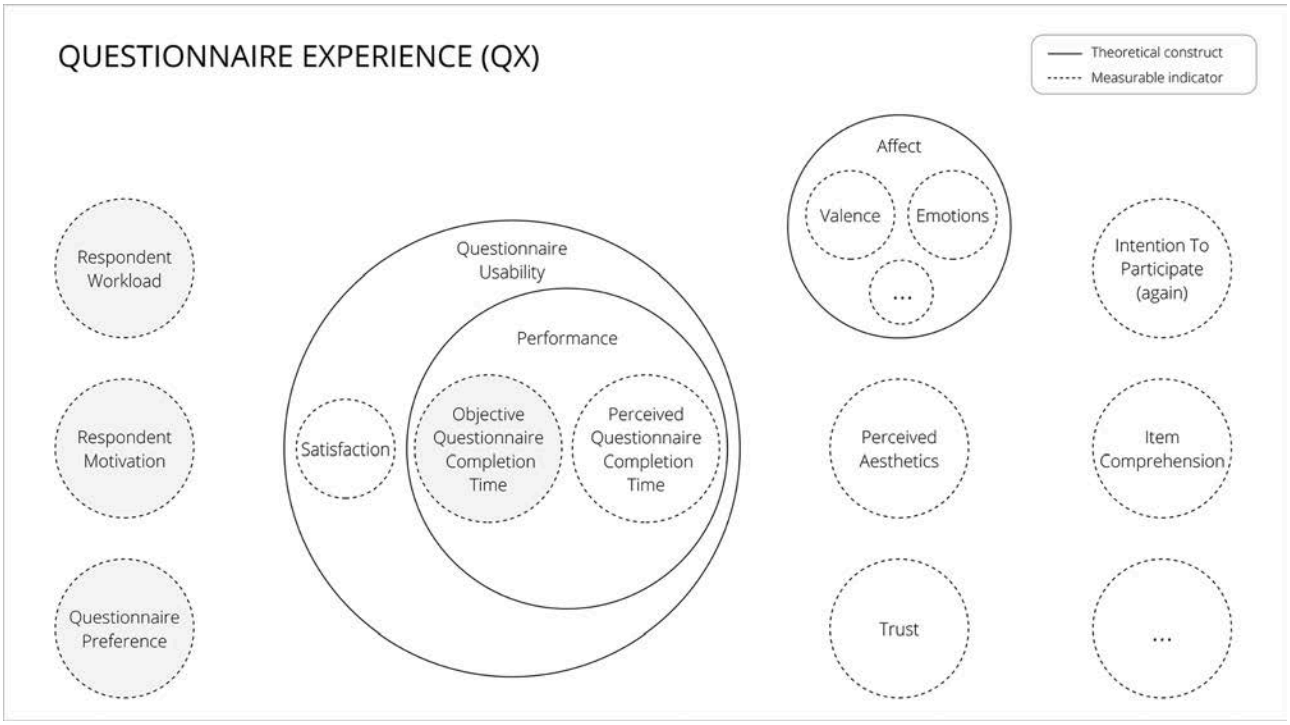


Fig. 1. The conceptualisation of the new term 'questionnaire experience' describing its constituting elements (grey circles denote indicators that were measured in empirical study).

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 verbal and 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 delay 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 satisfying 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 and 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, this article aims 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 (such as validity, reliability, and sensitivity), 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 Fig. 2).

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 the corresponding SUS item. A male and a female version of the avatar were developed with identical content to increase respondents' 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 by comparing it to the well-established verbal SUS. The psychometric properties assessed included convergent validity, divergent validity, criterion-related validity, reliability in the form of internal consistency, and sensitivity. 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

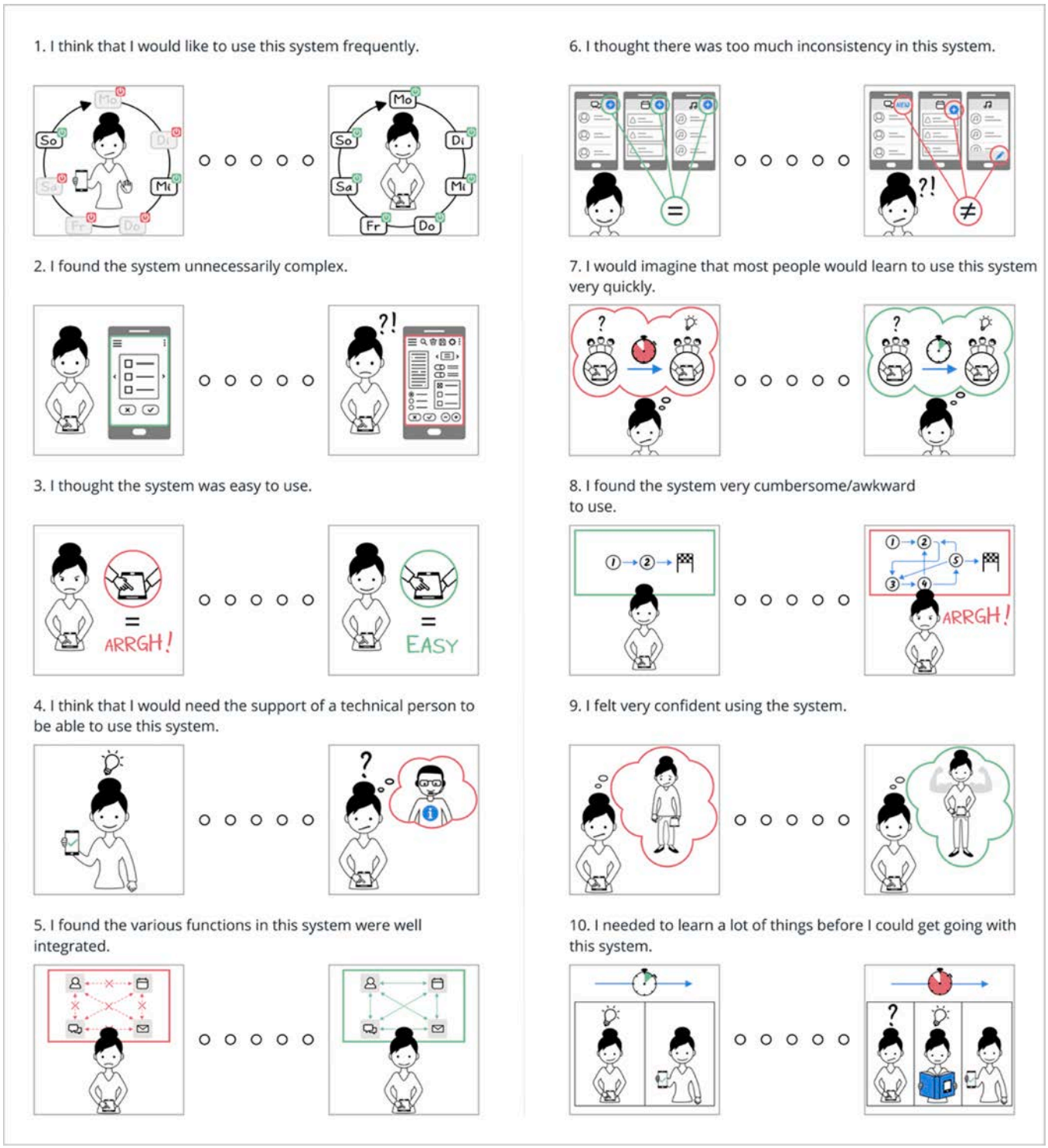


Fig. 2. H-SUS items (female version) with verbal content and the five-point rating scale using pictorial representations for the positive and negative end points.

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. Besides, participants were asked at the end of the study

to forward the link to their friends. Five vouchers worth €50 each were raffled to increase participant motivation.

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 having 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. They 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, measures of QX were considered, such as (6) respondent workload, (7) respondent motivation to complete the questionnaire, (8) questionnaire preference, and (9) questionnaire completion time.

Convergent Validity. Convergent validity 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. This instrument 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. Divergent validity 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. Affect and visual aesthetics were assessed to obtain a measure of divergent validity. Affect was measured using the AniSAM (Sonderegger et al., 2016), which is a nonverbal instrument based on the Self-Assessment Manikin (SAM; Bradley and Lang, 1994). The instrument 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 an indicator for physiological arousal. The intensity of arousal is indicated by the frequency by which the heart beats. For the assessment of visual aesthetics, the short version of the Visual Aesthetics of Websites Inventory (VisAWI-S) was used (Moshagen and 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 $\alpha = .81$).

Criterion-related Validity. Criterion-related validity 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, 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 and 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. The short version of the Intrinsic Motivation Inventory (IMI; Wilde et al., 2009) was used to assess the motivation of questionnaire completion. 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 ($\alpha = .85$).

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. The online questionnaire automatically recorded the completion time for each item. Completion times of 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 and colleagues (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 the 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 an '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 and 15 minutes to carry out the tasks and to complete the online questionnaire. On the first page, an image of a male and a female avatar was presented to the participants. 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. Participants completed subsequently 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. Before 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, and to determine respondent workload and motivation, and questionnaire completion time by using Wilcoxon signed-rank tests. 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

The psychometric criteria of the tested instruments are described in the following paragraphs. Fig. 3 summarises the main results of analysing the psychometric criteria.

3.3.1.1. Convergent validity. In Fig. 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).

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.

Concerning 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.

3.3.1.3. Criterion-related validity. For the assessment of criterion-related validity, correlations of performance measures (task completion time and the number of interactions) with both H-SUS 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 the number of interactions, correlations were similar for the H-SUS and SUS, at around r = −.600.

3.3.1.4. Internal consistency. Fig. 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 (α = .91). Similarly, a high internal consistency score was found for the SUS (α = .91).

3.3.1.5. Sensitivity. In Fig. 3c, usability scores in low and high-usability conditions are presented for H-SUS and SUS. A Mann-Whitney test was carried out to assess whether there is a difference between low and high usability. The 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. Fig. 4 summarises the main results of analysing the different QX measures.

3.3.2.1. Respondent workload and motivation. Fig. 4a summarises 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

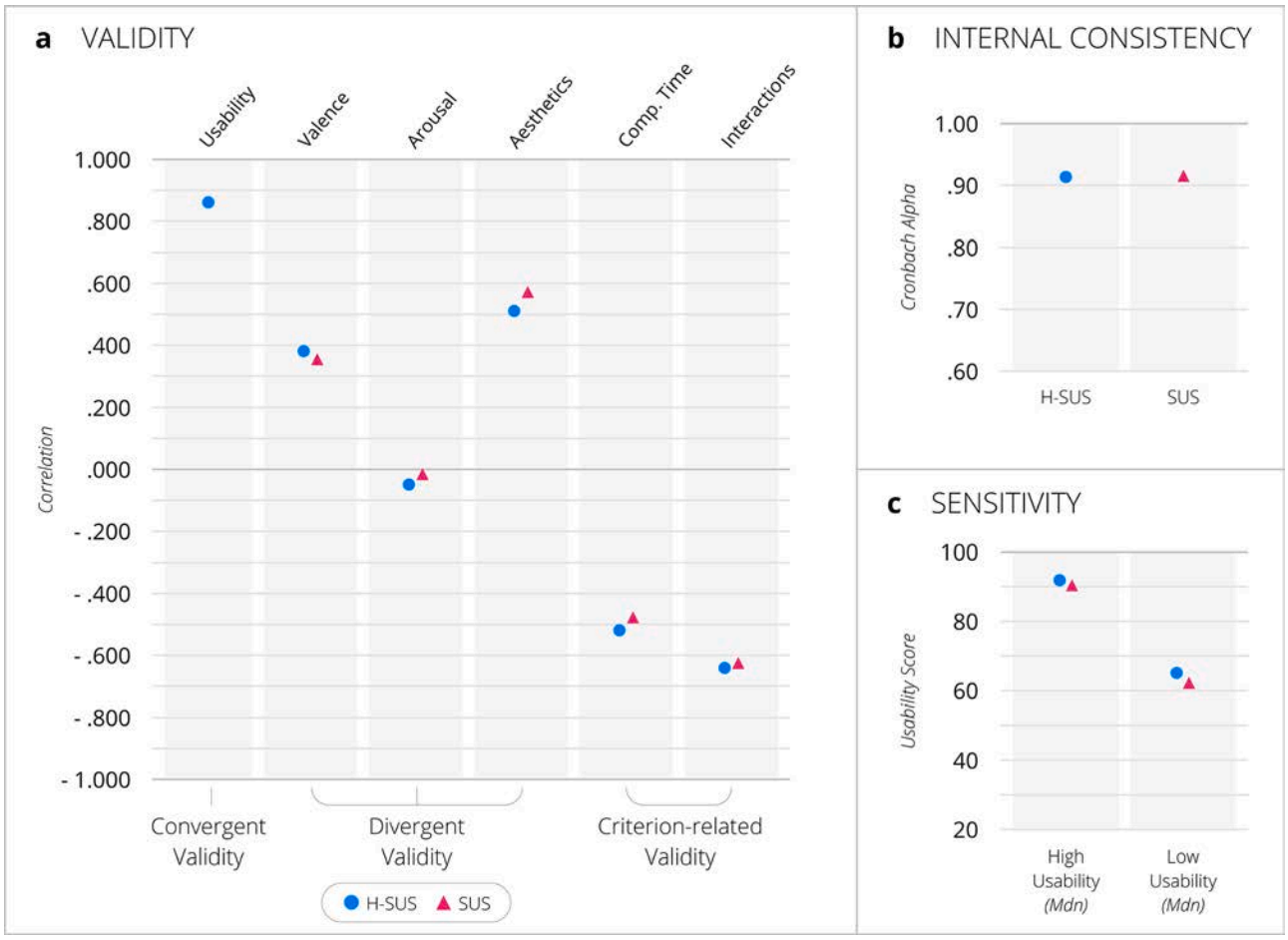


Fig. 3. Comparative analysis of psychometric criteria of H-SUS and SUS: (a) Correlations of usability scores with scores of convergent, divergent and criterion-related validity, (b) internal consistency score and (c) sensitivity score.

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

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

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

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

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

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

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.

3.3.2.2. Questionnaire preference. The results of the questionnaire preference rating (see Fig. 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.

