



## To Animate or Not to Animate Usability Scales: The Effect of Animation on Questionnaire Experience and Psychometric Properties

Juergen Baumgartner, Andreas Sonderegger & Juergen Sauer

**To cite this article:** Juergen Baumgartner, Andreas Sonderegger & Juergen Sauer (15 Apr 2024): To Animate or Not to Animate Usability Scales: The Effect of Animation on Questionnaire Experience and Psychometric Properties, International Journal of Human-Computer Interaction, DOI: [10.1080/10447318.2024.2338333](https://doi.org/10.1080/10447318.2024.2338333)

**To link to this article:** <https://doi.org/10.1080/10447318.2024.2338333>



© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 15 Apr 2024.



Submit your article to this journal [↗](#)



Article views: 298




View related articles [↗](#)



View Crossmark data [↗](#)

# To Animate or Not to Animate Usability Scales: The Effect of Animation on Questionnaire Experience and Psychometric Properties

Juergen Baumgartner<sup>a,b</sup> , Andreas Sonderegger<sup>a,c</sup>, and Juergen Sauer<sup>a</sup>

<sup>a</sup>Department of Psychology, University of Fribourg, Fribourg, Switzerland; <sup>b</sup>We Are Cube, Bern, Switzerland; <sup>c</sup>Business School, Institute for New Work, Bern University of Applied Sciences, 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. This study aimed to extend the static pictorial representation by using animations. Previous research has not yet addressed the positive or negative outcomes of using animations in pictorial questionnaires. We hypothesized that an animated questionnaire would have an additional positive effect on respondents' questionnaire experience, motivation, and preferences without impinging psychometric properties. The goal of the present study was to compare the static HUI with an animated version (AniHUI) in an online test setting. Respondent-centered aspects (questionnaire experience) and 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 the static and animated scales were superior to the SUS regarding respondent-centred measures. Findings suggest that the HUI and the AniHUI are engaging 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-centered 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 favorably on respondent-centered aspects (i.e., motivation, aesthetics, and perceived completion time).
- The animated questionnaire did not emerge to be more engaging than the static one, being at the same level as the hybrid questionnaire.

## 1. Introduction

The presumably most common and cost-effective 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. Standardized 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 verbal questionnaires comes with certain limitations: (1) Only the literate population can answer them (Sonderegger et al., 2016). (2) Validated instruments are often unavailable in languages other than English, making 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, even fewer questionnaires 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

**CONTACT** Juergen Baumgartner  [juergen.baumgartner@unifr.ch](mailto:juergen.baumgartner@unifr.ch)  Department of Psychology, University of Fribourg, Rue P.-A.-de-Faucigny 2, Fribourg, 1700, Switzerland

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

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 into 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-centered development process is considered 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 standardized questionnaires. Over the past 30 years, more than 20 standardized 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 for developing and using pictorial scales is to provide users with inclusive access to questionnaires and facilitate usability evaluation. Especially hybrid scales have been shown in past studies to be 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 visualization of abstract concepts (e.g., usability) and 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 joy, and increase the respondents’ motivation to complete this kind of scale (e.g., Desmet, 2003; Haddad et al., 2012). There are also some disadvantages: (1) When completing hybrid scales, participants typically need 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 does not match well, there is the risk of ambiguity, which might lead to confusion and wrong answers. (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 that hybrid instruments have promising advantages, and potential drawbacks, we searched for ways to improve their characteristics to take advantage of their positive features and mitigate the negative ones. In this work, we considered the inclusion of animations as a promising next step in the evolution 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 various contexts, such as arts, advertising, and marketing, but also in learning environments, such as computer animations for medical education (Knapp et al., 2022; Ruiz et al., 2009). In the context of questionnaire design, an animated scale brings motion into play as an additional element. Consequently, we 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 a few validated questionnaires match the definition of an animated instrument. In emotion research, PREMIO (Product Emotion Measurement Tool, Desmet, 2003; Laurans & Desmet, 2017) was created to assess 14 emotions toward 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 the activity limitations of patients with hip or knee osteoarthritis. A further animated scale was developed by Setty et al. (2019) to assess 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 a smile to a frown.

