Beyond Aesthetics: Evaluating Response Widgets for Reliability & Construct Validity of Scale Questionnaires

Habiba Farzand
h.farzand.1@research.gla.ac.uk
University of York
United Kingdom
University of Glasgow
United Kingdom

David Al Baiaty Suarez
2474364a@student.gla.ac.uk
University of Glasgow
United Kingdom

Thomas Goodge
2605440g@student.gla.ac.uk
University of Glasgow
United Kingdom

Shaun Alexander Macdonald
shaun.macdonald@glasgow.ac.uk
University of Glasgow
United Kingdom

Karola Marky
karola.marky@rub.de
Ruhr University Bochum
Germany

Mohamed Khamis
mohamed.khamis@glasgow.ac.uk
University of Glasgow
United Kingdom

Paul Cairns
paul.cairns@york.ac.uk
University of York
United Kingdom

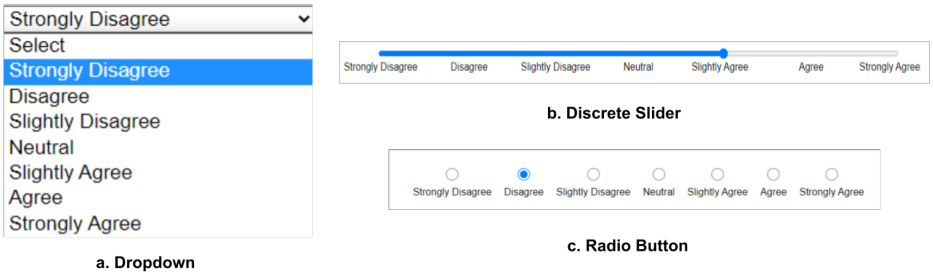

Figure 1: This paper investigates UI response widgets – Dropdown, Discrete Slider, and Radio Button – for their impact on a scale questionnaire's overall reliability and validity using 7-point Likert scales and user experience.

ABSTRACT

Scale questionnaires are psychometric tools that capture perspectives and experiences. Consequently, these tools need to be reliable and valid. In this paper, we investigate the impact of response widgets - the UI elements that allow users to answer scale items - on the overall scale reliability and construct validity of three varied length scale questionnaires in a user study (N=30). Our results reveal that optimum reliability was achieved using radio buttons and dropdowns in all varied-length questionnaires. Further, valid results were produced utilising the slider and dropdown. No significant differences were found in time consumption, but click count was significantly higher with dropdown. Radio buttons scored lower in format satisfaction than others, and dropdown was the least effective in ease of selection and quick completion. In light of these results, we conclude that response widgets are more than just aesthetics and should be selected as per the researcher's aims.

CCS CONCEPTS

• **Human-centered computing**; • **Human Computer Interaction**; • **Interaction Techniques**;

KEYWORDS

UI response styles, scale questionnaires, reliability, validity, user experience

ACM Reference Format:
Habiba Farzand, David Al Baiaty Suarez, Thomas Goodge, Shaun Alexander Macdonald, Karola Marky, Mohamed Khamis, and Paul Cairns. 2024. Beyond Aesthetics: Evaluating Response Widgets for Reliability & Construct Validity of Scale Questionnaires. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May*

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650751

11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3650751

1 INTRODUCTION

Scale questionnaires are frequently administered on a large scale to collect information from participants in HCI and related research fields such as [4, 12]. The questionnaire life cycle comprises three stages: (1) design, (2) distribution, and (3) evaluation. The designing phase includes taking care of user interface elements such as the style of interaction [29], and the distribution phase involves utilizing different methods to invite participants to complete the questionnaires, for example, by providing incentives and reminders. Lastly, in the evaluation phase, the data gathered is analysed using different quantitative or qualitative methods depending on the researcher's aim. In particular, the design phase, which includes taking care of the user interface elements, is the most pivotal, as it directly impacts the quality of responses and errors in this phase propagate to all later phases. Among the various aspects of the user interface, one of the most important decisions is to choose the most appropriate response widget, such as radio buttons or sliders.