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

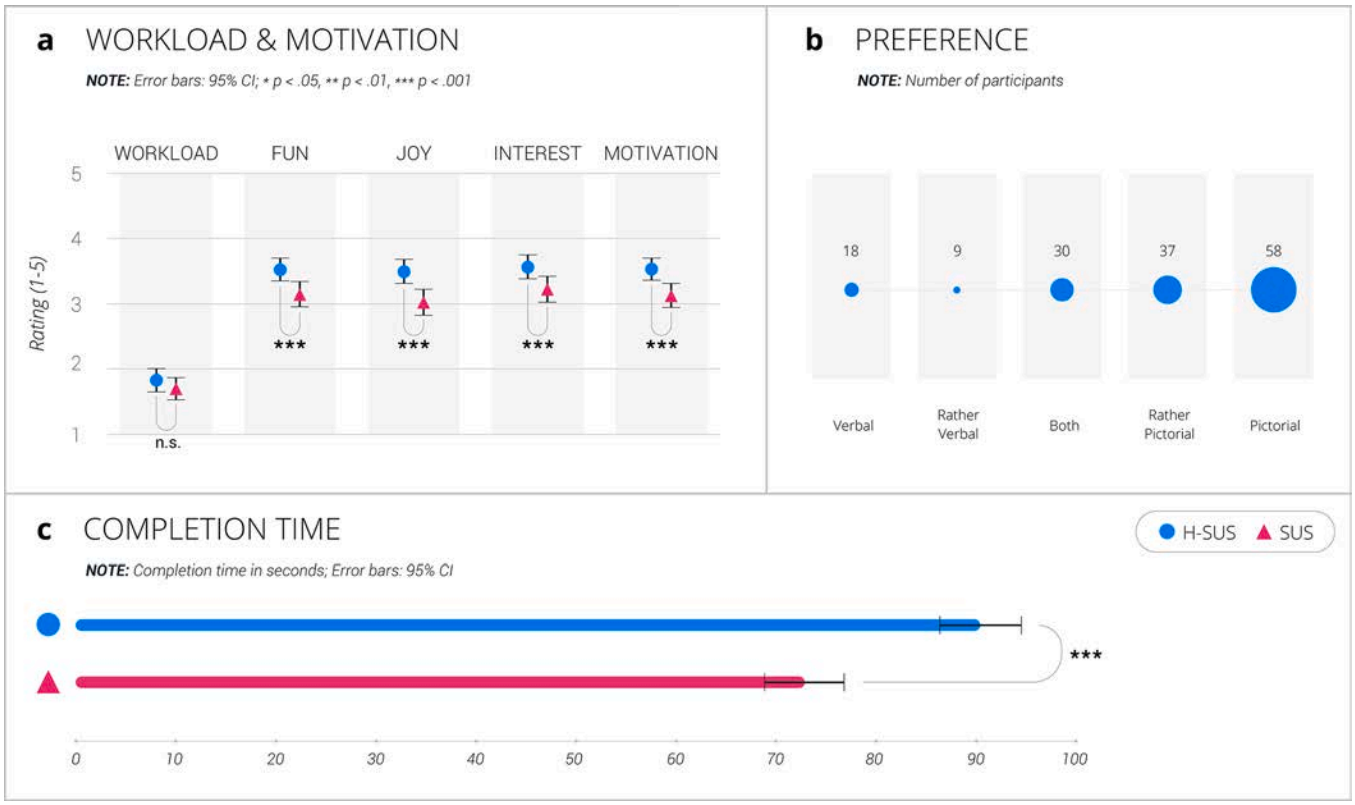


Fig. 4. Comparative analysis of different dimensions of questionnaire experience for H-SUS and SUS: (a) Respondent workload and motivation, (b) preference and (c) questionnaire completion time.

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

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

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

This study aimed to compare H-SUS to the verbal SUS concerning 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 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).

Concerning 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). These effect sizes may be less surprising given that the H-SUS shares many elements with the SUS. However, the high convergent validity score may suggest that one of the 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 concerning 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 ($r \approx -.05$). Particularities of this concept may explain the reason why aesthetics as a measure of divergent validity had a rather high score. 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 and colleagues 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. However, we believe that this finding demonstrates sufficient divergent validity, though it is conceded that future research in the usability domain needs to reconsider the choice of aesthetics as a measure of divergent validity.

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 the 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 and Levy, 1994; Sauro and 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 as 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). Concerning sensitivity, we found for both instruments highly significant differences between low and high-usability condition. Therefore, both instruments are considered sufficiently sensitive to distinguish 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, the 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. We assume that this type of questionnaire was new to most participants (even if a sample item had been given for practice in the beginning). 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 some limitations. The first limitation refers to the test setting. Since the H-SUS was tested in an online study, it was not possible to standardise 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 (such as visual and auditory distractions, and short interruptions). 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 some 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 and Sauro (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. One outcome could be 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 subject is essential 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). The methodological approach also considered the identification of the respondent with the gender of the avatar by allowing them to choose between different options, and a large and heterogeneous sample (comprising students, professionals and pensioners).

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 positively and help to gain more confidence when choosing an appropriate instrument.

CRediT authorship contribution statement

Juergen Baumgartner: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Nicole Ruettgers:** Conceptualization, Investigation, Writing - original draft, Writing - review & editing. **Annigna Hasler:** Conceptualization, Investigation, Writing - original draft. **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.

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 and 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.

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. J. Hum.–Comput. Interact. 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. In: 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. Int. J. Hum.-Comput. Stud. 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. J. Behav. Therapy Exp. Psychiatry 25 (1), 49–59.
Brooke, J., 1996. SUS-A quick and dirty usability scale. Usability Eval. Ind. 189 (194), 4–7.
Brooke, J., 2013. SUS: a retrospective. J. Usability Stud. 8 (2), 29–40.
Chu, S., 2003. Cross-cultural comparison of the perception of symbols. J. Vis. Lit. 23 (1), 69–80.
Cohen, J., 1988. The effect size. Stat. Power Anal. Behav. Sci. 77–83.
Coolican, H., 2017. Research Methods And Statistics in Psychology. Psychology Press.
Deci, E.L., Ryan, R.M., 2003. Intrinsic motivation inventory. Self-Determ. Theory 267.
Desmet, P., 2003. Measuring emotion: development and application of an instrument to measure emotional responses to products. Funology. Springer, pp. 111–123.
DeVellis, R.F., 2016. Scale Development: Theory and Applications, 26. Sage Publications.
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.
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. Commun. 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. Adv. Hum.-Comput. Interact. 2014, 1–13. https://doi.org/10.1155/2014/946239.
Harley, A., 2014. Icon Usability. July 27. 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. J. Manage. 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. Org. Sci. 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. Psychol. Trauma 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. Educ. Media Int. 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. Comput. Hum. Behav. 68, 170–179.
Kortum, P.T., Bangor, A., 2013. Usability ratings for everyday products measured with the System Usability Scale. Int. J. Hum.-Comput. Interact. 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. Interact. Comput. 23 (5), 473–483.
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., 2018. The system usability scale: past, present, and future. Int. J. Hum.–Comput. Interact. 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. J. Usability Stud. 13 (1), 38–46.
Messick, S., 1979. Test validity and the ethics of assessment. ETS Res. Rep. Ser. 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. Behav. Inf. Technol. 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. 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. Personal. Soc. Psychol. Bull. 27 (2), 151–161.
Rummel, B., 2015. System Usability Scale – jetzt auch auf Deutsch. January 12. SAP User Experience Community. https://experience.sap.com/skillup/system-usability-scale-jetzt-auch-auf-deutsch/.
Sauer, J., Baumgartner, J., Frei, N., & Sonderegger, A. (2020). Pictorial scales in research and practice: a review. Eur. Psychol.
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. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1609–1618.
Sonderegger, A., Heyden, K., Chavaillaz, A., Sauer, J., 2016. AniSAM & AniAvatar: animated visualizations of affective states. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 4828–4837.
Sonderegger, A., Uebelbacher, A., Sauer, J., 2019. The UX construct–does the usage context influence the outcome of user experience evaluations?. In: IFIP Conference on Human-Computer Interaction, pp. 140–157.
Thielsch, M.T., Lenzner, T., Melles, T., 2012. Wie gestalte ich gute items und interviewfragen. Praxis Der Wirtschaftspsychologie II 221–240.
Tractinsky, N., Katz, A.S., Ikar, D., 2000. What is beautiful is usable. Interact. Comput. 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. Comput. Hum. Behav. 28 (5), 1596–1607.
Wanous, J.P., Reichers, A.E., Hudy, M.J., 1997. Overall job satisfaction: How good are single-item measures? J. Appl. Psychol. 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. Behav. Inf. Technol. 18 (2), 68–82.
Wilde, M., Bätz, K., Kovaleva, A., Urhahne, D., 2009. Überprüfung einer Kurzskala intrinsischer Motivation (KIM). Z. 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.

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

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

International Journal of Human - Computer Studies

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

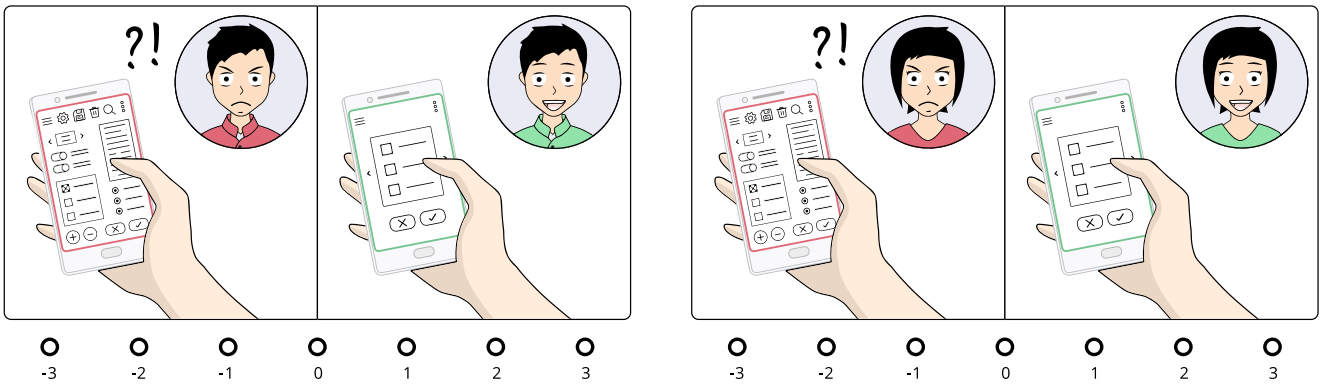


Fig. 1. Example of PUI item with male and female avatar referring to interface complexity of a smartphone.

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

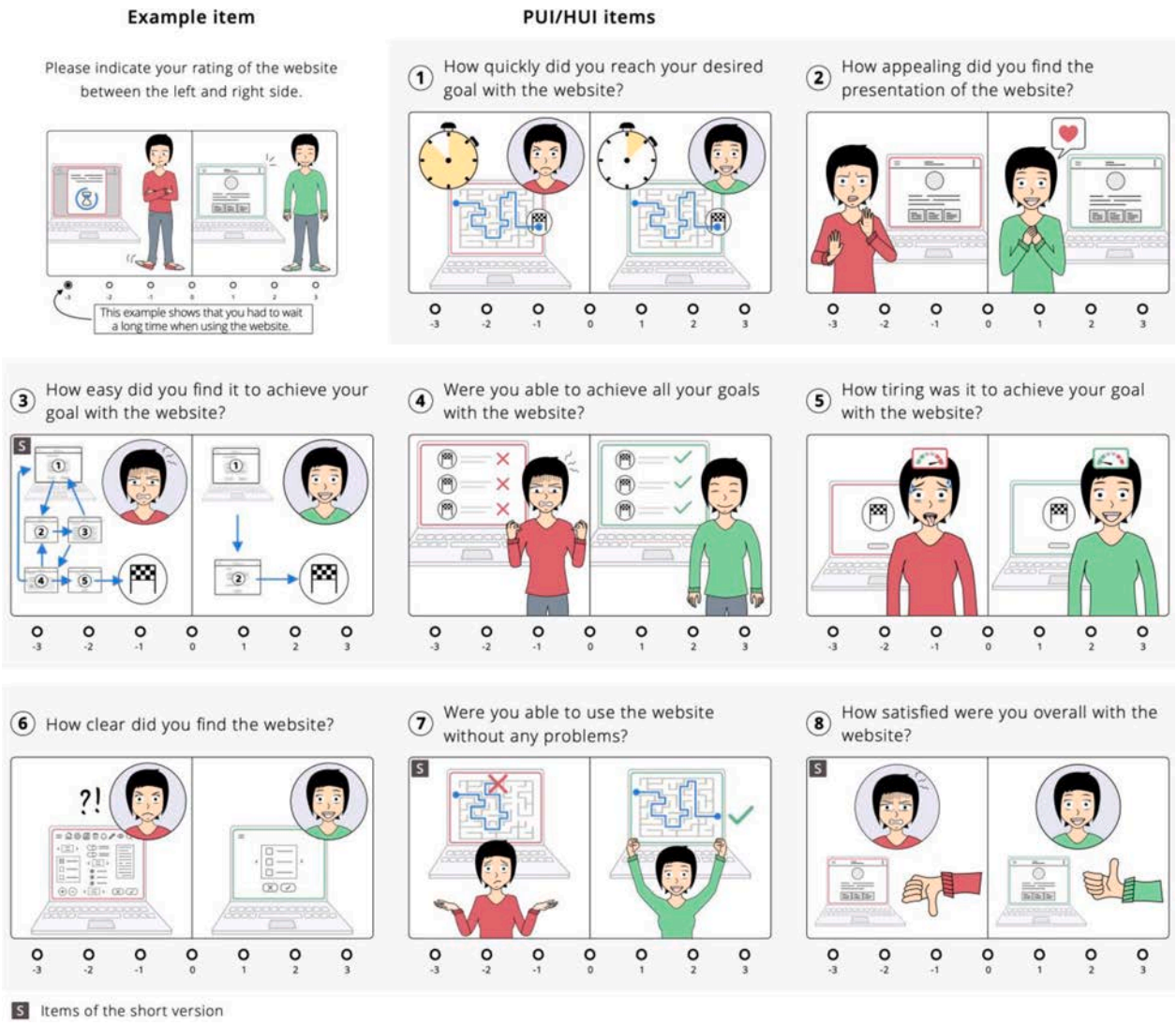


Fig. 2. Example item and complete set of items of the Pictorial Usability Inventory (PUI) in a female version. The verbal question was only shown for the hybrid version (HUI). The wording was translated from German to English.