Several potential disadvantages are related to the use of animated questionnaires: (1) Rebetz et al. (2010) argue that animations could have an overwhelming effect on the

working memory since change between frames needs to be memorized and processed 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 must wait until the animation ends to have all the information ready for subsequent interpretation. This might lead to a longer item completion time. (3) Finally, creating and implementing 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 expressing 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) and hence might contribute to an improved experience of answering a questionnaire.

#### 1.4. Questionnaire experience

Questionnaire experience (QX) is a recently introduced concept aiming to capture respondents' subjective experiences when answering 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 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 developed to assess perceived usability. It consists of a verbal question (e.g., "How quickly did you achieve your goal with the website?") and pictorial information that visually expresses the corresponding answer options. The pictorial content is based on the PUI (Pictorial Usability Inventory, Baumgartner et al., 2020). However, it uses a

subset of the original 12 items to make the instrument more economical and less time-consuming (see Figure 1). The selection of the six items was based on the results of a comprehension test that was conducted for a previous study (cf. Baumgartner et al., 2023). The six items with the highest comprehension rates were selected for the present study.

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 to the user the handling of the scale ("Use the buttons -2 to 2 to select the option that most applies to you").

Several design strategies were applied to distinguish adequately between answer options, consisting of (1) a change in the avatar's facial expression (e.g., frowning vs. smiling), (2) the use of colors for key elements (i.e., red, grey and green tones), and (3) the application of Weber's law using geometric progression to express the 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 to overcome binary stereotypes and avoid the need to implement 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-s primary animation representing the main idea of the item by manipulating graphic elements (e.g., completing the path to a goal or counting the time on the stopwatch). Figure 3 shows the animation sequence. The animation is repeated once to ensure that the respondent does not miss any information. A secondary 1-s 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.

#### 1.6. Aim of the research and hypotheses

This study aimed to systematically compare respondent-centered aspects and psychometric properties of the HUI with an animated version of the same instrument (AniHUI) in an online test setting (i.e., a website usability test). The primary goal consisted of gaining insights into 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.



