Improving the usability of online usability surveys with an interactive Stripe scale

Matevž Pesek, Alja Isaković, Gregor Strle, Matija Marolt
University of Ljubljana
Faculty of Computer and Information Science
Laboratory for Computer graphics and Multimedia
{matevz.pesek,matija.marolt}@fri.uni-lj.si

ABSTRACT

The paper introduces Stripe, an interactive continuous scale for online surveys that makes it easy to compare multiple answers on a single screen. The Stripe is evaluated as an alternative to the n-point Likert scale, which is commonly used in online usability questionnaires like the System Usability Scale (SUS). The paper presents the results of a user study, which confirmed the validity of results gained with the proposed Stripe interface by applying both the Stripe and the Likert interface to an online SUS questionnaire. Additionally, the results of our study show that the participants favor the Stripe interface in terms of intuitiveness and ease of use, and even perceive the Stripe interface as less time consuming than the standard Likert scaled interface based on radio buttons.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Miscellaneous

Keywords

user interfaces, feedback gathering, human computer interaction, system usability score, user study, measurement scales

1. INTRODUCTION

Questionnaires are a common tool for usability evaluation in HCI research. For the purposes of our own usability testing, we developed Stripe, a more interactive and compact scale that fits on smaller screens and supports the comparison of answers across different questions. Knowing that the design of a user interface can affect the gathering procedure, and in some cases influence (or bias) the results [6, 8], we performed a user study that compared the validity of the newly proposed Stripe interface with the standard Likert scale.

The user study tested both user interfaces on the System Usability Scale (SUS) questionnaire for two well-known products. This gave us the ability to compare the SUS scores attained through both user interfaces to SUS scores reported by other studies. To further evaluate the potential of Stripe, we also performed a usability survey on both interfaces.

2. RELATED WORK

Usability is defined as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use [12]. There are many standard methodology tools available for measuring various usability aspects, ensuring the validity and comparability of results gained by a methodologically sound and well-structured approach. The tools vary in size and scope, but they all commonly use the Likert scale as the de facto standard for user-feedback gathering.

The NASA-Task Load Index (NASA-TLX) is a multi-dimensional scale designed to obtain workload estimates from a user performing a specific task [9, 10]. The ATTRAKDIFF questionnaire [11] is often used for qualitative evaluation of the pragmatic and hedonic aspects of a product or service. For measuring the usability aspects, the Software Usability Measurement Inventory (SUMI), Questionnaire for User Interaction Satisfaction (QUIS), System Usability Scale (SUS) and Usability Metric for User Experience (UMUX) are commonly used [13]. SUMI is a 50-item Likert scale questionnaire that measures five aspects of user satisfaction and scores them against expected industry norms. QUIS consists of a 27-item Likert scale and is similar to SUMI, but measures attitude towards 11 interface factors. SUS [2] is a 10-item Likert scale questionnaire measuring the usability and an overall satisfaction with a product or service. Finally, the UMUX [6] is a 4-item Likert scale questionnaire used for a subjective assessment of perceived usability.

For the purpose of testing new user interfaces for surveys, the SUS provides the right balance between length and precision with its 10 questions. Like other standard usability measurement methodology approaches, the SUS was carefully constructed from the beginning in order to achieve high reliability, validity and repeatability of results [2]. The result of the SUS is a single score, between 0 and 100.

2.1 Scales used in online questionnaires

Paper-based questionnaires have a long history of experimentation with different styles of rating scales, especially in the field of psychology. Visual analogue scales (VAS) appeared back in 1921 and were improved upon by graphic rating scales (GRS) in 1923 [4]. Both scales include an anchored horizontal line, with extreme values of the measured property listed at each end [4]. The user can place a mark anywhere along the continuous line.

In 1932 psychologist Rensis Likert introduced his own scale, which limits the number of available options to 5 in the original scale and no longer provides a continuum of choices along the line [5]. Since then, the Likert scale has been

adapted to different types of questionnaires, including online versions that use standard HTML input radio buttons.

In contrast, continuous line-based scales have not been supported by the HTML standard until recently. HTML5 introduced a new "range" input type, which creates a slider scale with a handle that can be moved along the line to select a value[1]. The slider can be configured to support discrete steps or to act as a continuous scale. A potential problem with this approach is that the initial slider position can influence the response and can even lead to a different response distribution when compared to traditional scales based on radio buttons [7]. Luckily, the wide adoption of the JavaScript programming language in modern web browsers offers new opportunities for more interactive user interfaces that can bypass the limitations of standard HTML input types.