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.	.903
	The questionnaire was entertaining.	
	The questionnaire was interesting.	
Questionnaire comprehension	The questionnaire was comprehensible.	.871
	The questions were clear.	
	The questionnaire was easy to fill in.	
Questionnaire workload	The questionnaire was too long.	.738
	The questionnaire was complicated.	
	The questionnaire was tedious to fill in.	
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.

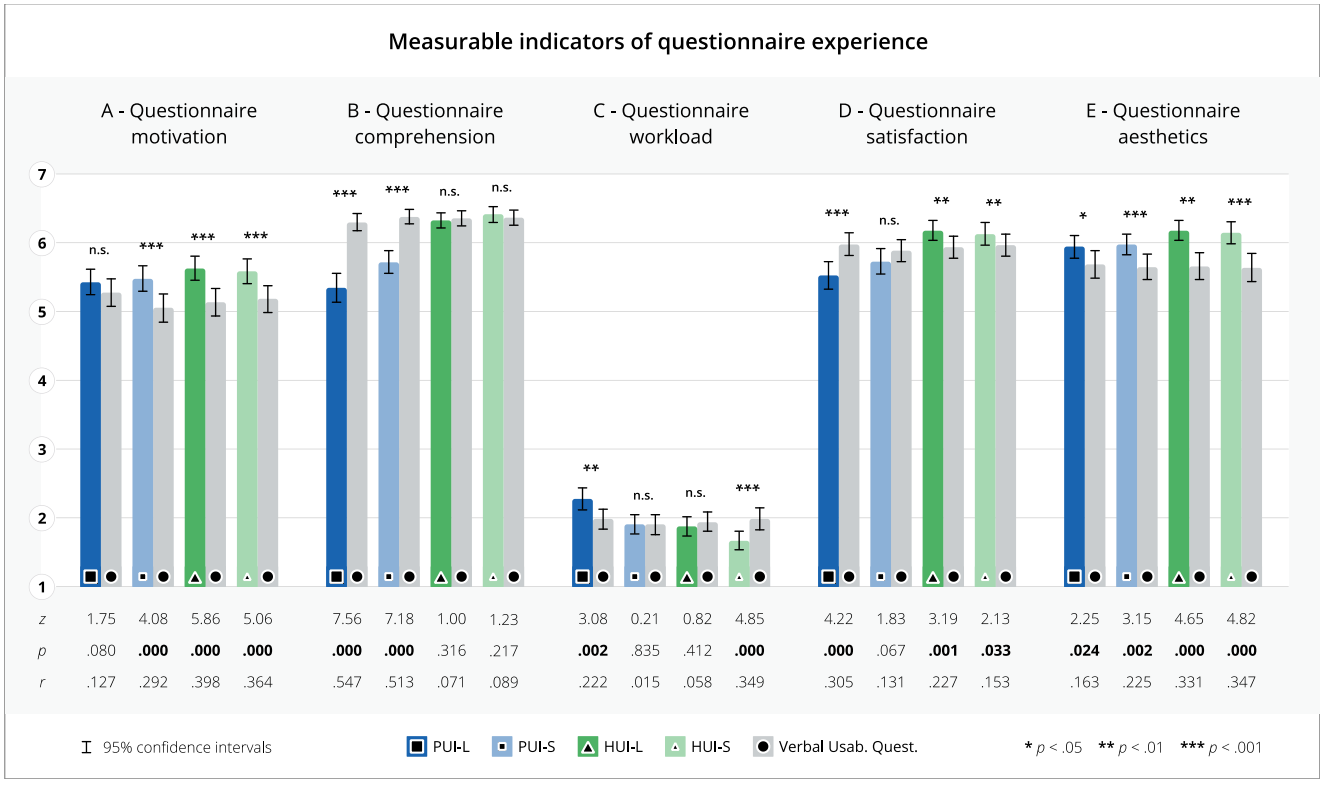


Fig. 3. Overview of QXQ indicators, including statistical parameters of Wilcoxon test between PUI-L, PUI-S, HUI-L, HUI-S and verbal usability questionnaires. Verbal Usability questionnaires comprised SUS, UMUX-LITE and three single-item scales (effectiveness, efficiency, and satisfaction).

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

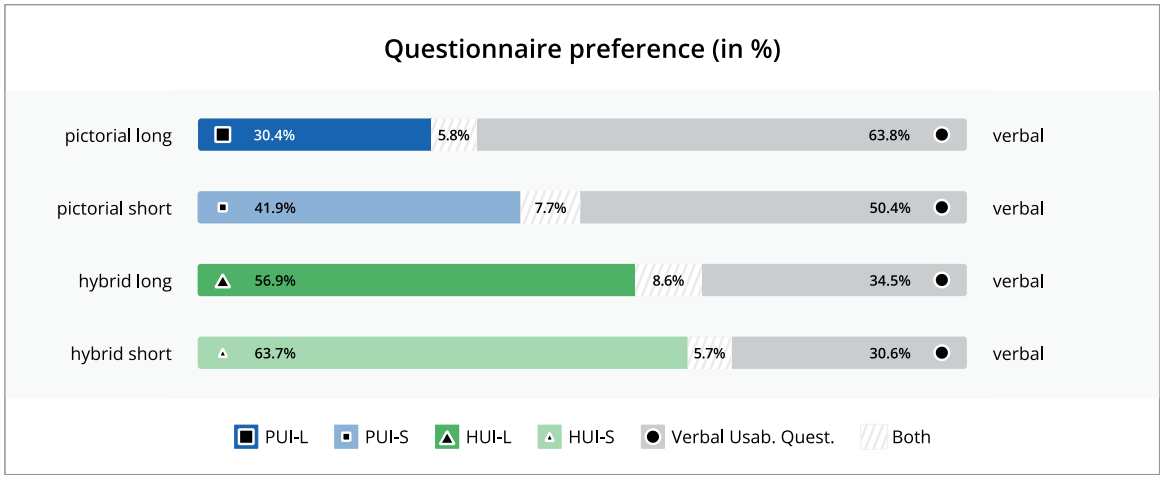


Fig. 4. Overview of questionnaire preference for all PUI versions and the verbal usability questionnaires.

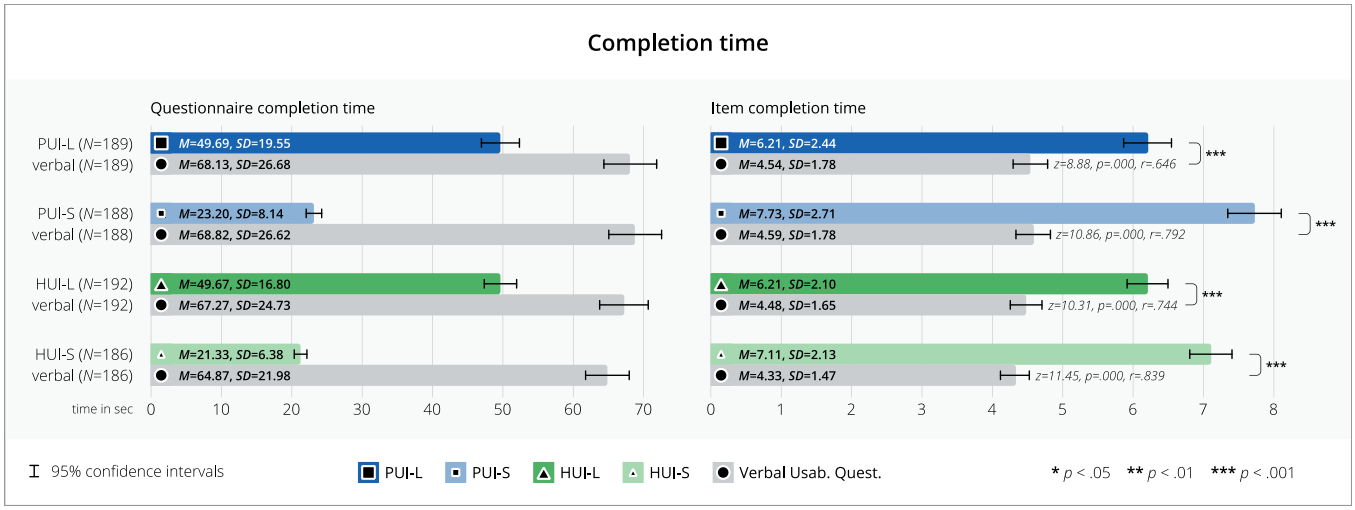


Fig. 5. Overview of the questionnaire and item completion time for all PUI versions and the verbal usability questionnaires.
Notes: Data of N=22 participants (2.83% of the overall sample) were excluded from data analysis since it was identified as outliers (i.e. completion time per item >16s)

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.

6. Study Three – To move or not to – a comparison of static and animated usability scales

To move or not to: a comparison of static and animated usability scales

Juergen Baumgartner[1ab], Andreas Sonderegger[ac], Juergen Sauer[a]

[a] Department of Psychology, University of Fribourg, Rue P.-A.-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, 3005 Bern, Switzerland

Abstract

The Hybrid Usability Inventory (HUI) is a usability questionnaire that uses a combination of pictorial and verbal information to express the meaning of its items. The aim of this study was to extend the static pictorial representation by using animations. Previous research has not yet addressed positive or negative outcomes of animations in questionnaires. We hypothesised that an animated questionnaire would have an additional positive effect on respondents' motivation and preference, without drawbacks on psychometric properties. The goal of the present study was to compare the HUI with an animated version (AniHUI) in an online test setting. Respondent-centred aspects (questionnaire experience) as well as psychometric properties (sensitivity, validity, reliability) were assessed. Participants (N=192) interacted with a website prototype (either high or low usability) and subsequently assessed the website's usability either with HUI or AniHUI, the System Usability Scale (SUS) and further measures of interest. Results suggest that AniHUI did not differ substantially from HUI. However, both static and animated scale were superior to the SUS regarding respondent-centred measures. Findings suggest that the HUI and the AniHUI are motivating and reliable scales that can be used in research and practice.

Keywords: usability; animated scales; hybrid scales; animated questionnaire; questionnaire experience; consumer product evaluation

Highlights

- This study is the first that systematically compares a static and an animated hybrid usability scale regarding respondent-centred aspects (questionnaire experience) and psychometric properties.
- The static and the animated hybrid usability scale achieved psychometric results comparable to the SUS but were rated more favourably on respondent-centred aspects (i.e. motivation, aesthetics and perceived completion time).
- The animated questionnaire did not emerge to be more motivating than the static one, being at the same level as the hybrid questionnaire.

1 Introduction

The presumably most common and economic way of collecting information about individuals is by means of questionnaires. They were introduced in the first half of the 19th century (Gault, 1907) and made ever since a meteoric rise in empirical research and practice. Standardised questionnaires are also popular in the domain of usability evaluation, where they are frequently used during or after usability tests (Sauro & Lewis, 2016). However, the use of

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

verbal questionnaires comes with certain limitations: (1) Only the literate population can answer them (Sonderegger et al., 2016). (2) Validated instruments are often not available in other languages than English, which makes them difficult to use across language barriers (Baumgartner et al., 2020). (3) Participants' motivation might suffer when answering long questionnaires or a battery of multiple questionnaires, leading to inadequate answering behavior such as random answers (Robins et al., 2001). To overcome these limitations, alternative questionnaire types using pictures (pictorial) or a combination of pictures and words (hybrid) have been proposed (Baumgartner et al., 2021, 2023). While the number of established image-based tools is relatively modest, there are even fewer questionnaires that use animations. The scope of this article is to investigate whether there are advantages associated with the use of animated questionnaires and whether they are useful in the context of a usability evaluation. Previous research on pictorial scales has shown that difficulties might appear regarding the reliable and understandable communication of meaning through images alone (Baumgartner et al., 2023). Therefore, it was suggested that animations in hybrid scales could be used for easier communication of specific content (e.g. movement, changes over time, highlighting). Although the idea seems reasonable and understandable, the question arises how this might affect the experience of the respondents and as to what consequences this approach might have on the psychometric properties of the scale.

1.1 Usability evaluation

Usability is defined 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). Being integrated in the overall umbrella construct of user experience (UX, ISO 9241-210, International Organization for Standardization, 2019), usability plays a vital role for practitioners to assess the outcome of the interaction of a user with services and products. This is also reflected in the fact that usability is still routinely assessed in the context of interface development. The user-centred development process is considered to be the gold standard in system design. Prototypes and design variants of an interface are tested at regular intervals with actual users to find out whether they can efficiently and effectively interact with the design and whether the interaction is satisfactory (Gould & Lewis, 1985; Noyes & Baber, 1999; Salah et al., 2014). The method applied in such an iterative design and evaluation procedure is referred to as usability test (Nielsen, 1994). In a usability test various forms of data are recorded. In addition to interview and observational data, the collection of subjective usability-data is common (for more details see Sauer et al., 2020; Sonderegger et al., 2019). These data on subjective experiences are usually collected by means of standardised questionnaires. Over the past 30 years, more than 20 standardised instruments were published assessing usability in different forms (for an overview see Assila et al., 2016).