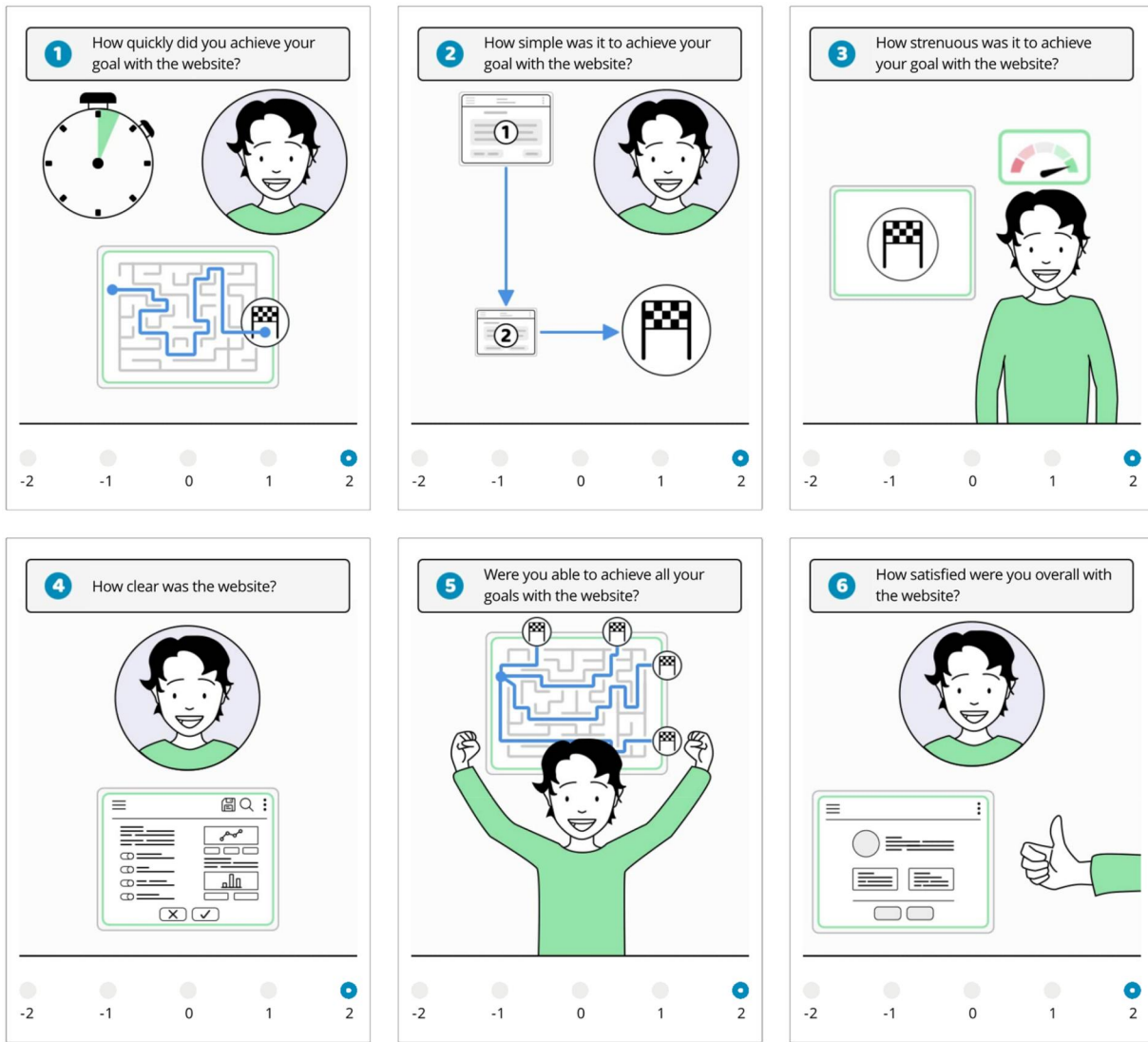


Figure 1. HUI items 1–6 with the most positive answer option selected.

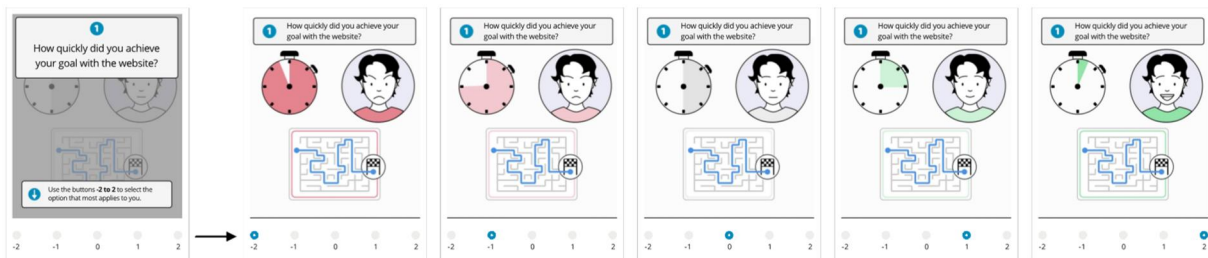


Figure 2. HUI item 1 with the initial display for the question and the answer options.

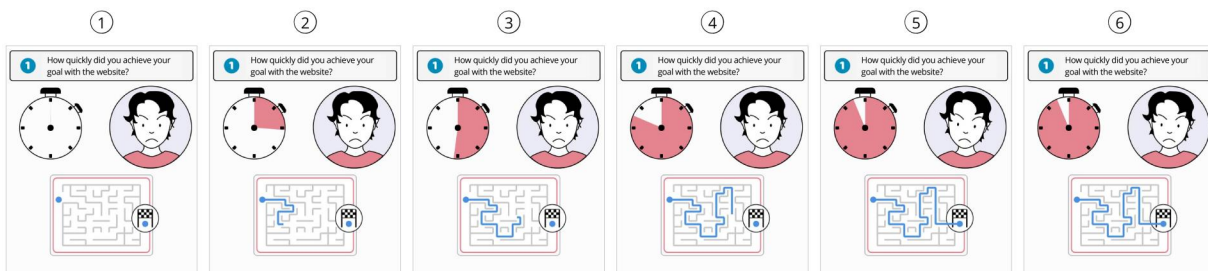


Figure 3. Primary animation sequence of AniHUI item 1, from beginning to end state.

In general, we expect HUI to have 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 the 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 most 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 times for hybrid items are generally longer than for 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 to have the shortest item completion time since only verbal content is shown.