Prior work has investigated several aspects of the transition from paper to online questionnaires, such as response rates [24], acquiescence bias [20], response biases [17], and different questionnaire modes [5]. Among this research, it has also been shown that the usability of *response widgets* – the UI elements that study participants use to select their answers – impacts the responses of participants [29]. Based on that, this paper takes a closer look at different response widgets in the context of questionnaire design. In this scope, we also investigate two important research constructs in questionnaire design: reliability and validity. Reliability refers to the extent to which the items of a scale are consistent with each other [7], while validity indicates if the questionnaire indeed measures the construct that it aims to [6, 19]. Consequently, reliability and validity ensure the integrity and quality of the measurement tool. While prior work has investigated the impact of individual scale items [3], this paper examines the impact of response widgets on the overall reliability and validity of the measured construct. For this, we explore our first research question:

RQ1: *How does using response widgets offering different selection methods impact the reliability and validity of a scale questionnaire?*

To complement the findings of selecting the most appropriate response widget, we additionally explore the user experience of the widgets. This leads us towards our next research questions:

RQ2: *Which response widgets require more user time and effort to complete a questionnaire?*

RQ3: *How does the user experience vary across interchangeably used response widgets?*

To answer our research questions, we conducted an in-the-wild study, allowing participants to fill out the questionnaire in the naturalistic setting with 30 participants who filled out validated and reliable scales from the literature. We investigated three styles of response widgets for Likert scales in online questionnaires that are widely used in HCI research [18]: (1) radio buttons, (2) discrete sliders, and (3) dropdowns.

Our results show that radio buttons and dropdowns are appropriate for optimum reliability as a single selection response widget in any length scale questionnaire. Valid results can be produced using any widget style in long and medium-length questionnaires. However, care should be taken when using radio buttons in small-length questionnaires. All widgets acquire equal time to complete the questionnaire regardless of the length of the questionnaire. Still, dropdowns require the most clicks, which may lead to selection errors and cause user fatigue. Sliders and radio buttons are seen as offering quick completion. All are seen as equal in terms of understanding and clarity of meaning. Based on these findings, we recommend selecting a response widget with the study's focus in perspective, as each widget has its pros and cons. On the whole, this paper makes the following core contributions:

- We contribute an evaluation of the commonly used response widgets, radio buttons, sliders, and dropdowns focusing on a scale questionnaire's reliability and validity in a user study with 30 participants on three standard 7-point Likert scale questionnaires.
- To strengthen the persuasiveness of selecting the appropriate response widget considering scale metrics, we assess and compare the user experience of response widgets on questionnaires of varied lengths, emphasising five attributes: ease of selection, format satisfaction, quick completion, understanding, and clarity in the meaning of three lengths of standard scale questionnaires (short (19-items), medium (26-items), and long (38-items)).

2 BACKGROUND & RELATED WORK

2.1 Reliability & Validity in Designing Questionnaires

When questionnaires are administered electronically, there are various issues to be dealt with while aiming for reliability and validity, for example, whether all response items should be labelled or just the endpoints. It has been demonstrated that there is no difference between a) labelling each response item and b) labelling the endpoints only [11]. Full labelling and having a midpoint on a scale with five response items is the most recommended [6]. However, reliability is not affected by the number of scale points [16]. Consequently, seven-point and five-point scales result in the same reliability. Validity is also independent of the number of steps involved in the scale [16]. Hence, any numbered-point scale can be used per the researcher's choice.

The effect of response option orientation and directionality, such as vertical, horizontal, ascending or descending, on internal consistency and factorial validity has been investigated [21]. The results showed that internal consistency is consistent in all configurations and that altering the orientation of response options did not impact factorial validity. Further, the impact of radio buttons and sliders on reliability in a probability-based panel in Norway was explored [3]. It also included exploring the effects of smartphone and PC/tablet responses. The study showed that radio buttons and sliders have a similar measurement quality and can be used interchangeably [3]. However, if the device is not a smartphone, a marker indicating the slider should be placed in the middle rather than on the left side. This study, however, focused on only two commonly used styles of widgets and explored the reliability and validity of individual questionnaire items. Moreover, the study was restricted to Norway;

only two questionnaires from a specific field were investigated. Since questionnaires are designed to measure different concepts in different fields, there is a need to confirm if guidelines about designing questionnaires are consistent across all disciplines and geographical regions.