Research on online survey interfaces tends to focus on the validity of results and user performance (completion time), but fails to evaluate other usability aspects of alternative interfaces. For example, Couper et al. [4] compared online questionnaires that used VAS to ones with different styles of radio buttons and surveys with numeric input fields. Their experiment found that while VAS surveys took longer to complete and contained more missing data, they produced the same response distributions as other types. Cook et al. [3] compared a slightly different style of online graphic rating slider scales with surveys based on radio buttons and found that both provided reliable scores, but also noticed that sliders took a bit longer to manipulate. User satisfaction and subjective perceptions were not evaluated in these studies, which calls for more HCI research that takes a wider range of usability aspects into account when evaluating new interfaces for online surveys.

In the following Section we propose the Stripe, an alternative to the Likert scale that aims to take advantage of the benefits of continuous scales while offering a more compact interactive user interface that makes it easy for users to compare answers, even on a smaller screen.

3. THE STRIPE: A DYNAMIC INTERFACE

The Stripe is a user interface developed to provide an interactive and intuitive continuous-scale alternative to the standard multi-point scale interfaces. It is implemented as a canvas with one horizontal dimension (Figure 1). The dimension represents the presence of a variable, ranging between two extremes (e.g. negative/positive, absent/significantly expressed, completely disagree/agree). This is similar to the standard VAS scale. But unlike the VAS or the Likert scale, the Stripe interface accommodates drag-and-drop functionality for multiple labels, as well as annotation of multiple categories on the same canvas. In its simplest form, the user is provided with a set of labels describing different nominal values of the variable. By dragging the labels onto different positions of the canvas, the user marks their perception of each individual label on a continuous scale. The positions of placed items can subsequently be quantized to discrete values, if so desired. The amount of information retrieved by the Stripe interface is therefore at least equal to the amount of information gathered by a radio button matrix (for example, a set of 5-point scales) commonly used to capture similar information

[1]http://www.w3schools.com/html/htmlforminputtypes.asp

The Stripe and its extended version were already used in an online survey on multi-modal perception of music [14], and later evaluated in terms of usability, using a modified version of the NASA TLX questionnaire [15]. However, in order to fully evaluate the potential of Stripe, it is necessary to compare it with the standard multi-point Likert scale approach, typically used in online surveys.

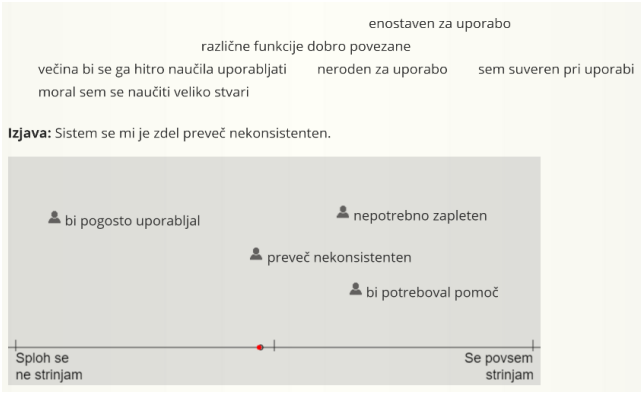

Figure 1: The Stripe interface. The statements are shortened into phrases for improved readability, but the full statement for each label is shown on 'mouse-over'.

4. EVALUATION

The goals of our experiment were: 1) to evaluate the validity of SUS scores gathered with the Stripe interface using the Likert scale as control and 2) to compare the usability of the Stripe interface with the usability of the standard Likert scale. The Stripe interface was designed for online questionnaires, so the experiment was conducted online. The user study was conducted on 2 different groups of participants, two months apart, to provide additional verification of the results.

4.1 Participants

A total of 68 participants, recruited from students and faculty members at the University of Ljubljana, fully completed the user study. Only participants with previous experience with the subject of the SUS were included in the survey, to obtain feedback from users with pre-existing experience and regular interaction with the chosen SUS subject.

For the first, Gmail survey, we collected feedback from 41 participants, 12 were male and 29 were female. Their average age was 29.4 years, with 7.1 standard deviation. For the second, Microsoft Exchange survey, we collected 27 responses from 21 female and 6 male participants. The average participant's age was 31.5 years, 7.9 standard deviation.