**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 tend to lose track of time when engaged in a pleasant or motivating activity (Conti, 2001; Nakamura & Csikszentmihalyi, 2014). Since we expect completing the animated questionnaire as a pleasant activity, we assume that time flies faster for the participants during questionnaire completion. We expect a similar effect to happen for the hybrid questionnaire but to a lesser extent.

## 2. Methods

### 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. The sample consisted of 192 participants (75.5% female, 24.5% male)

with 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 who did not report their professional status (5.2%). Two participants ( $\approx 1\%$ ) reported having some form of color 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 they had seen the website before.

An a priori sample size estimation was conducted using effect sizes obtained in a previous study ( $d = .868$ ; Baumgartner et al., 2023). According to the power calculation, 76 participants would be required to achieve a power of  $1 - \beta = .95$  assuming an error probability of  $\alpha = .050$ . Being aware of the issues of using exemplary data to estimate population effect sizes (e.g., Anderson, 2019), we considerably increased our sample size to be able to detect smaller effects.

### 2.2. Website prototype and user tasks

In the present study, participants interacted with a website of a fictitious leisure center, 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 successful. In contrast to the previous study, participants completed only two instead of three tasks to minimize 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 center. Participants were able to navigate on the webpage to solve the tasks freely. Furthermore, they were instructed to move to the next task in case they could not find the solution within 4 min. A browser script was used to record whether participants interacted with the website.

### 2.3. Measures and instruments

The measures and instruments used in this study are divided into respondent-centered and psychometric measures. Respondent-centered 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-centered measures

To assess respondent-centered aspects of the usability questionnaires, the Questionnaire Experience Questionnaire (QXQ; Baumgartner et al., 2023) was presented after the 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.

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 to distinguish 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). This study assessed sensitivity 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 the 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). Sauro and Lewis (2016) have introduced a grading system, ranging from "A" to "F" for easier interpretation of scores. This study used a validated German version of the SUS (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 well the items of a questionnaire relate to each other (Coolican, 2017). Internal consistency is expected to be high when a questionnaire is assumed to measure a one-dimensional construct. Internal consistency of HUI, AniHUI, and SUS was assessed by 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 the 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 the 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

**Table 1.** Items of the questionnaire experience questionnaire (QXQ) and Cronbach's alpha values for multi-item scales (based on Baumgartner et al., 2023).

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?	–

The wording was translated from German to English.

either the low or the high usability version of the website of the fictitious leisure center. 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 already knew the webpage. On the subsequent pages, participants completed the post-test usability questionnaires, consisting of either HUI or AniHUI, and SUS. The sequence of presenting hybrid and verbal usability questionnaires was counterbalanced to prevent order effects (i.e., half of the participants completed HUI/AniHUI first, and the other half SUS first). After each usability questionnaire, respondent-centered measures were assessed using the QXQ. On the last pages, participants were asked how they perceived the avatar (i.e., the gender evaluation of the avatar, the 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 who genuinely interacted with the website prototype, (3) participants without multiple study participation, and (4) participants who responded “yes” to the question of whether they completed the study seriously. Out of 243 participants, 192 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-centered measures (Mann–Whitney  $U$  test, Wilcoxon signed-rank test), correlational analyses for convergent measures (Spearman’s rank correlation), analysis of variance to check for potential order effects, 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-centered measures

#### 3.1.1. QXQ

Wilcoxon tests for all six measurable indicators of the QXQ were conducted to identify differences in respondent-centered 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.