To sum up, the literature recommends a method for designing a questionnaire, i.e. full labelling of response items [11], including a midpoint and five response items in total for each statement [6, 16], using Item Response Theory for ordinal data achieved from Likert statements [6], opting for any response option configuration such as horizontal, vertical, bidirectional, ascending or descending [21], and various interaction styles for response option selection [29]. However, it is still unclear how different response widgets may impact the reliability and validity of the complete questionnaire.

2.2 Styles of Response Widgets

Selecting the most appropriate response widget for the statements is another essential step of questionnaire design. While numerous widgets exist, sliders, dropdowns, and radio buttons are the most frequently used. Comparing these generally, sliders offer multiple responses but may be less suitable for labelling. Long drop-downs allow respondents to see and scroll through all available options but may also be impractical due to the requirements use of screen retail and longer navigation by users; radio buttons are simple and conventional. Different styles might consume various amounts of time to get selected by participants and may demand different user efforts. Increased user effort may lead to more errors.

Radio buttons and sliders have been studied and compared extensively. Sliders are more challenging to use than radio buttons and, thus, more inclined towards response biases [27]. The initial position of sliders handles impacts responses, especially on smartphones [22]. Radio buttons have also been compared to pictorial answer categories, such as smileys, which performed poorly by comparison [28]. Pictorial answer categories bring along the challenge of cognitive load and usability, impacting their use. Different slider styles have been proposed and investigated for different purposes, such as entering uncertain data. Examples include fixed/flexible range sliders and flexible and advanced flexible range best estimate sliders [14]. Furthermore, the precision of entered values is affected by the orientation and visual style when the slider is presented on a touchscreen [9]. Moreover, visual appearance, such as decorations, for example, tick marks or labels, along the slider, contributes to response biases [23]. Despite this, generic-styled sliders like those investigated in this work remain the most commonly used.

In sum, prior work investigating sliders has focused on how the initial positioning of the handle, the presence of numeric labels, response rate, response time, and visual appearance impact response biases and participant confidence. This paper, in contrast, explores how reliable and valid results can be produced using different response widget styles, specifically comparing radio buttons, sliders, and dropdowns.

3 METHODOLOGY

We conducted a study with 30 participants to answer the research questions. The response widget presented served as the independent variable, resulting in three conditions: (1) radio buttons, (2) discrete sliders, and (3) dropdowns.

Design. To ensure participants were not biased, we adopted a simple cover story where participants were asked to play a game and then answer a questionnaire. For the game, we chose *Fruit-Salad Slice* [15] due to its ease of play and requiring minimal training. To heighten validity, we selected three well-established and validated gaming evaluation questionnaires, namely the Game Engagement Questionnaire (GEQ) [4], the Gameplay Scale Questionnaire (GSQ) [25], and Gaming Experience Questionnaire (CEGEQ) [8]. Each features a different number of items (19, 26, and 38), allowing us to study the impact of questionnaire length. We opted for a similar methodology used in related prior work to report the impact on reliability with different presentation styles of responses [21].

We constructed one questionnaire per response widget, leading to nine total questionnaires: three versions of the GEQ, GSQ, and CEGEQ. We deployed each questionnaire using a 7-point scale because (1) 7-point scales were used in the original study, and (2) 5-point and 7-point scales both produce valid results [16]. Participants answered the questionnaires in counterbalanced order (Latin square) to mitigate sequential effects. The questions in each questionnaire were also randomized to mitigate these effects further. We further counterbalanced the combinations of widgets and questionnaires. In summary, participants experienced all response widgets by completing three scale questionnaires. Alongside the scale questionnaires, a further short questionnaire was administered that assessed participant experience of filling out the scale questionnaires, comprising five questions and referred to as the Response Widget Questionnaire (RWQ) (see Appendix A). The questions used in the RWQ were inspired by previous work [14]. The RWQ also used a 7-point Likert scale similar to the scale questionnaires. RWQ was deployed after each main questionnaire and utilised the same response widget as the previous scale questionnaire not to impact user experience. Participants were allowed to use any device to participate in the study. Completion time and the number of clicks were recorded to investigate RQ2. Before running the study, we pilot-tested it with three participants and ensured it ran smoothly.