1.2 Alternative questionnaire types

In recent years, alternative questionnaire types for usability assessment were created, such as pictorial and hybrid usability questionnaires. A pictorial scale may be defined as 'an instrument that makes use of image-based elements to convey the meaning of its items' (Sauer et al., 2020, p.1). A hybrid scale adds verbal elements (i.e. a question or a description) to the image-based elements to convey the underlying meaning (Baumgartner et al., 2021). The rationale to develop and use pictorial scales is to provide users with inclusive access to questionnaires and to facilitate usability evaluation in general. Especially, hybrid scales have proven themselves in past studies as more convenient for participants and were preferred when directly compared with verbal scales (Baumgartner et al., 2021, 2023). There are several advantages related to the use of hybrid questionnaires (as compared to verbal scales), with the

most important being: (1) They provide a concrete visualisation of abstract concepts (e.g. usability) and therefore give the respondent context (e.g. showing a specific usage situation). (2) The visual information is complemented by a verbal statement or a question, which makes it easier for participants to understand the intended meaning (e.g. Ghiassi et al., 2011; Sauer et al., 2020). (3) They stimulate interest, provide pleasure or even joy and therefore increase the respondents‘ motivation to complete this kind of scales (e.g. Desmet, 2003; Haddad et al., 2012). There are also some disadvantages: (1) When completing hybrid scales, participants need normally more time per item compared to using verbal items. In the wake of a growing need for more economic instruments, the number of items needs to be reduced to a reasonable number to compete with verbal instruments. (2) If verbal and pictorial information do not match well, there is the risk of ambiguity to increase. (3) The development process is more complex and time consuming than creating verbal items, and specialist drawing skills are needed to visualize the items (e.g. Desmet et al., 2016). Given the fact that hybrid instruments have promising advantages, but also potential drawbacks, we searched for ways to improve their characteristics. In this work, we considered the inclusion of animations as a promising next step in the evolvement of image-based scales.

1.3 Animated questionnaires

An animation is an illusion of movement created by rapidly displaying a sequence of static images (Harrison & Hummell, 2010). The first film animations became popular in the 19th century and primarily served amusement purposes (Bendazzi, 2015). Besides entertainment, animations are used today in a variety of contexts such as arts, advertising, marketing, but also in learning environments, such as computer animations for medical education (Ruiz et al., 2009). In the context of questionnaire design, an animated scale brings motion into play as an additional element. We therefore define an animated scale as an instrument that uses image-based elements enhanced with motion to convey the meaning of its items. To our knowledge, only few validated questionnaires match the definition of an animated instrument. In emotion research, PREMO (Product Emotion Measurement Tool, Desmet, 2003; Laurans & Desmet, 2017) was created to assess 14 emotions towards a product using an animated hand-drawn avatar and specific sounds for each emotion. Another instrument in this field is the AniSAM (Animated Self Assessment Manikin, Sonderegger et al., 2016), which is a dynamic version of the original SAM (Bradley & Lang, 1994) using animations to express arousal (i.e. a heartbeat with low or high intensity). In the medical field, the Animated Activity Questionnaire (AAQ, Peter et al., 2015) was developed using animated video sequences to assess activity limitations of patients with hip or knee osteoarthritis. A further animated scale was developed by Setty and colleagues (2019) for the assessment of dental anxiety in children. Addressing a similar population, the Computer Face Scale (Gulur et al., 2009) assesses pain and mood using an animated face that ranges from smile to frown.

Several potential disadvantages are related to the use of animated questionnaires: (1) Rebetez and colleagues (2010) argue that animations could have an overwhelming effect on the working memory since change between frames needs to be memorised and processed in order to understand the item’s meaning. (2) Another argument is that not all graphical elements are instantly present but appear in a sequence of time. Participants therefore must wait until the animation ends to have all information ready for subsequent interpretation. This might lead to a longer item completion time. (3) Finally, the creation and implementation of animations in a questionnaire requires a lot of time and effort.
There are also potential advantages of using animated questionnaires. (1) Animations provide more information than a static representation (Tversky et al., 2002). In consequence, item comprehension could be facilitated due to the availability of more detailed information. (2)

They serve well as support for certain representations such as the expression of emotions (Caicedo & Van Beuzekom, 2006), reducing the abstraction level by showing a concrete representation from beginning to end. (3) Animations have the potential to enhance intrinsic motivation (Bülbül & Abdullah, 2021) and were found to be more intuitive and much more enjoyable (Desmet, 2003).

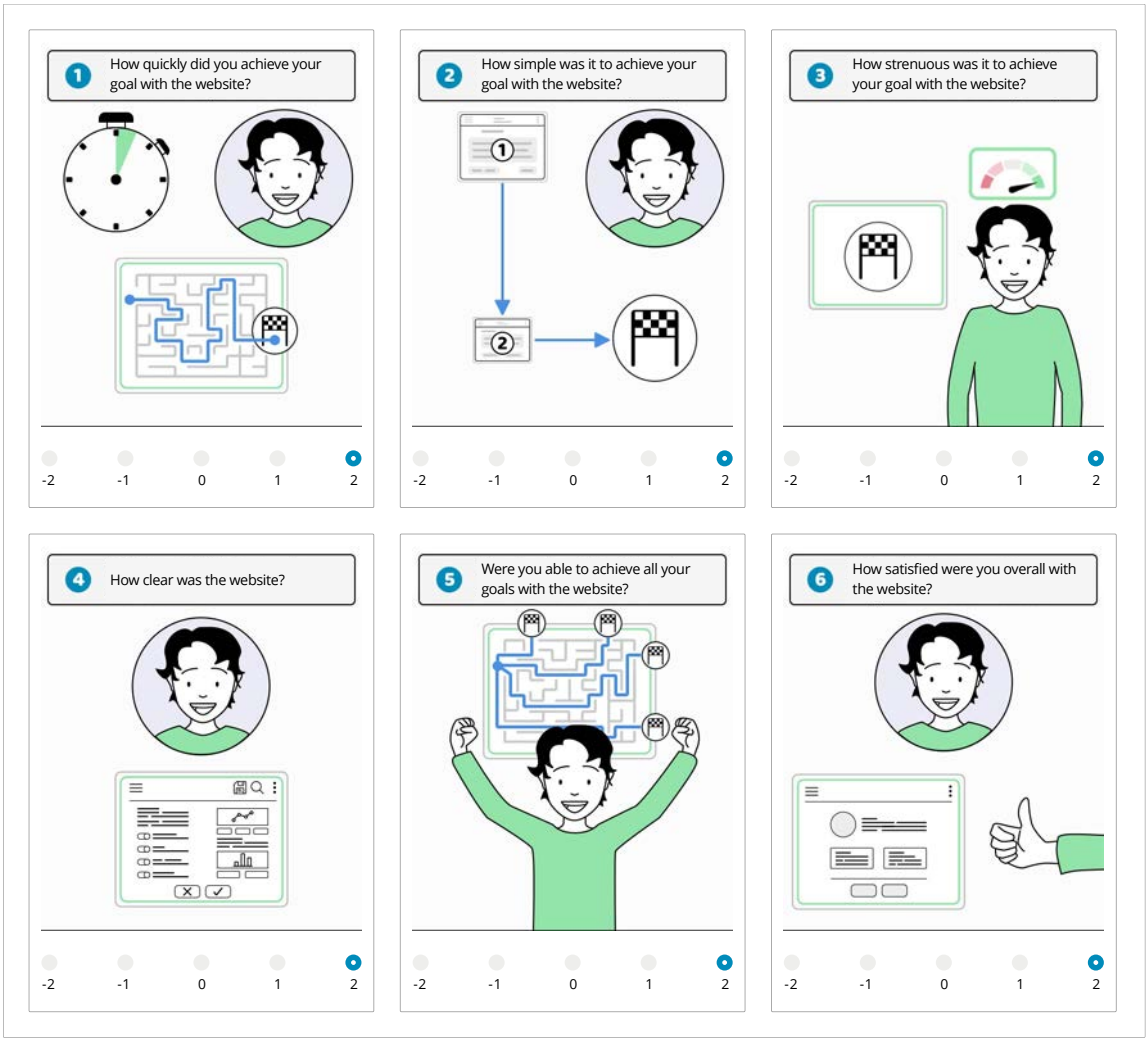
1.4 Questionnaire experience

Questionnaire experience (QX) is a recently introduced concept aiming to capture respondents' subjective experiences when answering to a questionnaire. QX bears some resemblance to the underlying ideas of the concept of user experience (UX) and was defined as a comprehensive experiential process that respondents undergo when completing a questionnaire or a test (Sauer et al., 2020). It is considered an extension to the traditional psychometric properties of a scale with the purpose of providing a more wide-ranging assessment of a given questionnaire (Baumgartner et al., 2021). The assessment of QX offers insights on (1) how engaged the participants were (motivation), (2) how comprehensible the scales were (comprehension), (3) how demanding it was to complete the scales (workload), (4) how satisfied the participants were with the questionnaire (satisfaction), (5) how aesthetically appealing the questionnaire was (aesthetics), and (6) how much time they participants thought they needed to complete the questionnaire (perceived time). Assessing these aspects alongside classical psychometric properties helps to identify experiential issues of instruments. Furthermore, they represent a valuable complement when comparing two or more instruments.

1.5 Development of the Hybrid and Animated Usability Inventory

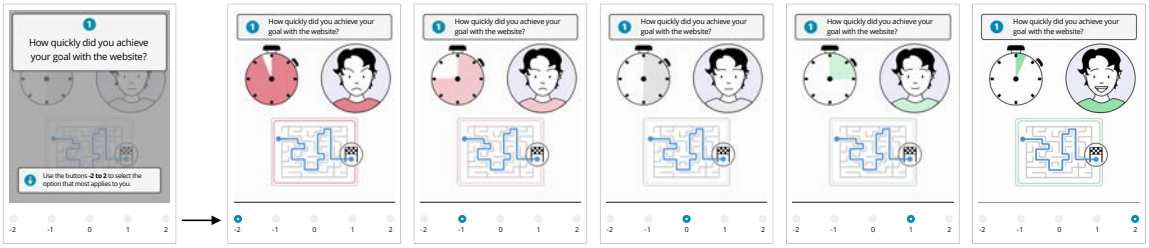
The Hybrid Usability Inventory (HUI) is a so-called hybrid instrument that was developed for the assessment of perceived usability. It consists of a verbal question (e.g. 'How quickly did you achieve your goal with the website?') and a pictorial information that visually expresses the corresponding answer options. The pictorial content is based on the PUI (Pictorial Usability Inventory, Baumgartner et al., 2020), but uses a subset of the original 12 items to make the instrument more economic and less time consuming (see figure 1). The selection of the six items was based on results of a comprehension test that was conducted for a previous study (cf. Baumgartner et al., 2023). For the present study, the six items with the highest comprehension rates were selected.

Figure 1. HUI items 1-6 with most positive answer option selected.



In contrast to previous versions of the PUI, the answer options were reduced from a 7-point to a 5-point Likert scale, and all answer options are depicted instead of only the extreme points. Consequently, five representations were created for each item, each one representing one of the scale points. Radio buttons with numerical anchors are used for displaying the corresponding answer option. Figure 2 shows the initial display consisting of the question and the five answer options. In addition to the question, a call to action is shown to explain the user the handling of the scale (‘Use the buttons −2 to 2 to select the option that most applies to you’).

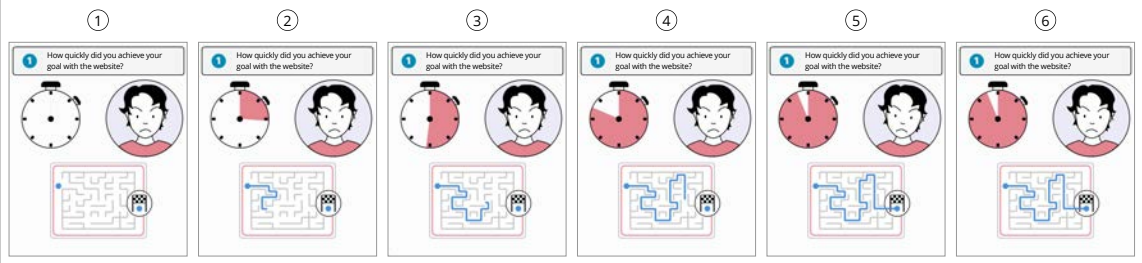
Figure 2. HUI item 1 with initial display for the question and the answer options.



To distinguish adequately between answer options, several design strategies were applied, consisting of (1) a change in the avatar's facial expression (e.g. frowning vs smiling), (2) the use of colours for key elements (i.e. red, grey and green tones), and (3) the application of Weber's law using geometric progression to express change in a given stimulus (e.g. the varying degree the time of the stopwatch is filled; Kunin, 1955). In addition to these design strategies, we designed a gender-fluid avatar in order to overcome binary stereotypes and to avoid the need of implementing two or more gender representations (e.g. Ku et al., 2005; Sonderegger et al., 2016). Several pilot studies were run with students to develop the gender-fluid avatar.

For the purpose of this study, an enhanced version of HUI was created using animations (AniHUI). The animation consists of a 3-second primary animation representing the main idea of the item by manipulating graphic elements (e.g. complete the path to a goal or counting the time on the stopwatch). Figure 3 shows the animation sequence. The animation is repeated once to make sure that the respondent does not miss any information. A secondary 1-second animation is played when the pictorial representation is in its end state (i.e. after running the main animation twice) and consists of slight movements of the avatar to make it appear alive and to motivate the respondent to complete the rating.

Figure 3. Primary animation sequence of AniHUI item 1, from beginning to end state.



1.6 Aim of the research and hypotheses

The aim of this study was to systematically compare respondent-centred aspects and psychometric properties of the HUI with an animated version of the same instrument (AniHUI) in an online test setting (i.e. a usability test of a website). The primary goal consisted of gaining insights of whether AniHUI would have benefits on an experiential level (e.g. motivation, preference) and whether psychometric properties were acceptable. The System Usability Scale (SUS, Brooke, 1996) was used as an additional measure of comparison of which questionnaire experience and psychometrics were assessed as well. The secondary goal consisted of testing a gender-fluid representation of the avatar.

In general, we expect HUI having psychometrics close to those of the SUS. A previous study (Baumgartner et al., 2023) with the preceding version of HUI showed very similar results for sensitivity and high coefficients of convergent validity (r=.773). Furthermore, we do not expect considerable differences between HUI and AniHUI since both scales use the same pictorial and verbal content. Instead, we expect differences rather on an experiential level. Therefore, we put the following hypotheses forward for questionnaire experience:

H1: *Higher motivation and stronger preferences for HUI and AniHUI compared to SUS, with AniHUI having the highest ratings (AniHUI>HUI>SUS).* Previous studies (Baumgartner et al., 2021, 2023) showed increased motivation ratings for the hybrid questionnaire type and the majority of participants preferred a hybrid questionnaire over a verbal one. Furthermore we

assume that the animated version gives a further motivation boost because of the inclusion of motion, which makes the questionnaire more vivid and pleasant to interact with (Bülbül & Abdullah, 2021).