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

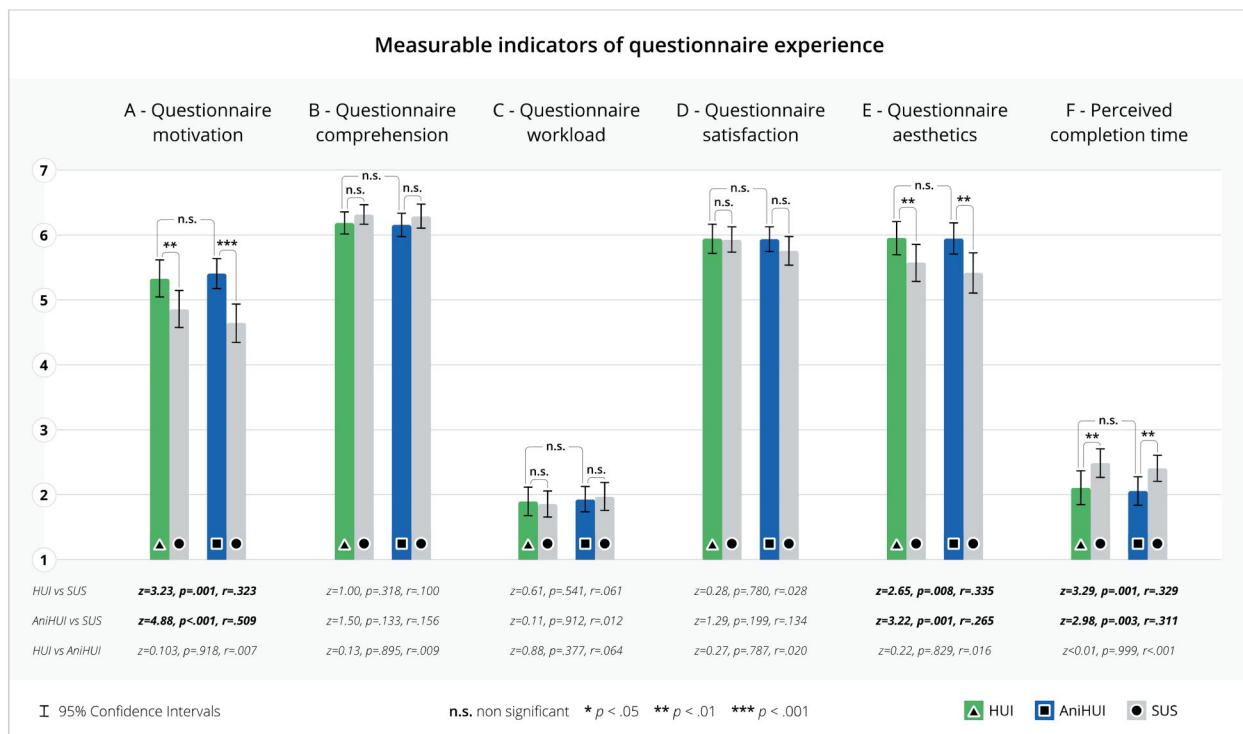


Figure 4. Overview of QXQ indicators, including statistical parameters of Wilcoxon test (HUI vs. SUS, AniHUI vs. SUS) and Mann–Whitney  $U$  test (HUI vs. AniHUI).



The analysis of the between-subjects comparisons showed no significant difference for any of the respondent-centered 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%).

### 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 ( $p > .05$ ).

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 s 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. Order effects

Analyses of variance were conducted to assess whether the order of questionnaire administration had an effect on the usability ratings. The analysis showed no significant main effects of order on HUI [ $F(1, 90) = 0.86, p = .412, \eta^2_{\text{partial}} = .007$ ] and SUS [ $F(1, 90) = 0.20, p = .412, \eta^2_{\text{partial}} = .007$ ] and AniHUI [ $F(1, 98) = 1.38, p = .244, \eta^2_{\text{partial}} = .014$ ] and SUS [ $F(1, 98) = 0.67, p = .415, \eta^2_{\text{partial}} = .007$ ].

### 3.2.2. 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 conditions, with AniHUI showing a large effect size ( $r = .500$ ), and HUI and SUS showing medium effect sizes ( $r \approx .370$ ).

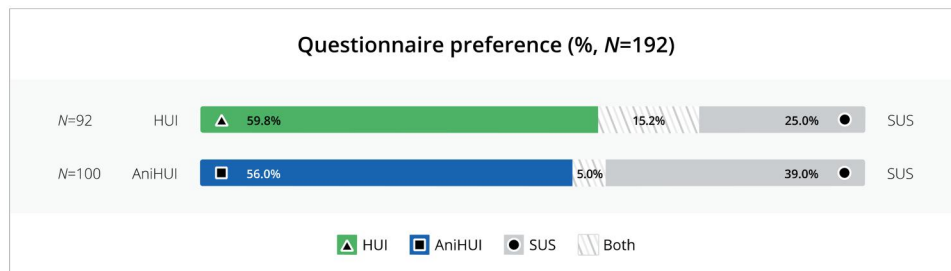


Figure 5. Overview of questionnaire preference for HUI, AniHUI, and SUS.