4.2 Experiment procedure

The user study was conducted online, with participants filling in all questionnaires on their own, using their own computers and their web browser of choice. At the beginning,

each participant was asked to confirm their familiarity with the product being evaluated in the SUS: Gmail for the first group of participants and Microsoft Exchange for the second group. Participants that passed this initial step continued to filling is the SUS questionnaire twice. The formal Slovenian translation of the SUS questionnaire was used [1].

The website randomly assigned either the Stripe or Likert version of the SUS first, followed by the other version, displayed on a separate page. The user interface used in the SUS questionnaire was the independent variable, the two configurations were the Stripe interface and the 5-point Likert scale. The resulting SUS score was the dependent variable. This part of the experiment lasted on average approximately 7 minutes per participant, no time limits were imposed.

After the SUS evaluation, the participants were presented with 3 additional usability questions on a 7-point scale:

- By comparing both, the Stripe and the 5-point scale interfaces, which of the interfaces was more intuitive and comprehensible? (1 - 5-point scale, 7 - Stripe)
- By comparing both, the Stripe and the 5-point scale interfaces, which of the interfaces takes more time to fill-in? (1 - 5-point scale, 7 - Stripe)
- Is it easier or more difficult to express your opinion with the Stripe interface (due to the visual comparison of your answers)? (1 - easier, 7 - more difficult)

Basic demographic data (age and gender) of participants with optional written feedback was also collected during the final step. All questions were asked in Slovenian language.

5. RESULTS AND DISCUSSION

The scores of the SUS questionnaire for both groups of participants and both interfaces are shown in Table 1. For both experiments, results indicate consistent responses gathered with each interface. However, the standard deviation of responses gathered by the Stripe interface is smaller. This is due to the use of a continuous scale, which allows for a more fine-grained positioning of the labels, unlike restricted options on traditional n-point scales. When we performed a quantization of continuous responses into a 5-point scale (row 3 in Table 1), the scores were very similar for both interfaces. The average SUS score for Gmail was close to the average SUS score of 83.5 from [3], further confirming the roboustness of the SUS questionnaire and the validity of our results for both interfaces.

To further explore the consistency of results for both interfaces, we performed a two sample t-test for each question given in the Stripe and the 5-point Likert interface. The variances for all 10 SUS questions appear statistically consistent within each pair of variables for a given question. Thus, we rejected the null hypothesis of unequal variances for each pair of question variables for $\alpha = 0.01$. Consequently, we performed the analysis of variance for the cumulative scores. The variances for both services appear not to differ significantly. No group has marginal means statistically different from the other for $\alpha = 0.01$. The ANOVA shows no statistical differences between values gathered by both interfaces for both services, Gmail and Microsoft Exchange ($p = 0.32$). Furthermore, the ANOVA shows no statistical difference in variances between both services ($p = 0.44$).

The study also included questions on how both interfaces compare in terms of intuitiveness and comprehension, time perception and difficulty. The results showed that the participants found the Stripe more intuitive and comprehensible with the average values of 4.54 on a 7-point scale. In terms of time perception, the Stripe was rated as slightly less demanding than the 5-point Likert scale with an average value of 3.79 (Figure 2). Finally, the participants rated the Stripe interface as slightly easier for expressing opinions, with the average score of 3.42 on a 7-point scale (1 - the Stripe was easier than the Likert interface, 7 - the Stripe was more difficult than Likert).

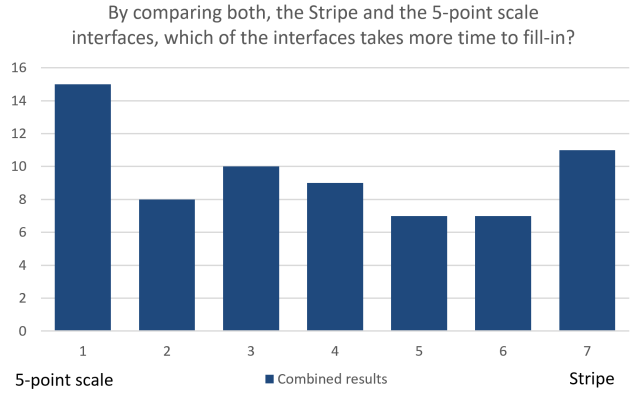

Figure 2: Comparison of the Stripe and Likert scale shows that the 5-point scale is perceived as more time consuming.

Overall, the results favor the Stripe interface over the 5-point Likert scale: mostly in terms of intuitiveness, slightly less in terms of simplicity of expressing an opinion. An unexpected result was the finding that participants found the 5-point Likert scale, which was implemented with standard radio buttons, as slightly more time consuming than the graphical and interactive Stripe interface. This result is at odds with research that shows that graphical scales take more time to fill-in than radio button scales, which leads us to the conclusion that the participants found the Stripe interface more enjoyable and engaging than the standard Likert scale interface.