H2: *Higher objective item completion time for HUI and AniHUI compared to SUS, with AniHUI having the longest completion time (AniHUI>HUI>SUS).* Two previous studies (Baumgartner et al., 2021, 2023) demonstrated that completion time of hybrid items are in general longer than verbal items, due to the additional pictorial information that has to be processed. We assume that item completion times are even longer for the animated versions since the animation must be played before giving a rating. The SUS is considered having the shortest item completion time, since only verbal content is showed.

H3: *Lower subjective questionnaire completion time for HUI and AniHUI compared to SUS, with AniHUI having the lowest questionnaire completion time (AniHUI<HUI<SUS).* We assume that time perception is biased when completing the animated and hybrid questionnaire. There is evidence from motivation and flow research that intrinsically motivated participants have a tendency of losing track of time when engaged in a pleasant or motivating activity (Conti, 2001; Nakamura & Csikszentmihalyi, 2014). Since we expect completing the animated questionnaire as an activity that is pleasant, we assume that time flies faster for the participants during questionnaire completion. We expect a similar effect happening for the hybrid questionnaire, but to a lesser extent.

2 Method

2.1 Participants

Participants were recruited by an email invitation sent to bachelor's and master's students of various fields of study at the University of Fribourg. Moreover, the study was advertised on the webpage of the Psychology Department. Ten gift vouchers (each 20 CHF) were raffled to increase participation. The study was conducted in German language. The sample consisted of 192 participants (75.5% female, 24.5% male) with their ages ranging from 17 to 84 years (M=25.76 , SD=8.64). Amongst the participants were 149 students (77.6%), 33 employees (17.2%) and 10 persons which did not report their professional status (5.2%). Two participants (≈1%) reported having some form of colour blindness. Participants rated their experience with websites in general above midscale (M=5.55, SD=1.11) on a 7-point Likert scale ranging from 1 (very low) to 7 (very high). Thirty-six participants (18.8%) indicated that they had seen the website before.

2.2 Website prototype and user tasks

In the present study, participants interacted with a website of a fictitious leisure centre which was manipulated in terms of usability (low vs high). The manipulation consisted of several violations of usability heuristics (Nielsen & Molich, 1990) such as excessively long delays when loading pages, or inappropriate form design. The same website was already used in a previous study (Baumgartner et al., 2023) in which the manipulation of usability proved to be successful. In contrast to the previous study, participants completed only two instead of three tasks to minimise study completion time and dropout rate. The two tasks consisted of (1) finding the opening hours of a specific sauna and (2) buying an annual subscription for the leisure centre. Participants were able to freely navigate on the webpage to solve the tasks. Furthermore, they were instructed to move to the next task in case they could not find the solution within four minutes.

2.3 Measures and instruments

The measures and instruments used in this study are divided in respondent-centred measures and psychometric ones. Respondent-centred measures involve aspects of QX (motivation, comprehension, etc.), preference, and questionnaire completion time. Psychometric measures consist of sensitivity, measures of convergent validity and internal consistency. The measures and instruments are described in the following sections in more detail.

2.3.1 Respondent-centred measures

To assess respondent-centred aspects of the usability questionnaires, the Questionnaire Experience Questionnaire (QXQ; Baumgartner et al., 2023) was presented after completion of the hybrid or animated scale and SUS. The QXQ consists of three multi-item scales assessing motivation, comprehension, and workload. Two single-item scales are used to measure satisfaction and aesthetics. The scales are rated on a seven-point Likert scale (1=totally disagree, 7=totally agree). In addition, a single-item scale for perceived questionnaire completion time was used in this study (1=very little time, 7=very much time). The QXQ was already used in a previous study (Baumgartner et al., 2023) with a large sample (*N*=777) in which the multi-item scales obtained acceptable to excellent reliability scores. Table 1 shows the wording of the scales and Cronbach's alpha values.

Table 1. Items of the questionnaire experience questionnaire (QXQ) and Cronbach's alpha values for multi-item scales (based on Baumgartner et al., 2023). The wording was translated from German to English.

Measurable indicator	Item	Cronbach's alpha
Questionnaire motivation	The questionnaire was fun.	.903
	The questionnaire was entertaining.	
	The questionnaire was interesting.	
Questionnaire comprehension	The questionnaire was comprehensible.	.871
	The questions were clear.	
	The questionnaire was easy to fill in.	
Questionnaire workload	The questionnaire was too long.	.738
	The questionnaire was complicated.	
	The questionnaire was tedious to fill in.	
Questionnaire satisfaction	Overall, I was satisfied with the questionnaire.	-
Questionnaire aesthetics	The questionnaire had an appealing design.	-
Questionnaire completion time	How much time did it take you to complete the questionnaire?	-

Moreover, respondents' questionnaire preference was assessed by using a bipolar seven-point Likert scale (1=verbal questionnaire, 7=image-based questionnaire) and questionnaire completion time in seconds was recorded by the online survey tool.

2.3.2 Sensitivity

Sensitivity refers to the ability of distinguishing between different levels of usability (Lewis, 2002). For an instrument being highly sensitive, large differences in usability scores are expected when websites are evaluated that vary regarding their design (e.g. a well-designed webpage is compared with an ill-designed webpage). In this study, sensitivity was assessed by comparing scores of the various scales assessing a well-designed or an ill-designed webpage.

2.3.3 Convergent validity

Convergent validity refers to the idea that when two independent instruments measure the same construct, high correlations between them are to be expected (Messick, 1979). As main convergent measure for this study, the System Usability Scale (Brooke, 1996) was chosen, a ten-item verbal scale that is answered with a 5-point Likert scale (1=strongly disagree,

5=strongly agree). The SUS is a prominent and frequently used instrument in the field of usability evaluation, with translations in various languages and good psychometric properties (for an overview see Lewis, 2018). For easier interpretation of scores, Sauro and Lewis (2016) have introduced a grading system, ranging from ‘A’ to ‘F’. For this study, a validated German version of the SUS was used (Gao et al., 2020).

In addition to the SUS, a self-created single-item scale for overall satisfaction was used (‘Overall, I was satisfied with this website.’). The scale was rated on a 7-point Likert scale (1=strongly disagree, 7=strongly agree).

2.3.4 Internal consistency

Internal consistency is one measure of reliability and estimates how good the items of a questionnaire relate to each other (Coolican, 2017). When a questionnaire is assumed to measure a one-dimensional construct, internal consistency is expected to be high. Internal consistency of HUI, AniHUI and SUS was assessed calculating Cronbach’s alpha.

2.3.5 Related variables to the avatar’s gender

At the end of the study, two items were used to assess gender-related perception of the avatar in the HUI and AniHUI. The first item asked for the gender the participant would attribute to the avatar, using a 7-point Likert scale. The adjective anchors ‘very male’ (left extreme) and ‘very female’ (right extreme) represented the extreme values, and ‘neutral’ was used as middle category. The second item asked how important it is to the participant that the avatar represents the participant’s own gender. A 7-point Likert scale was used with adjective anchors ‘not at all important’ (left extreme) and ‘very important’ (right extreme).

2.4 Experimental design

A two-factorial between-subjects design was employed in this study, with questionnaire type as the first independent variable (AniHUI vs HUI), and system usability as second independent variable (low vs high). The latter permitted to estimate sensitivity.

2.5 Procedure

Participants who clicked on the link in the study invitation were redirected to an online questionnaire where they received information about the study procedure and data privacy. After giving informed consent and completing a page with initial questions (demographics, website experience), participants were randomly assigned and redirected to either the low or the high usability version of the website of the fictitious leisure centre. They were asked to solve two tasks using the website. After interacting with the website, participants were redirected to the online questionnaire, where they had to indicate how many tasks they could complete and whether they had already known the webpage or not. On the subsequent pages, participants completed the post-test usability questionnaires, consisting of either HUI or AniHUI, and SUS. To prevent order effects, the sequence of presenting hybrid and verbal usability questionnaires was counterbalanced (i.e. half of participants completed HUI/AniHUI first, the other half SUS first). After each usability questionnaire, respondent-centred measures were assessed using the QXQ. On the last pages, participants were asked how they perceived the avatar (i.e. gender evaluation of the avatar, importance of gender representation), which post-test usability questionnaire they preferred most (HUI/AniHUI, SUS), if they completed the questionnaire seriously and whether they want to participate in the raffle. Finally, they were thanked for their participation.

2.6 Inclusion criteria and data treatment

The following criteria were used to include data sets for the analysis: (1) participants with complete data sets, (2) participants without multiple study participation, (3) participants that responded 'yes' to the question whether they completed the study seriously. Out of a total of 243 participants, 192 participants were included for data analysis according to these criteria.

Non-parametric tests were used for data analysis in case requirements for normal distribution and homogeneity of variance were not met. The following analyses were made: Comparisons of group means to determine sensitivity and respondent-centred measures (Mann-Whitney U-test, Wilcoxon signed-rank test), correlational analyses for convergent measures (Spearman's rank correlation), calculation of internal consistency (Cronbach's alpha), and frequency analyses for questionnaire preference and avatar-related analyses (descriptive percentages). The significance level for all analyses was set to 5%.

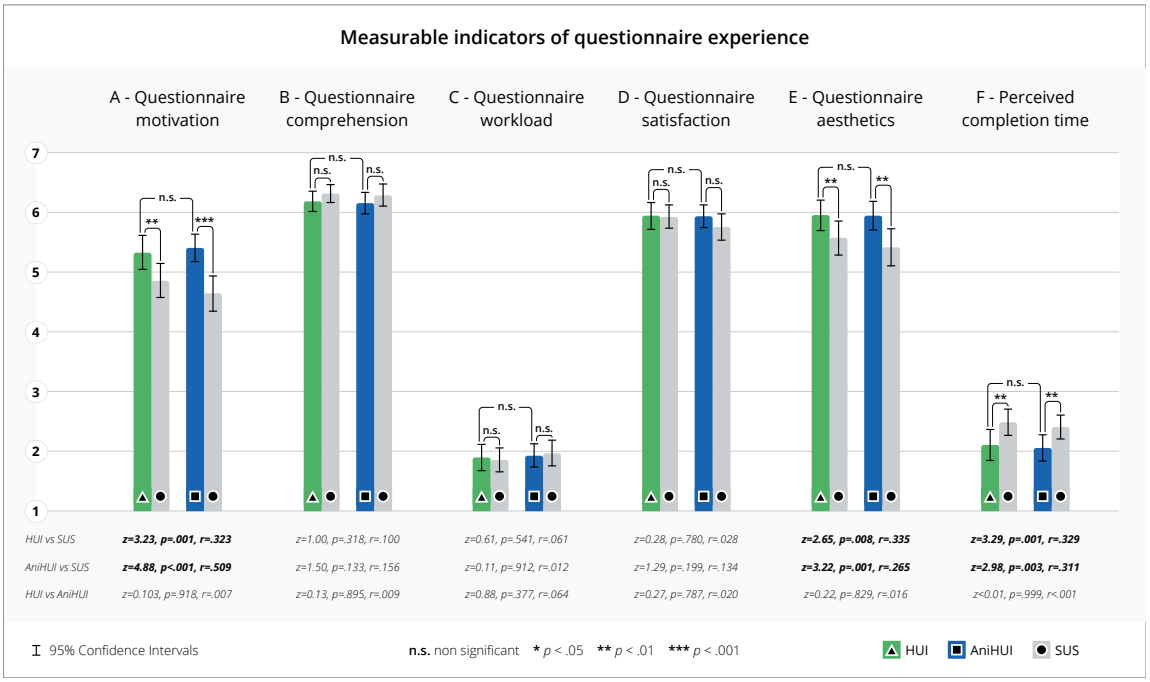
3 Results

3.1 Analysis of respondent-centred measures

3.1.1 QXQ

Wilcoxon tests for all six measurable indicators of the QXQ were conducted to identify differences of respondent-centred aspects between HUI and SUS, and between AniHUI and SUS (within-subjects comparisons). In addition, Mann-Whitney U-tests were conducted to test whether there are significant differences between HUI and AniHUI (between-subjects comparisons). Figure 4 gives an overview of the results.

Figure 4. Overview of QXQ indicators, including statistical parameters of Wilcoxon test (HUI vs SUS, AniHUI vs SUS) and Mann-Whitney U-test (HUI vs AniHUI).

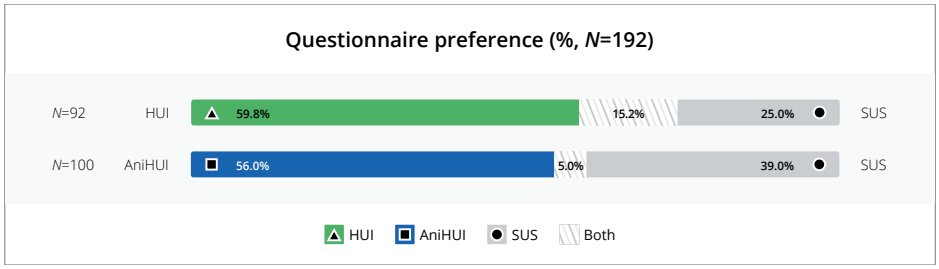


The analysis of the within-subjects comparisons showed significant differences on questionnaire motivation, questionnaire aesthetics and perceived completion time. This effect pattern emerged for HUI and AniHUI in a similar way. They both were rated higher in motivation, were perceived as more aesthetically pleasing and less time consuming than their verbal counterpart SUS. With regard to questionnaire comprehension, questionnaire workload and questionnaire satisfaction, no significant differences were found (all p>.05). The analysis of the between-subjects comparisons showed no significant difference for any of the respondent-centred aspects (all p>.05).

3.1.2 Preference

The results of the questionnaire preference are presented in figure 5. The analysis showed that a majority of participants preferred the HUI (59.8%) over the SUS (25.0%). The AniHUI was also preferred by most participants (56.0%) compared to the SUS (39.0%).