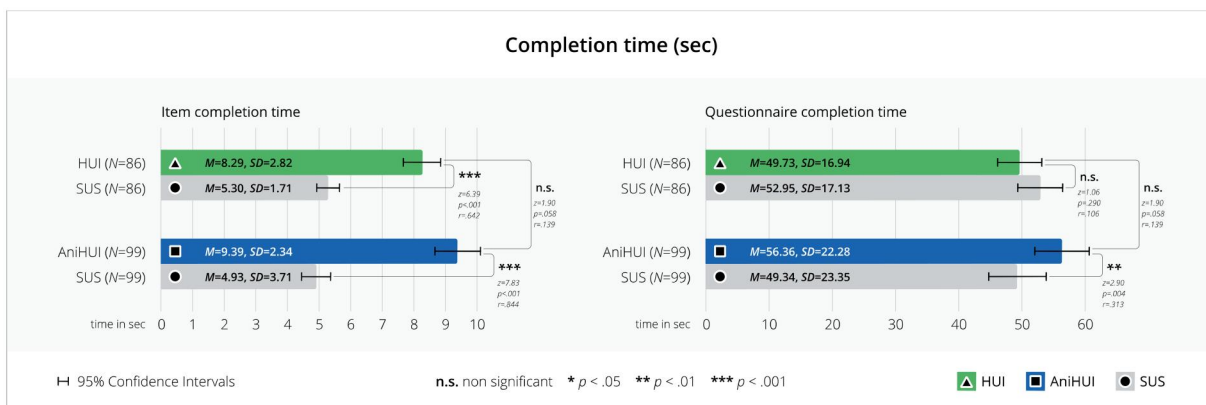


Figure 6. Overview of item and questionnaire completion time for HUI, AniHUI, and SUS. Notes: Data of  $N = 7$  participants (3.76% of the overall sample) were excluded from data analysis since it was identified as outliers.

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); \*\*\* $p < .001$ .

### 3.2.3. Convergent validity

Correlations were calculated to determine convergent validity (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.

### 3.2.4. 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 ( $\alpha = .827$ ), AniHUI ( $\alpha = .814$ ), and SUS ( $\alpha = .886$ ).

### 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 perceive 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 12% being undecided.

## 4. Discussion

This study systematically compared a static hybrid usability questionnaire (HUI) with an animated hybrid questionnaire (AniHUI), focusing on respondent-centered aspects of questionnaire experience and psychometric properties. In addition, both instruments were compared with a standardized instrument that measures perceived usability (i.e., the SUS). Findings indicate that respondent-centered 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 questionnaires obtained fairly similar results regarding psychometric properties (i.e., high sensitivity, high convergent validity, and good internal consistency).

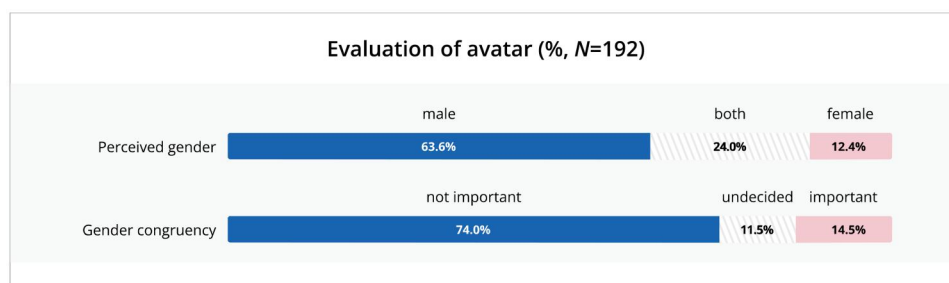
With regard to respondent-centered measures, we assumed in our first hypothesis (H1) that motivation and preference were highest for AniHUI, followed by HUI and SUS (AniHUI > HUI > SUS). Results indicated that HUI and AniHUI obtained considerably higher motivation ratings than SUS. Although we expected the AniHUI to be more engaging 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 much between them (HUI: 59.8%; AniHUI: 56.0%). Therefore, the 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 a mechanism to convey the intended meaning and 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 bias in 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 an auditive supplement that emphasizes 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 ( $\approx 5$  s). However, no significant difference was found between HUI ( $\approx 8$  s) and AniHUI ( $\approx 9$  s), although results are pointing toward 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 from the hybrid version.

**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***

Note: \*\*\* $p < .001$ .