Procedure. Participants were informed that the study aimed to measure gameplay elements. The actual aim of the study was kept hidden to mitigate response biases. The participants were then asked to sign the consent form. Participants played the game Fruit-Salad Slice [15] for three minutes. After the game session, participants were asked to fill out three main questionnaires, each followed by the RWQ that assessed the user experience of the widgets. After completing the six tasks mentioned above, a demographic survey was provided, a final debriefing section, and a comment of appreciation. We developed a web application using Django that included the game and follow-up questionnaires. Respondents were then informed that the real aim of the experiment was to evaluate and compare different response widgets. Our study received Ethical Approval at our institute. Participants optionally participated in a one out of five £20 Amazon voucher raffle.

Recruitment & Participants. We opted for convenience sampling, and the study was advertised via email distribution lists and university platforms. N=30 participants took part in the survey. Participants reported ages ranged from 21 to 47 years (mean = 26.0, SD = 7.7), with 20 identifying as male, nine as female, and one as "other". Respondents were asked to indicate their preferred device when completing questionnaires. All participants chose either a laptop (66.7%) or a desktop PC (33.3%), with all respondents except one having at least some experience completing questionnaires. This was also reflected in the participants' device selection for the study, where all participants chose a desktop or a PC to participate in the study, and none used a smartphone. Moreover, 20% of the sample reported completing questionnaires frequently. 43.3% of participants had completed a bachelor's degree, and 20.0% had finished high school. This was closely followed by master's and doctorate degrees, with 16.7% and 13.3%, respectively.

Limitations. We opted for an in-the-wild survey that allowed participants to use their preferred device in their naturalistic environment. Though all our participants used either a desktop or a laptop to participate, further studies comparing smartphones with desktops/PCs should be conducted. Second, our study serves as an exploratory study for the impact of widgets of scale reliability and validity. Further studies involving users with diverse backgrounds and a large sample size should be conducted. Third, compared with the slider and radio button, the dropdown menu has a larger space to click on. This may have impacted how users responded to the questionnaires. Next, each participant filled in one version of all three questionnaires, which might have caused fatigue and impacted user experience. Moreover, while we only investigated three commonly used widgets, studies examining and comparing more widgets should be conducted. Despite these limitations, our study takes a crucial step in evaluating the impact of response widgets on questionnaire reliability, construct validity, and user-centred evaluation. Future research in this field should consider and address the limitations highlighted by (1) conducting studies that explore questionnaires from other disciplines, (2) controlling the prior questionnaire experience of participants, and (3) recruiting a larger group of participants with diverse backgrounds.

4 ANALYSIS

To determine the internal consistency as the original studies used this method for the scales we used, we calculated reliability using Cronbach's Alpha (α) [6, 10]. A reliability coefficient of 0.70 or higher is typically considered *"acceptable"* [10].

Validity refers to whether the measured concept fully corresponds to the intended construct [6] and is the foundation of any questionnaire [7]. This paper measures validity by checking if the scales with different response widgets received similar responses by calculating scale scores ranging from one to seven. Similar mean values of versions of the same questionnaire indicate the validity, and the items correlate with some reference criteria of what the concept means [6]. As the scales used are validated and well-established questionnaires, we assume they will produce valid results.

Finally, linear mixed-effects models were fitted to the data to predict the effect of each widget on the overall score, as well as for responses to each of the five RWQ questions. Additionally, a generalised linear mixed effects model was fitted to predict the effect of widgets on click count. These models were estimated using Maximum Likelihood (ML) and the bobyqa optimizer. Confidence Intervals (CI = 95%) and p-values were computed using a Wald t-distribution approximation. In keeping with guidance from Barr et al. [2], the simplest models that explained the greatest variation were retained and presented here. More complex models that did not significantly improve the variation explained are omitted here for relevance, as they do not provide any greater explanation of the data than the simpler models. Similarly, models that did not reveal any significant effects were also omitted for conciseness.

5 RESULTS

5.1 Internal Consistency Reliability

Table 1 shows the results of reliability tests and compares these results with the reliability achieved in the original study of each scale. Radio buttons and dropdowns have alpha values comparable to the original scale questionnaire studies, and the 95% CI includes that value. No formal test was applied to these results because there are no commonly accepted tests for Cronbach's Alpha. The results suggest that radio buttons and dropdowns can be used for optimum reliability. However, discrete sliders are observed as less reliable.

Key Takeaway: The reliability test suggests that radio buttons and dropdowns can be reliably used.