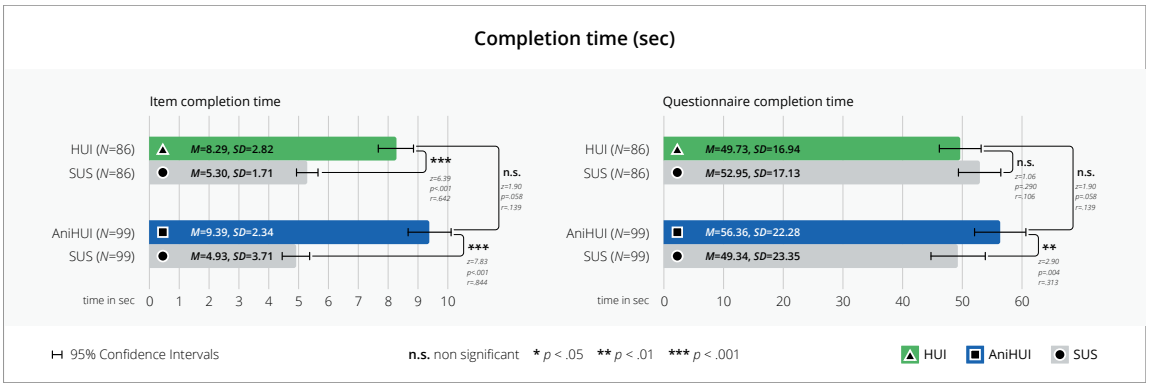
Figure 5. Overview of questionnaire preference for HUI, AniHUI and SUS.



3.1.3 Completion time

The analysis of completion time is illustrated in figure 6. The results for average item completion time show large significant differences between HUI and SUS, and between AniHUI and SUS (all p < .001). No significant difference was found between HUI and AniHUI (p>.05).

Figure 6. Overview of item and questionnaire completion time for HUI, AniHUI and SUS.



Notes: Data of N=7 participants (3.76% of overall sample) were excluded from data analysis, since it was identified as outliers.

Regarding questionnaire completion time, no significant difference was spotted between HUI and SUS (p>.05). However, a significant difference was found between AniHUI and SUS, with AniHUI requiring on average 7 seconds longer to process than SUS (p<.01). However, no significant difference was obtained between HUI and AniHUI (p>.05).

3.2 Analysis of psychometric properties

3.2.1 Sensitivity

Mann-Whitney U-tests were carried out to assess the difference between low and high usability for HUI, AniHUI and SUS. The analysis showed significant differences for all instruments (cf. table 2). All usability questionnaires were highly sensitive, distinguishing well between low and high-usability condition, with AniHUI showing a large effect size (r=.500), and HUI and SUS showing medium effect sizes (r≈.370).

Table 2. Sensitivity of HUI, AniHUI and SUS as a function of usability levels, including means, grades, and statistical parameters of Mann-Whitney U-test.

	Low usability *M (SD),* grade	High usability *M (SD),* grade	*U*	*z*	*p*	*r*
HUI (*N*=92)	70.02 (19.87), C	85.42 (15.26), A+	593.50	3.64	<.001***	.379
SUS (*N*=92)	70.57 (20.04), C	84.15 (13.58), A+	613.00	3.47	<.001***	.362
AniHUI (*N*=100)	72.28 (16.17), C+	87.01 (15.58), A+	528.50	5.00	<.001***	.500
SUS (*N*=100)	69.85 (18.73), C	82.70 (14.12), A	705.50	3.76	<.001***	.376

Notes: Grades range from 'A' to 'F' (cf. Sauro & Lewis, 2016); * $p < .05$; ** $p < .01$; *** $p < .001$

3.2.2 Convergent validity

To determine convergent validity, correlations were calculated (cf. table 3). The analysis showed a strong correlation between HUI and SUS, and a slightly lower correlation between AniHUI and SUS. Comparing the two correlations using Fisher's Z indicates a small effect (Cohen's q=0.156). The correlation with the single-item scale for satisfaction was similarly high for HUI and SUS, and again slightly lower for AniHUI.

Table 3. Correlations between HUI, AniHUI and measures of convergent validity (SUS, single item for satisfaction).

	SUS	Satisfaction (single item)
HUI (*N*=92)	.827.***	.763***
SUS (*N*=92)	-	.801***
AniHUI (*N*=100)	.771***	.642***
SUS (*N*=100)	-	.765***

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

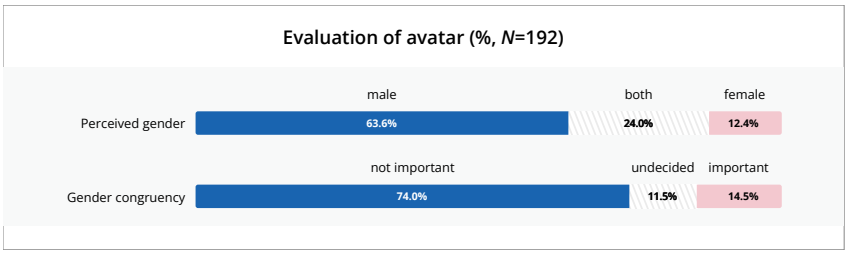
3.2.3 Internal consistency

For the analysis of internal consistency, all items of the respective questionnaire were used. The results showed good Cronbach alpha values for HUI (α=.827), AniHUI (α=.814) and SUS (α=.886).

3.3 Evaluation of avatar

The evaluation of how the participants perceive the gender of the avatar is shown in figure 7. The results show that almost two thirds of the participants perceive the avatar as male, slightly more than 10% see it as female and only a quarter perceives it as both male and female. When asked if it is important to present an avatar with the same gender as the respondent, three quarters of the participants do not think gender congruence is important and 15% think it is important, with about 10% being undecided.

Figure 7. Evaluation of avatar's gender and importance of gender-congruent representation in percentages.



4 Discussion

This study compared systematically a static hybrid usability questionnaire (HUI) with an animated hybrid questionnaire (AniHUI), focusing on respondent-centred aspects of questionnaire experience and psychometric properties. In addition, both instruments were compared with a standardised instrument that measures perceived usability (i.e. the SUS). Findings indicate that respondent-centred aspects were very similar for HUI and AniHUI, with both having advantages on motivation, aesthetic appeal, and perceived completion time over the SUS. Moreover, static and animated questionnaire obtained fairly similar results regarding psychometric properties (i.e. high sensitivity, high convergent validity, and good internal consistency).

With regard to respondent-centred measures, we assumed in our first hypothesis (H1) that motivation and preference were highest for AniHUI, followed by HUI and SUS (AniHUI>HUI>SUS). Results indicated that HUI and AniHUI obtained considerably higher motivation ratings compared to SUS. Although we expected the AniHUI to be more motivating than the HUI, no such effect was observed. The same holds true for questionnaire preference, which was clearly higher for HUI and AniHUI compared to the SUS but did not differ between much them (HUI: 59.8%; AniHUI: 56.0%). Therefore, findings are partially in line with H1 since no clear advantage of the animated questionnaire over the static one could be found. One explanation might lie in the animation itself. Comics or cartoons often use exaggeration as mechanism to convey the intended meaning and to create an entertaining experience (Eisner, 1985). It could be that the animations were too subtle to promote a more enjoyable experience. However, it must also be mentioned in this context that using too much exaggeration might risk to bias the rating (Reynolds-Keefer et al., 2011). Another explanation could be that the animations alone have a similar impact as the static pictures, because they lack of an auditive supplement that emphasises the animated content such as sound effects (Bülbül & Abdullah, 2021). Other instruments such as the PREMO use sounds that correspond to the emotion the avatar represents (Caicedo & Desmet, 2009; Desmet, 2003). Hence, it is possible that additional auditory stimuli would lead to an even more positive evaluation of the AniHUI in terms of participant motivation.

Our second hypothesis (H2) stated that HUI and AniHUI would require increased item completion times compared to SUS, with AniHUI requiring the most time to be completed (AniHUI>HUI>SUS). In line with our hypothesis, results showed that verbal items were completed the fastest (≈5s). However, no significant difference was found between HUI (≈8s) and AniHUI (≈9s), although results are pointing towards that direction (p=.058). Again, our assumptions were only partially met. We conclude that verbal content is processed faster than hybrid content, and that the additional animation also needs some extra time but does not differ significantly with the hybrid version. Looking at questionnaire completion time, results suggest that HUI and SUS need about the same amount of time to complete (≈50s), and

AniHUI needs a couple of seconds longer (≈56s). Even if HUI and AniHUI have 4 items less than the SUS, both instruments are completed in under one minute on average, which makes them still very time-efficient in administration.

In our third hypothesis (H3) we assumed that subjective time perception is different between hybrid and verbal questionnaires, with the animated questionnaire having the shortest perceived completion time, followed by the hybrid questionnaire, and lastly the verbal questionnaire (AniHUI<HUI< SUS). Results suggest that there is a significant difference between the hybrid and verbal questionnaire in the expected direction, but no difference between the hybrid and the animated questionnaire. Again, our hypothesis is partially in line with our assumptions and underlines that AniHUI and HUI behave very alike. This finding is interesting, since objective completion time for HUI and SUS is about the same (≈50s), and even longer for AniHUI (≈56s), but from the respondents' point of view it is perceived as faster as completing the verbal questionnaire. One explanation might be, that participants are in general more engaged when processing pictorial questionnaires and thereby tend to lose track of time (e.g. Conti, 2001). It does not seem to matter whether the image-based elements are animated or not. Another explanation could be that participants simply used the number of items as argument for comparison and therefore evaluated the instruments with less items as less time-consuming.

With regard to psychometric properties, the analysis of sensitivity between usability conditions revealed a medium effect for HUI and SUS (r≈.370) and a large effect for the AniHUI (r≈.500). In this regard, HUI and SUS behave very similarly, whereas differences using the AniHUI seem to be more pronounced, especially in the high-usability condition. Analyses for convergent validity showed that correlations between HUI and SUS were in general high (r=.827), and correlations with the single-item scale for satisfaction were substantial and in the same range as those with the SUS (r≈.780). This finding is also reflected in the obtained average usability score that is almost the same for HUI and SUS. For the AniHUI, correlations with SUS (r=.771) and the satisfaction scale (r=.642) were of slightly lower magnitude. Finally, internal consistency turned out to be good for HUI (α=.827), AniHUI (α=.814) and SUS (α=.886), indicating the items of the questionnaires relate well to each other. Taken together, analysis of data indicates good psychometric values for HUI that are comparable to an established instrument such as the SUS. Results of the AniHUI are generally somewhat lower. Given that the correlation between AniHUI and SUS indicates a strong agreement between measures (r>.700, e.g. Aron & Aron, 1999), we conclude that there is sufficient evidence that perceived usability is adequately measured. However, we might not dismiss the possibility that the animations impacted the results in some way.

Another finding that is worth mentioning is about the perceived workload when completing a questionnaire. There are concerns mentioned in the literature that animations could have an overwhelming effect on the respondent (e.g. Rebetez et al., 2010). In this study, we did not find any evidence to support this assumption. No significant differences were observed between AniHUI and HUI or SUS concerning relevant respondent-centred aspects such as questionnaire workload or questionnaire comprehension (all p>.05). We therefore conclude that these concerns are unfounded, at least in the context of this study with this particular sample.

The secondary goal of this study consisted of testing a gender-fluid version of the pictorial scales. Despite attempts to design a gender-neutral representation, two thirds of participants evaluated the avatar as male, and only a quarter perceived it as both female and male.

Interestingly, when participants were asked how important the correct gender representation is for them (i.e. whether the gender of the avatar corresponds with the gender of the respondent), almost three quarters of participants reported that it is not important for them. We believe that using a gender-fluid avatar in pictorial questionnaires is a viable way of representing the protagonist in a questionnaire because it removes the need of designing and implementing multiple versions of a scale. Nevertheless, further design iterations with a more stringent evaluation procedure are needed to develop such an avatar.

The present study has some limitations. Three quarters of the participants were students, which means that most participants are highly educated. We assume that students are more efficient at completing questionnaires compared to non-students, and that this might have influenced some of the results (e.g. completion time). Furthermore, roughly a fifth of the participants reported already have seen the webpage. Since we cannot know which version of the website (i.e. low or high usability) they interacted in the preceding study, there is the possibility that the previous interaction shaped their experience somehow. However, we do not believe that this preceding experience had a considerable influence on the results, since the previous study was conducted more than one year before this study.

Future research may look further into the direction whether animated questionnaires coupled with sound effects have a more positive impact on questionnaire experience than silent animated scales. In this context, it would also be important to assess whether the perceived attractiveness of the sound effects might bias the actual rating in some way, leading to a measurement error. Similar concerns have been raised earlier with regard to the attractiveness of pictorial scales (cf. Haddad et al., 2012). Another promising line of research might lie in the idea of determining better which target groups benefit from hybrid or animated scales. There is a list of assumptions concerning favourable conditions for administering pictorial scales to groups such as non-native speakers or people with poor language skills (see Sauer et al., 2020), but no research has yet examined whether usefulness and subjective perception of hybrid scales differ systematically between important demographic variables (age, gender, or other variables of interest).

5 Conclusion

Results of this study imply that AniHUI showed increased motivation compared to a verbal scale (i.e. SUS), but it did not differ considerably from the static scale (i.e. HUI). In fact, most measures assessed in this study showed a pattern very similar to the static scale. Therefore, we conclude that the animated questionnaire – as it was implemented in this study – did not provide additional benefits that are not already covered by the hybrid scale. However, considering the findings of respondent-centred measures and psychometric properties, we suggest for practitioners and scientists alike that both instruments are suitable to assess perceived usability.

6 Acknowledgements

This study was funded by a research grant (No 100019_188808) from the Swiss National Science Foundation (SNSF). Furthermore, we are very grateful to We Are Cube and Puzzle ITC for the support in design and technical matters, to Gaëlle Meyer and Oriane Clerc for the help in scale development and data collection, and to Veronica Solombrino for the numerous design and animation reviews.

7 References

Aron, A., & Aron, E. N. (1999). *Statistics for psychology*. Prentice-Hall, Inc.

Assila, A., De Oliveira, K. M., & Ezzedine, H. (2016). Standardized usability questionnaires: Features and quality focus. *Electronic Journal of Computer Science and Information Technology: eJCIST*, *6*(1), 15–31.

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. *International Journal of Human-Computer Studies*, *147*, 102575.

Baumgartner, J., Sauer, J., & Sonderegger, A. (2020). Pictorial usability inventory (PUI) a pilot study. *Proceedings of the Conference on Mensch Und Computer*, 43–52.