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

Looking at questionnaire completion time, results suggest that HUI and SUS need about the same amount of time to complete ( $\approx 50$  s), and AniHUI needs a couple of seconds longer ( $\approx 56$  s). Even if HUI and AniHUI have four items less than the SUS, both instruments are completed in under 1 min 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 ( $\text{AniHUI} < \text{HUI} < \text{SUS}$ ). Results suggest 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 the objective completion time for HUI and SUS is about the same ( $\approx 50$  s) and even longer for AniHUI ( $\approx 56$  s). However, from the respondents' point of view, it is perceived as faster than completing the verbal questionnaire. One explanation might be that participants are generally more engaged when processing pictorial questionnaires and tend to lose track of time (e.g., Conti, 2001). Whether the image-based elements are animated or not does not seem to matter. Another explanation could be that participants used the number of items as an argument for comparison and evaluated the instruments with fewer 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 ( $r \approx .370$ ) and a large effect for the AniHUI ( $r \approx .500$ ). In this regard, HUI and SUS behave very similarly, whereas differences using the AniHUI seem more pronounced, especially in the high-usability condition. Analyses for convergent validity showed that correlations between HUI and SUS were generally high ( $r = .827$ ), and correlations with the single-item scale for satisfaction were substantial and in the same range as those with the SUS ( $r \approx .780$ ). This finding is also reflected in the obtained average usability score, which is almost the identical 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 ( $\alpha = .827$ ), AniHUI ( $\alpha = .814$ ), and SUS ( $\alpha = .886$ ), indicating the items of the questionnaires relate well to each other. Taken together, data analysis indicates good psychometric values for HUI 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). We did not find any evidence to support this assumption in this study. No significant differences were observed between AniHUI and HUI or SUS concerning relevant respondent-centered aspects, such as questionnaire workload or questionnaire comprehension. Therefore, we 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. Since about 15% of participants indicate that adequate representation is important to them (and an additional 12% are unsure about this), the question of gender-appropriate representation in pictographic scales is relevant. 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 to design and implement 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 were highly educated. We assume that students are more efficient at completing questionnaires than non-students, which might have influenced some of the results (e.g., completion time). In addition, the size of the sample in this study allows to detect medium to large effects. It would be interesting for future studies to address the non-significant effects of interest using a larger and more diverse sample. Furthermore, roughly a fifth of the participants reported already having seen the webpage. Since we cannot know which version of the website (i.e., low or high usability) they interacted with 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. As an additional limitation, we note the mixed experimental design employed in this study, in which the assessment of HUI and AniHUI was conducted between-subjects, while the SUS was assessed within-subjects. For a more stringent experimental design, the inclusion of a control group exclusively assessing the SUS would have been beneficial. But since there were no effects of questionnaire order detected, we believe that the potential impact on results was minimal.

Future research may look further into the direction of 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). Scale developers should be aware, however, that animating a questionnaire is labor-intensive, and whether the additional effort ultimately pays off should be considered. Another promising line of research may be to determine better which audiences benefit from hybrid or animated scales. There is a list of assumptions concerning favorable conditions for administering pictorial scales to groups, such as non-native speakers or people with poor language skills (see Sauer et al., 2020). However, research has not yet examined whether the 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-centered measures and psychometric properties, we suggest for practitioners and scientists alike that both instruments are suitable to assess perceived usability.

## Acknowledgments

We are very grateful to We Are Cube and Puzzle ITC for the support in the 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.

## Ethical approval

This study obtained ethical approval from the Internal Review Board within the Psychology Department of the University of Fribourg (approval no. 546).

## Disclosure statement

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

## Funding

This study was funded by a research grant (no. 100019\_188808) from the Swiss National Science Foundation (SNSF).