6. CONCLUSION

Usability questionnaires like the SUS are still widely based on the traditional n-point Likert scale, which has also been adopted in online surveys due to its simple implementation with HTML radio buttons. And while there is some existing research that compares Likert scales with continuous scales, most research focuses on time performance and reliability of results. For this reason, we decided to conduct a user study that would also evaluate the usability of an alternative user interface for online usability surveys. In addition to providing the benefits of a continuous scale, the proposed

Table 1: Comparison of average SUS scores and their deviations for the 5-point Likert scale and Stripe interfaces.

	Gmail		Exchange	
User interface	Avg. SUS score	σ of SUS scores	Avg. SUS score	σ of SUS scores
(1) 5-point Likert SUS	79.88	18.03	72.03	20.32
(2) Continuous Stripe SUS	79.02	16.61	70.03	21.44
(3) 5-point Stripe (quantized)	80.55	17.27	70.37	22.36
Δ 1 vs. 2	0.86	1.42	2.00	1.12
Δ 1 vs. 3	1.67	0.76	1.66	2.05

Stripe scale also aims to provide a more compact alternative that could work well across different devices and smaller screens.

The results of the user study, which was conducted online on two separate groups of participants, show that both the Stripe and Likert scale interfaces provide consistent SUS scores, confirming the Stripe interface as a viable alternative. The Stripe interface was favored in terms of intuitiveness and chosen as easier for expressing opinions. The most surprising result was seeing the Stripe interface score slightly better in terms of perceived time. While surveys based on graphical interfaces like the Stripe usually take more time to complete, the participants in our study rated the standard 5-point Likert scale as taking slightly more time. Overall, our results show that the Stripe interface was the participant's favorite interface across all tested usability aspects.

7. REFERENCES

[1] B. Blažica and J. R. Lewis. A slovene translation of the system usability scale: The sus-si. *International Journal of Human-Computer Interaction*, 31(2):112–117, January 2015.

[2] J. Brooke. Sus: A 'quick and dirty' usability scale. *Usability Evaluation in Industry*, 189(164):7, 1996.

[3] C. Cook, F. Heath, R. L. Thompson, and B. Thompson. Score reliability in webor internet-based surveys: Unnumbered graphic rating scales versus likert-type scales. *Educational and Psychological Measurement*, 61(4):697–706, August 2001.

[4] M. P. Couper, R. Tourangeau, F. G. Conrad, and E. Singer. Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, 24(2):227–245, 2006.

[5] R. Cummins and E. Gullone. Why we should not use 5-point likert scales: The case for subjective quality of life measurement. In *Proceedings, Second International Conference on Quality of Life in Cities*, pages 74–93, Singapore, 2000. National University of Singapore.

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

[7] F. Funke. A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, 34(2):244–254, April 2016.

[8] F. Funke and U. D. Reips. Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, 24:310–327, 2012.

[9] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50:904–908, 2006.

[10] S. G. Hart and L. E. Staveland. *Human Mental Workload*, chapter Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, pages 139–183. North Holland Press, Amsterdam, 1988.

[11] M. Hassenzahl, M. Burmester, and F. Koller. Attrakdiff: A questionnaire to measure perceived hedonic and pragmatic quality. *Mensch & Computer*, 2003.

[12] Iso/Iec. *ISO/IEC, 9241-11 Ergonomic requirements for office work with visual display terminals (VDT)s - Part 11 Guidance on usability*, 1998.

[13] A. Madan and S. K. Dubey. Usability evaluation methods: a literature review. *International Journal of Engineering Science and Technology*, 4:590–599, 2012.

[14] M. Pesek, P. Godec, M. Poredos, G. Strle, J. Guna, E. Stojmenova, and M. Marolt. Introducing a dataset of emotional and color responses to music. In *Proceedings of the International Society for Music Information Retrieval, Taipei*, pages 355–360, 2014.

[15] M. Pesek, P. Godec, Poredoš, G. M. Strle, J. Guna, E. Stojmenova, and M. M. Capturing the mood: Evaluation of the moodstripe and moodgraph interfaces. In *Management Information Systems in Multimedia Art, Education, Entertainment, and Culture (MIS-MEDIA), IEEE Internation Conference on Multimedia & Expo (ICME)*, pages 1–4, 2014.