5.2 Construct Validity

A between-group ANOVA comparing the effect of the Widget on the Score for each questionnaire found that the Widget had a significant effect for the GEQ ($F(2) = 7.914$, $p = 0.002$). Post hoc Tukey tests showed that the Radio Button Widget scores were significantly higher than the Dropdowns ($p = 0.003$) and the slider (GEQ $p = 0.013$). However, there were no significant differences in GSQ or the CEGEQ scores for any of the Widgets.

A linear mixed model was fitted to predict the main effect of Widget on Score across all questionnaires, including participant and Questionnaire Type as random effects. The model's total explanatory power was substantial (conditional $R^2 = 0.71$). Within this model, Scores for the Dropdown Widget were significantly higher than the Radio Button Widget (-0.32, $p = 0.009$), but there were no significant differences with other comparisons.

Key Takeaway: Similar results were obtained in the CEGEQ and GSQ questionnaire versions, suggesting that valid results can be produced using any widgets in medium and long-length questionnaires. However, results differed for the radio buttons version of GEQ from the dropdown and slider version of GEQ. This points out that the role of radio buttons in validity requires further investigation.

5.3 Response Time & Click Counts

The time taken to complete the questionnaire with each response widget was recorded in seconds to see if one response widget took more time than another, as this would help save the respondents' time. The time taken was documented for each questionnaire individually. The effect of the Widget on time taken was analysed

Questionnaire	Original Study	Widget Style	Our Study		
			Lower Bound	α	Upper Bound
CEGEQ (38-item questionnaire)	0.79	Radio Button	0.653	0.839	0.952
		Dropdown	0.521	0.777	0.934
		Discrete Slider	0.43	0.735	0.921
GSQ (26-item questionnaire)	0.9	Radio Button	0.616	0.823	0.947
		Dropdown	0.537	0.787	0.937
		Discrete Slider	0.346	0.699	0.91
GEQ (19-item questionnaire)	0.85	Radio Button	0.557	0.798	0.94
		Dropdown	0.657	0.844	0.954
		Discrete Slider	0.261	0.663	0.9

Table 1: The Table shows the Cronbach's Alpha (α) values for each scale questionnaire and widget with the three widget styles: Radio Buttons, Dropdown, and Discrete Slider.

using a between-groups ANOVA and a linear mixed-effects model, but there were no significant differences. Next, we took notes of click counts.

Click counts represented the number of clicks made while filling out the three game-play-related scale questionnaires. A between-group ANOVA compared the effect of the Widget on the Click Count for each Questionnaire Type. There was a significant effect of the Widget on the number of clicks for the GEQ ($F(2) = 96.78$, $p < 0.001$), GSQ ($F(2)= 190.1$, $p < 0.001$) and CEGEQ ($F(2) = 111.6$, $p < 0.001$). As expected, post hoc Tukey tests showed that the click counts from using the dropdowns were significantly higher in all questionnaires than both the slider (GEQ $p < 0.001$, GSQ $p < 0.001$, CEGEQ $p < 0.001$) and the radio button (GEQ $p < 0.001$, GSQ $p < 0.001$, CEGEQ $p < 0.001$). However, there were no significant differences in the number of clicks between the radio button menu and the slider on any questionnaire type.

A generalised linear mixed model was fitted to predict the main effects of Widget on Click Count across all Questionnaires, including participant (b_{P}) and Questionnaire Type (b_{h}) as random effects..

The model's total explanatory power was substantial (conditional $R^2 = 0.88$). Within this model, Scores for the Dropdown Widget were significantly higher than the Radio Button Widget (-0.71, $p < 0.001$) and the Slider Widget (-0.76, $p < 0.001$). There was no significant difference between the Slider and Radio Button widgets.

Key Takeaway: All widgets require a similar amount of time to fill in the three varied lengths of questionnaires. However, dropdowns require the most number of clicks in all three varied-length questionnaires.

5.4 User Experience of Response Widgets

Following the scale questionnaires, participants were presented with an additional questionnaire (Response Widget Questionnaire (RWQ)) to capture their interaction experience with the response widgets. Inspired by previous work [28], the interaction experience was assessed on five attributes on a 7-point scale: ease of selection, clarity in meaning, understanding, format satisfaction, and quick completion. The reader is referred to Appendix A for the Response Widget Questionnaire. Linear mixed-effects models were fitted to predict the main effects of each UX question in the RWQ, including Participant ID (b_{h}) as a random intercept for each group. These models were fitted using the procedure outlined above in Section 4.