## ORCID

Juergen Baumgartner  <http://orcid.org/0000-0003-1341-7502>

## References

- Anderson, S. F. (2019). Best (but oft forgotten) practices: Sample size planning for powerful studies. *The American Journal of Clinical Nutrition*, 110(2), 280–295. <https://doi.org/10.1093/ajcn/nqz058>
- 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*, 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. <https://doi.org/10.1016/j.ijhcs.2020.102575>
- 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. (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. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9)
- Brooke, J. (1996). SUS: A “Quick and Dirty” Usability Scale. In *Usability evaluation in industry* (pp. 207–212). <https://doi.org/10.1201/9781498710411-35>
- Bülül, A. H., & Abdullah, K. (2021). Emotional design of educational animations: Effects on emotion, learning, motivation and interest. *Participatory Educational Research*, 8(3), 344–355. <https://doi.org/10.17275/per.21.69.8.3>
- Caicedo, D. G., & Desmet, P. M. A. (2009). *Designing the new PrEmo*. Dissertação apresentada à Delft University of Technology.
- 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 product research*. Delft University of Technology, Department of Industrial Design.
- 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. <https://doi.org/10.1504/JDR.2016.079751>
- Eisner, W. (1985). *Theory of comics and sequential art*. 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. <https://doi.org/10.1080/10447318.2020.1801173>
- Gault, R. H. (1907). A history of the questionnaire method of research in psychology. *The Pedagogical Seminary*, 14(3), 366–383. <https://doi.org/10.1080/08919402.1907.10532551>
- 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. *Communications of the 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. <https://doi.org/10.1016/j.jpain.2008.08.005>
- 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-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>
- Knapp, P., Benhebl, N., Evans, E., & Moe-Byrne, T. (2022). The effectiveness of video animations in the education of healthcare practitioners and student practitioners: A systematic review of trials. *Perspectives on Medical Education*, 11(6), 309–315. <https://doi.org/10.1007/s40037-022-00736-6>
- 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. <https://doi.org/10.1089/cpb.2005.8.493>
- 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. *Journal of Design Research*, 15(3/4), 214. <https://doi.org/10.1504/JDR.2017.089903>
- Lewis, J. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, 14(3), 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. <https://doi.org/10.1080/10447318.2018.1455307>
- Messick, S. (1979). Test validity and the ethics of assessment. *ETS Research Report Series*, 1979, 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. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 249–256). <https://doi.org/10.1145/97243.97281>
- 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. <https://doi.org/10.1007/s00296-015-3230-4>
- Rebetz, C., Bétrancourt, M., Sangin, M., & Dillenbourg, P. (2010). Learning from animation enabled by collaboration. *Instructional Science*, 38(5), 471–485. <https://doi.org/10.1007/s11251-009-9117-6>
- Reynolds-Keefe, L., Johnson, R., & Carolina, S. (2011). Is a picture 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. <https://doi.org/10.1177/0146167201272002>
- 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. <https://doi.org/10.1111/j.1365-2923.2009.03429.x>
- Salah, D., Paige, R. F., & Cairns, P. (2014). A systematic literature review for agile development processes and user centred design integration. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (pp. 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*, 26(2), 112–130. <https://doi.org/10.1027/1016-9040/a000405>
- Sauer, J., Sonderegger, A., & Schmutz, S. (2020). Usability, user experience and accessibility: Towards an integrative model. *Ergonomics*, 63(10), 1207–1220. <https://doi.org/10.1080/00140139.2020.1774080>
- 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. <https://doi.org/10.17245/jdapm.2019.19.4.227>
- Sonderegger, A., Heyden, K., Chavallaz, 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, September 2–6). The UX construct—Does the usage context influence the outcome of user experience evaluations? In *Human-Computer Interaction-INTERACT 2019: 17th IFIP TC 13 International Conference, Proceedings, Part IV 17*, Paphos, Cyprus (pp. 140–157).
- Tversky, B., Morrison, J. B., & Bétrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, 57(4), 247–262. <https://doi.org/10.1006/ijhc.2002.1017>

## About the authors

**Juergen Baumgartner** is a PhD student in Psychology at the University of Fribourg (CH) and a Senior UX Consultant at Puzzle ITC in Bern (CH). He received his MSc in Work and Organisational Psychology from the University of Fribourg (CH) in 2015.

**Andreas Sonderegger** is Professor at Bern University of Applied Sciences (Business Department, Institute for New Work) and Lecturer at the University of Fribourg (Department of Psychology) (both CH). He received his PhD in 2010 and his MSc in Work and Organisational Psychology in 2003 from the University of Fribourg (CH).

**Juergen Sauer** is Full Professor of Cognitive Ergonomics at the Department of Psychology at the University of Fribourg (CH). He received an MSc in Occupational Psychology from the University of Sheffield (UK) in 1990 and a PhD in Psychology from the University of Hull (UK) in 1997.