Baumgartner, J., Sonderegger, A., & Sauer, J. (2023). Questionnaire experience of the pictorial usability inventory (PUI) – a comparison of pictorial and hybrid usability scales. *International Journal of Human-Computer Studies*, *179*, 103116. https://doi.org/10.1016/j.ijhcs.2023.103116

Bendazzi, G. (2015). *Animation: A World History: Volume I: Foundations-The Golden Age*. Routledge.

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.

Bülbül, A. H., & Abdullah, K. (2021). Emotional Design of Educational Animations: Effects on Emotion, Learning, Motivation and Interest. *Participatory Educational Research*, *8*(3), 344–355.

Caicedo, D. G., & Desmet, P. M. A. (2009). *Designing the new PrEmo*. Dissertação apresentada à Delft University of Tecnology.

Caicedo, D. G., & Van Beuzekom, M. (2006). How do you feel? *An Assessment of Existing Tools for the Measurement of Emotions and Their Application in Consumer Products Research*.

Conti, R. (2001). Time flies: Investigating the connection between intrinsic motivation and the experience of time. *Journal of Personality*, *69*(1), 1–26. https://doi.org/10.1111/1467-6494.00134

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

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

Desmet, P., Vastenburg, M., & Romero, N. (2016). Mood measurement with Pick-A-Mood: Review of current methods and design of a pictorial self-report scale. *Journal of Design Research*, *14*(3), 241–279.

Eisner, W. (1985). *Theory of Comics and Sequential Art*. F.: Poorhouse press.

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

Gault, R. H. (1907). A history of the questionnaire method of research in psychology. *The Pedagogical Seminary*, *14*(3), 366–383.

Ghiassi, R., Murphy, K., Cummin, A. R., & Partridge, M. R. (2011). Developing a pictorial Epworth Sleepiness Scale. *Thorax*, *66*(2), 97–100. https://doi.org/10.1136/thx.2010.136879

Gould, J. D., & Lewis, C. (1985). Designing for Usability: Key Principles and What Designers Think. *Commun. ACM*, *28*(3), 300–311. https://doi.org/10.1145/3166.3170

Gulur, P., Rodi, S. W., Washington, T. A., Cravero, J. P., Fanciullo, G. J., McHugo, G. J., & Baird, J. C. (2009). Computer Face Scale for measuring pediatric pain and mood. *The*

Journal of Pain, *10*(2), 173–179.
Haddad, S., King, S., Osmond, P., & Heidari, S. (2012). Questionnaire design to determine children's thermal sensation, preference and acceptability in the classroom. *Proceedings-28th International PLEA Conference on Sustainable Architecture+ Urban Design: Opportunities, Limits and Needs-towards an Environmentally Responsible Architecture*.
Harrison, H. L. H., & Hummell, L. J. (2010). Incorporating animation concepts and principles in STEM education. *Technology and Engineering Teacher*, *69*(8), 20.
International Organization for Standardization. (2019). *ISO 9241-210:2019*. ISO. https://www.iso.org/standard/77520.html
Ku, J., Jang, H. J., Kim, K. U., Kim, J. H., Park, S. H., Lee, J. H., Kim, J. J., Kim, I. Y., & Kim, S. I. (2005). Experimental results of affective valence and arousal to avatar's facial expressions. *CyberPsychology & Behavior*, *8*(5), 493–503.
Kunin, T. (1955). The Construction of a New Type of Attitude Measure. *Personnel Psychology*, *8*(1), 65–77. https://doi.org/10.1111/j.1744-6570.1955.tb01189.x
Laurans, G., & Desmet, P. M. (2017). Developing 14 animated characters for non-verbal self-report of categorical emotions. *J. Des. Res*, *15*(3–4), 214–233.
Lewis, J. (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. (2018). The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, *34*(7), 577–590.
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
Nakamura, J., & Csikszentmihalyi, M. (2014). The concept of flow. In *Flow and the foundations of positive psychology* (pp. 239–263). Springer.
Nielsen, J. (1994). *Usability Engineering*. Elsevier.
Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 249–256.
Noyes, J., & Baber, C. (1999). *User-Centred Design of Systems*. Springer Science & Business Media.
Peter, W. F., Loos, M., van den Hoek, J., & Terwee, C. B. (2015). Validation of the Animated Activity Questionnaire (AAQ) for patients with hip and knee osteoarthritis: Comparison to home-recorded videos. *Rheumatology International*, *35*(8), 1399–1408.
Rebetez, C., Bétrancourt, M., Sangin, M., & Dillenbourg, P. (2010). Learning from animation enabled by collaboration. *Instructional Science*, *38*(5), 471–485.
Reynolds-Keefer, L., Johnson, R., & Carolina, S. (2011). Is a picture is worth a thousand words? Creating effective questionnaires with pictures. *Practical Assessment, Research & Evaluation*, *16*(8), 1–7.
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.
Ruiz, J. G., Cook, D. A., & Levinson, A. J. (2009). Computer animations in medical education: A critical literature review. *Medical Education*, *43*(9), 838–846.
Salah, D., Paige, R. F., & Cairns, P. (2014). A Systematic Literature Review for Agile Development Processes and User Centred Design Integration. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 5:1-5:10. https://doi.org/10.1145/2601248.2601276
Sauer, J., Baumgartner, J., Frei, N., & Sonderegger, A. (2020). Pictorial scales in research and practice. *European Psychologist*.
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.
Setty, J. V., Srinivasan, I., Radhakrishna, S., Melwani, A. M., & DR, M. K. (2019). Use of an animated emoji scale as a novel tool for anxiety assessment in children. *Journal of Dental Anesthesia and Pain Medicine*, *19*(4), 227–233.
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? *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part IV 17*, 140–157.
Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, *57*(4), 247–262.

7. Overall discussion

The primary goal of the three studies presented in this work was to explore the benefits and drawbacks of using image-based scales for usability assessment. Psychometric properties and respondent-centred aspects served as criteria to assess scale quality and their strengths and weaknesses. Study one aimed to test a hybrid version of an established usability questionnaire. Study two assessed differences between purely pictorial and hybrid scales and between long and short versions. Study three aimed to investigate whether including animations would have additional benefits over static representations.

7.1. Main results and interpretation

This section integrates the key findings from all three studies to identify overarching patterns with regard to psychometric properties and QX, enabling us to draw conclusions from a broader perspective. One essential finding of this work was that the studies successfully demonstrated that the concept of usability with its core components of effectiveness, efficiency and user satisfaction can be visualised and combined with a Likert scale to measure users' perceived usability of a given interactive system. This fact is reflected in the high levels of psychometric quality obtained by the image-based scales. As table 3 summarises, all instruments showed strong correlations with the SUS ($r \approx .800$), they were highly sensitive to changes in usability conditions, indicated by mainly strong effect sizes ($r \approx .500$), and showed good to excellent internal consistency (all $\alpha > .800$). It is noteworthy that purely pictorial and hybrid scale versions were quite similar in their results, which corroborate findings from previous pilot studies (Baumgartner et al., 2020, 2019a, 2019b). Results are also comparable with the relationship of other instruments with the SUS, for instance, UMUX-LITE correlated similarly high in two studies ($r = .810$, Lewis et al., 2013).

Table 3. Overview of characteristics and psychometric properties of image-based usability scales used in the three studies.

Study	Instrument	Test setting	Item number	*N*	Convergent validity (*r*)	Sensitivity (*r*)	Cronbach Alpha (α)	Reference
One	H-SUS	Online	10	152	.862	.626	.910	Baumgartner et al., 2021
Two	PUI-L (long version)	Online	8	191	.857	.635	.944	Baumgartner et al., 2023a
	PUI-S (short version)		3	196	.784	.559	.875	
	HUI-L (long version)		8	197	.774	.605	.912	
	HUI-S (short version)		3	193	.773	.618	.896	
Three	HUI-M (medium version)	Online	6	92	.827	.379	.827	Baumgartner et al., 2023b
	AniHUI		6	100	.771	.500	.814	

Notes: SUS was used in all studies to measure convergent validity.

Consequently, aggregated usability scores between image-based scales and SUS were similar in most cases. This fact might not be surprising for the H-SUS since it used the same verbal components as the SUS. However, it is noteworthy for the other image-based scales (i.e. PUI and HUI versions) since their items were derived from different usability questionnaires (for a detailed overview, see Baumgartner et al., 2020). The interpretation of the effect sizes (Cohen, 1988) outlined in table 4 shows no effects or only small ones (all d<0.400), suggesting that no fundamentally different usability experience was measured with one instrument or the other. One can even argue that the CGS (curved grading scale, Sauro & Lewis, 2016) could be used for the image-based scales to interpret scores using grades. Considering all psychometric findings, the results of these three studies suggest that image-based scales provide a robust and sound measurement of perceived usability.

Table 4. Overview of instruments' usability scores of the three studies as a function of usability condition, including effect sizes for differences between image-based instruments and SUS.

Study	Instrument	*N*	Low usability, score (0-100) *M (SD)*	Cohen's *d*	High usability, score (0-100) *M (SD)*	Cohen's *d*
One	H-SUS	152	63.09 (21.68)	0.018	89.31 (10.28)	0.147
	SUS		62.70 (20.88)		87.86 (9.48)	
Two	PUI-L	191	63.85 (22.05)	0.063	89.97 (9.67)	0.072
	SUS		65.21 (20.83)		89.30 (9.01)	
	PUI-S	196	62.77 (22.52)	**0.330***	86.60 (14.74)	0.069
	SUS		69.71 (19.47)		85.66 (12.53)	
	HUI-L	197	65.65 (22.93)	0.072	90.57 (9.21)	**0.383***
	SUS		67.30 (23.04)		86.88 (10.03)	
	HUI-S	193	63.83 (24.08)	**0.220***	89.29 (13.90)	**0.226***
	SUS		68.75 (20.58)		86.24 (13.08)	
Three	HUI-M	92	70.02 (19.87)	0.028	85.42 (15.26)	0.088
	SUS		70.57 (20.04)		84.15 (13.58)	
	AniHUI	100	72.28 (16.17)	0.139	87.01 (15.58)	**0.290***
	SUS		69.85 (18.73)		82.70 (14.12)	

Notes: * d≈0.2 → small effect; ** d≈0.5 → medium effect, *** d≈0.8 → large effect

The next finding relates to questionnaire experience. As mentioned in the theory part, several assertions in the literature concerning the advantages of pictorial scales over traditional verbal scales have not been empirically tested. These assertions often have an anecdotal foundation, rely on unsystematic observations, or are based on assumptions:

- *Increased motivation.* Participants reported pictorial scales being pleasant or even enjoyable (Desmet, 2003). Pictorial scales would help motivate respondents (Haddad et al., 2012), and they would focus attention and stimulate interest (Valla et al., 1994).
- *Intuitive comprehension.* Pictorial scales are easy to complete since measurement is more direct than using words (Bradley & Lang, 1994). There is no 'necessity for translating feelings into words' (Kunin, 1955, p. 66). Pictures help to identify the questions more easily (Haddad et al., 2012).
- *Reduced workload.* A pictorial item creates less mental workload than a verbally anchored one (Weibel et al., 2015; Wissmath et al., 2010).
- *Reduced completion time.* Using pictorial measures enables participants to respond more quickly than verbal measures (Lang, 1985, as cited in Weibel et al., 2015).

Since respondent-centred aspects were assessed in the three studies with appropriately large sample sizes, more concrete conclusions can be drawn regarding their correctness. Table 5 summarises key advantages and whether they were met by the image-based scales used in the three studies.

Table 5. Overview of key advantages of image-based usability scales compared to verbal usability scales.

Study	Instrument	Motivation	Comprehension	Workload	Item completion time	Preference
One	H-SUS	☺	*n.m.*	😐	☹	☺
Two	PUI-L	😐	☹	☹	☹	☹
	PUI-S	☺	☹	😐	☹	☹
	HUI-L	☺	😐	😐	☹	☺
	HUI-S	☺	😐	☺	☹	☺
Three	HUI-M	☺	😐	😐	☹	☺
	AniHUI	☺	😐	😐	☹	☺

Notes: ☺=better than verbal questionnaire, ☹=worse than verbal questionnaire, 😐=same as verbal questionnaire, n.m.=not measured

Motivation. The findings from the three studies indicate that all image-based instruments were perceived as more motivating than the verbal scales, except for PUI-L, which did not differ significantly from the ratings of the verbal scales. This observation might be indicative that one or more other factors might have influenced the respondents' experience of completing PUI-L negatively. Nevertheless, the results largely support the assertion of increased motivation and highlight a key advantage of image-based instruments.

Comprehension. Results concerning comprehension showed that all hybrid instruments were perceived as equally comprehensible as the verbal scales. However, the purely pictorial scales (PUI-L, PUI-S) were rated less comprehensible than the verbal scales. To offer an explanation, hybrid instruments have the advantage of redundant information in the form of a verbal description (i.e. redundancy gain, cf. Baumgartner et al., 2021). On the contrary, purely pictorial instruments such as the PUI-L or the PUI-S do not have such a fallback. Hence, their items' meanings need to be deciphered visually with the available pictorial information. If this information does not align with the respondent's mental model, which might have happened in this case, it may lead to ambiguous interpretations (cf. Baumgartner et al., 2019b). Therefore, the findings of these studies support only that hybrid scales are as comprehensible as verbal scales, but not more.

Workload. Most of the instruments did not differ regarding perceived workload. Only the HUI-S was rated lower in workload than the verbal usability questionnaires, and PUI-L was rated higher. Given these findings, we imagine that workload is impacted by several factors, including the complexity of the depictions and how clear they are, whether there is additional information available, such as verbal cues, but also the length of the instrument (i.e. the shorter the scale, and the clearer the underlying concept, the lower the perceived workload). In conclusion, the claim of pictorial scales having lower workload holds only true under certain circumstances, as it was the case for HUI-S.

Completion time. The results for item completion time showed clearly that image-based scales needed more time to complete than verbal scales, which contradicts assertions made in the literature. Item completion times for H-SUS, PUI-L, and HUI-L were about 1.70 seconds longer, those of the short and medium versions (PUI-S, HUI-S, HUI-M) were about 3.00 seconds longer, and those of the animated scale items were about 5.00 seconds longer than completion time of a verbal usability item. The results suggest that item completion time decreases the more items the image-based scale has. This observation might suggest a learning effect (e.g. Karni & Sagi, 1991), indicating that participants need to become accustomed to this kind of scale to complete it efficiently. There might be occasions when pictorial items need less time to be completed, for instance, when the underlying meaning is more concrete or when scales are applied repeatedly. There is evidence in the literature that

recognition of image-based material is faster than recognition of words (e.g. Potter, 1976; Potter et al., 2004).

Preference. Although there is no explicit assertion about increased preference for image-based scales in the literature, one might assume that if all advantages were true, one would prefer image-based scales over verbal ones. Findings show that in all three studies, most participants preferred the hybrid scales compared to the verbal scales. Surprisingly, the pictorial scales (PUI-L, PUI-S) used in study two were less preferred than the verbal scales. This finding goes against the results of previous studies (e.g. Baumgartner et al., 2020, 2019b), in which purely pictorial scales were tested with samples of a similar composition.

In summary, while the psychometric properties revealed only minor differences between the instruments, the respondent-centred measures indicate that the unique characteristics of each instrument, such as scale type (pictorial, hybrid, animated) or scale length (long, medium, short), have an impact on respondents' experience. While some assertions made in literature, such as increased comprehension (Desmet, 2003; Ghiassi et al., 2011), lowered workload (Weibel et al., 2015; Wissmath et al., 2010) and reduced completion times of pictorial scales (Lang, 1985, as cited in Weibel et al., 2015) have been proven too optimistic in most cases, the present findings support the idea of an increased motivation when using image-based scales.

7.2. Implications for practitioners and researchers

The studies outlined in this work were among the first that systematically compared verbal scales with pictorial scales (or variations such as hybrid and animated scales), not only on psychometric quality but also on a subjective experiential level. Findings shed light on assertions about advantageous characteristics of such scales that were made in research but never tested systematically. Therefore, these studies contributed to fundamental research on pictorial scales and their advantages and disadvantages.

Furthermore, the findings of these studies were vital in demonstrating that image-based scales are valid and reliable alternatives to traditional verbal scales in usability assessment. The studies expanded the toolbox of verbal usability questionnaires with image-based ones that researchers and practitioners can use for assessing perceived usability. We hope that this

contribution fosters the advancement of new types of questionnaires in the usability domain that follow a more inclusive approach.

Another noteworthy implication is introducing the QX framework in scale development. We hope that QX encourages researchers and practitioners to use it as an additional source of information together with psychometric properties to ensure the creation of new instruments with excellent overall quality. The QXQ used in studies II and III can be instrumental in identifying general issues of existing questionnaires or new ones under development.

7.3. Limitations

The studies in this work have some limitations. The sample composition primarily consisted of highly educated participants, with the proportion of students ranging between 62.5% and 91.9%. Additionally, participants were relatively young, with mean ages ranging between 23 and 28 years, and 73-79% of participants were female. Even if we do not assume drastically different results with a more heterogeneous sample composition, we cannot assert this with absolute certainty.

Another limitation related to the sample is that the scales have not been tested with user groups that could particularly benefit from using image-based scales, such as dyslexic people or users with limited educational backgrounds (e.g. Sauer et al., 2020). Furthermore, the scales were only tested with samples from Switzerland (i.e. Western culture). Therefore, we cannot confidently generalise the findings of how useful or comprehensible these image-based scales would be for other user groups or participants with diverse cultural backgrounds.

The last limitation concerns the test setting. All three studies were conducted online. Participants' environment and their device settings (e.g. browser version) could not be fully controlled as it would have been possible in a lab setting. While the sample sizes were sufficiently large, we did not expect much bias from the online setting. However, we cannot exclude the possibility that external factors may have influenced some participants during study participation (e.g. distractions, interruptions, noise).

7.4. Future research

Future studies in this domain could validate image-based usability scales by testing them with more diverse samples, such as stratified samples that consider various subgroups based on

specific characteristics (e.g., age, gender, education level, native language, cultural background). Objectives could consist of testing whether findings can be generalised across different populations and whether these scales benefit user groups that have problems processing verbal questionnaires (cf. Sauer et al., 2020).

Another venue worth exploring is the scale representation and how participants identify with it. Several pictorial scales developed for children give characters a name (e.g. Darryl; Neugebauer et al., 1999) or use animal characters like a dog (King et al., 2017) or a koala (Muris et al., 2003). Furthermore, storytelling elements could be used to embed the character in a narrative and connect the respondent with the protagonist. It might be interesting to test whether such approaches would increase questionnaire experience also in adult participants.

While the animated scales used in study three did not reveal significant advantages over the static scales, it would be rushed to categorically rule out the possibility that animations in questionnaires could offer additional benefits. Other interaction patterns could be explored, for instance, using a slider that visually manipulates the degree of agreement instead of just playing an animation when a scale point is chosen. Such scales would give the respondent more control over the animation and could positively impact engagement and perception of the scale. Another idea might be using additional sound effects (cf. Caicedo & Desmet, 2009) to make the interaction with the animation more vivid and pleasant.

Finally, the concept of QX could be developed and enhanced from a theoretical point of view. As mentioned in study one, the outlined framework should be considered a starting point. There might be other components we did not assess yet, that could serve well as respondent-centred measures for QX. In this context, the adaption or extension of the QXQ needs to be considered. A challenge might be to extend the QXQ with suitable measures while at the same time maintaining a short instrument. To address this constraint, single-item scales could be considered (Wanous et al., 1997). Furthermore, establishing norm data would be very useful. This can serve as a yardstick to draw a conclusion on whether a questionnaire is acceptable concerning specific respondent-centred aspects (e.g. is the questionnaire comprehensible enough or is it not).

7.5. Conclusion

While further studies are needed to validate image-based instruments with more diverse samples, we can conclude that pictorial and hybrid scales achieve adequate psychometric properties comparable to established verbal instruments. However, the findings of this work suggest that hybrid scales score better on respondent-centred aspects than purely pictorial questionnaires or verbal ones, which is also backed by respondents' preference ratings. At least with the samples used in the three studies, findings underline that hybrid scales are the better choice in terms of QX. We can also conclude that the QX of a given instrument is largely shaped by their unique characteristics, such as the length of the instrument and the type of image-based scale (i.e. pictorial, hybrid, animated). Furthermore, when time is of the essence, the shorter the instrument, the better. All the better if the questionnaire engages the respondents with an element of enjoyment.

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).

Aubert, M., Setiawan, P., Oktaviana, A. A., Brumm, A., Sulistyarto, P. H., Saptomo, E. W., Istiawan, B., Ma'rifat, T. A., Wahyuono, V. N., Atmoko, F. T., Zhao, J.-X., Huntley, J., Taçon, P. S. C., Howard, D. L., & Brand, H. E. A. (2018). Palaeolithic cave art in Borneo. *Nature*, *564*(7735), Article 7735. https://doi.org/10.1038/s41586-018-0679-9

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*, 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. *International Journal of Human-Computer Studies*, *147*, 102575.

Baumgartner, J., Sauer, J., & Sonderegger, A. (2020). Pictorial usability inventory (PUI) a pilot study. *Proceedings of the Conference on Mensch Und Computer*, 43–52.

Baumgartner, J., Sonderegger, A., & Sauer, J. (2023a). Questionnaire experience of the Pictorial Usability Inventory (PUI)–a comparison of pictorial and hybrid usability scales. *International Journal of Human-Computer Studies*, 103116.

Baumgartner, J., Sonderegger, A., & Sauer, J. (2023b). *To move or not to: A comparison of*

static and animated usability scales.

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.

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.

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.

Caicedo, D. G., & Desmet, P. M. A. (2009). *Designing the new PrEmo*. Dissertação apresentada à Delft University of Tecnology.

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

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

Desmet, P. M., Vastenburg, M. H., & Romero, N. (2016). Mood measurement with Pick-A-Mood: Review of current methods and design of a pictorial self-report scale. *Journal of Design Research*, *14*(3), 241–279.

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

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

Gault, R. H. (1907). A history of the questionnaire method of research in psychology. *The Pedagogical Seminary*, *14*(3), 366–383.

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. *Proceedings-28th International PLEA Conference on Sustainable Architecture+ Urban Design: Opportunities, Limits and Needs-towards an Environmentally Responsible Architecture*.

Harari, Y. N. (2014). *Sapiens: A brief history of humankind*. Random House.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*(4), 549–559.

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

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

Karni, A., & Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences*, *88*(11), 4966–4970.

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.

Kunin, T. (1955). The Construction of a New Type of Attitude Measure 1. *Personnel Psychology*, *8*(1), 65–77.

Laurans, G., & Desmet, P. M. (2017). Developing 14 animated characters for non-verbal self-report of categorical emotions. *J. Des. Res*, *15*(3–4), 214–233.

Lewis, C., & Mack, R. (1982). Learning to Use a Text Processing System: Evidence from "Thinking Aloud" Protocols. *Proceedings of the 1982 Conference on Human Factors in Computing Systems*, 387–392. https://doi.org/10.1145/800049.801817

Lewis, J. (2018). The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, *34*(7), 577–590.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When There's No Time for the SUS. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2099–2102. https://doi.org/10.1145/2470654.2481287

Lin, H. X., Choong, Y.-Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology*, *16*(4–5), 267–277.

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

Muris, P., Meesters, C., Mayer, B., Bogie, N., Luijten, M., Geebelen, E., Bessems, J., & Smit, C. (2003). The Koala Fear Questionnaire: A standardized self-report scale for assessing

fears and fearfulness in pre-school and primary school children. *Behaviour Research and Therapy*, *41*(5), 597–617.

Neugebauer, R., Wasserman, G. A., Fisher, P. W., Kline, J., Geller, P. A., & Miller, L. S. (1999). Darryl, a cartoon-based measure of cardinal posttraumatic stress symptoms in school-age children. *American Journal of Public Health*, *89*(5), 758–761.

Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things*. Civitas Books.

Paunonen, S. V., Ashton, M. C., & Jackson, D. N. (2001). Nonverbal assessment of the Big Five personality factors. *European Journal of Personality*, *15*(1), 3–18.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 509.

Potter, M. C., Staub, A., & O'Connor, D. H. (2004). Pictorial and conceptual representation of glimpsed pictures. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 478.

Robert, J.-M., & Lesage, A. (2017). Designing and evaluating user experience. In *The handbook of human-machine interaction* (pp. 321–338). CRC Press.

Rosenberg, M. (1965). Rosenberg self-esteem scale. *Journal of Religion and Health*.

Sauer, J., Baumgartner, J., Frei, N., & Sonderegger, A. (2020). Pictorial scales in research and practice. *European Psychologist*.

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

Sonderegger, A., Heyden, K., Chavaillaz, A., & Sauer, J. (2016). *AniSAM & AniAvatar: Animated Visualizations of Affective States*. 4828–4837. https://doi.org/10.1145/2858036.2858365

Steymans, H. U., Staubli, T., & Ünal, A. (2012). *Von den Schriften zur (Heiligen) Schrift: Keilschrift, Hieroglyphen, Alphabete und Tora*. Bibel+ Orient Museum.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. Oxford University Press, USA.

Taylor, F. W. (1911). *The principles of scientific management*. NuVision Publications, LLC.

Terman, L. M., & Merrill, M. A. (1960). *Stanford-Binet intelligence scale: Manual for the third revision, form lM.*

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. *Bulletin of Sociological Methodology/Bulletin de Méthodologie*

Sociologique, *142*(1), 57–74.

Valla, J.-P., Bergeron, L., Bérubé, H., Gaudet, N., & St-Georges, M. (1994). A structured pictorial questionnaire to assess DSM-III-R-based diagnoses in children (6–11 years): Development, validity, and reliability. *Journal of Abnormal Child Psychology*, *22*, 403–423.

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

Weibel, D., Schmutz, J., Pahud, O., & Wissmath, B. (2015). Measuring spatial presence: Introducing and validating the pictorial presence SAM. *Presence: Teleoperators and Virtual Environments*, *24*(1), 44–61.

Wissmath, B., Weibel, D., & Mast, F. W. (2010). Measuring presence with verbal versus pictorial scales: A comparison between online- and ex post-ratings. *Virtual Reality*, *14*(1), 43–53. https://doi.org/10.1007/s10055-009-0127-0

Yerkes, R. M. (1921). *Memoirs of the National Academy of Sciences, vol xv: Psychological examining in the United States Army.*

9. Acknowledgements

Many thanks go out to my supervisor and co-supervisor (Jürgen, Andreas), my coworkers and former coworkers from the university (Alain, Sven, Carli, Simon, Quentin) and from Puzzle ITC/We Are Cube (Mayra, Phippu, Josh, Julian), the former Bachelor students (Nicole, Annigna, Mona, Tania, Gaëlle, Oriane), my family (Marlen, Marion, Günther, Nico), my better half (Vero), my friends (Dän, Ali, Kathrin, Sylvain, Nils), and all the other ones I did not mention here.

Without your support and help for developing crazy ideas, discarding some, giving feedback, helping me improve my drawing skills, helping me in technical matters, encouraging me, getting my back, listening, being there, and so forth, this work would not have been possible.

10. Publications

Baumgartner, J., Sonderegger, A., & Sauer, J. (2024). *To move or not to: a comparison of static and animated usability scales* [Manuscript submitted for publication].

Baumgartner, J., Sonderegger, A., & Sauer, J. (2023). Questionnaire experience of the Pictorial Usability Inventory (PUI)–a comparison of pictorial and hybrid usability scales. *International Journal of Human-Computer Studies*, 103116.

Baumgartner J., Ruettgers, N., Hasler, A., Sonderegger A. and Sauer J. (2021). Questionnaire experience and the hybrid system usability scale: using a novel concept to evaluate a new instrument. *International Journal of Human-Computer Studies*, 147, 102575.

Sauer, J., Baumgartner, J., Frei, N., & Sonderegger, A. (2020). Pictorial scales in research and practice. *European Psychologist*.

Baumgartner, J., Sauer, J., & Sonderegger, A. (2020). Pictorial usability inventory (PUI) a pilot study. In *Proceedings of Mensch und Computer 2020* (pp. 43-52).

Baumgartner, J., Frei, N., Kleinke, M., Sauer, J., & Sonderegger, A. (2019, May). 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., Sonderegger A. and Sauer J. (2019). No need to read: Developing a pictorial single-item scale for measuring perceived usability. *International Journal of Human-Computer Studies*, 122, 78-89.