For **Format Satisfaction**, the model's total explanatory power was substantial (conditional $R^2 = 0.39$). Within this model, there were significant effects of the Widget, with radio button ratings being significantly lower than both the Dropdown (-0.9,$p = 0.007$) and Slider (-0.933, $p = 0.006$) widgets. However, there was no difference between the Slider and Dropdown ratings. For **Ease of Selection**, the model's total explanatory power was substantial (condition $R^2 = 0.27$). Within this model, there was a significant effect of the Widget, with the Dropdown widget being rated significantly lower than the Radio button (-1.1, $p = 0.002$). However, there was no difference between the Slider and either Dropdown or Radio button menu ratings. For **Quick Completion**, the model's total explanatory power was substantial (conditional $R^2 = 0.46$). The Widget had significant effects, with Dropdown ratings being significantly lower than the Radio Button (0.9, $p = 0.004$) and Slider (0.933, $p = 0.003$) widgets. However, there was no difference between Slider and Radio Button ratings or Dropdown and Slider ratings. Models were also fitted for **Understanding** (conditional $R^2= 0.21$) and **Clarity in Meaning** (conditional $R^2= 0.36$), but these found no significant differences between ratings for any of the widgets and so are omitted here for conciseness.

Key Takeaway: Radio buttons were rated with less format satisfaction overall and dropdown were seen as offering least ease of selection. Both the Radio button and Slider widgets are rated higher for quick completion than the dropdowns. All three widgets are perceived equally for understanding and clarity in meaning.

6 DISCUSSION & FUTURE WORK

As a Late-Breaking Work, this research presents preliminary findings on how response widgets play a role beyond just aesthetics and hold crucial importance in determining the reliability and validity of scale questionnaires. This research paves the way for future research needed in this direction to extend and validate the results on a large scale using scale questionnaires from multiple fields.

A vital finding of this research is that radio buttons and dropdowns had acceptable levels of reliability across all questionnaires. However, great care should be taken when considering sliders as

they produce less reliability scores than other widgets. With respect to validity, radio buttons cannot be confidently used, and their role requires further investigation of the impact on validity. When examining the user experience performance of each widget, however, we found that each brought its own strengths and weaknesses for future researchers to consider when designing and deploying their instruments. For example, the discrete slider was less reliable across two questionnaires, although users perceived it as satisfying, easy, and quick to use. On the other hand, the radio buttons had significantly lower format satisfaction than the other two widgets. This may be due to their very simple design compared to other response widgets. Finally, the dropdowns were no less valid or reliable, but users perceived them as taking less time to complete despite there being no significant difference in time taken between widgets. Additionally, users felt the dropdowns were less easy to use than the Radio button, which aligns with a significantly higher click count.

Our results suggest that researchers should select the widget depending on the study's aims. For example, for surveys that are long or prioritise user experience, consider avoiding dropdowns. Meanwhile, researchers seeking to fully maximise reliability may wish to avoid discrete sliders. Our work opens up many other directions for future research. One exciting direction is how the user's input selection device or method impacts the widget choice regarding reliability and construct validity. Recent advances in technology have made gaze-based selections. More and more technologies have been using gaze-based interaction [1, 26] and touchless gesture interaction [13]. This raises the question of investigating novel input selection methods.

Q1: *How do the input selection methods, such as gaze-based interaction and touchless gesture interaction, influence the widget choice for optimum reliability and construct validity?*

Second, as seen in our study, questionnaires can be of varied lengths. Therefore, it is important to investigate how many questions per page should be displayed to participants and if they impact the scale questionnaire's reliability, validity, or user experience. This directs us to another research question for future investigation, i.e.,

Q2: *How does the presentation of questionnaire items impact scale reliability and validity?*

Lastly, in this study, we investigated three scale questionnaires that relate to digital games. However, it would be interesting to explore if the results of this study can be generalized to other domains as well. Further, in our study all participants used either a desktop or a PC to fill in the questionnaires. A cross-device comparison would yield interesting results and a recommendation for the best device to use for filling out questionnaires. This guides us to the next research question, i.e.,

Q3: *How does the impact of response widgets on scale reliability and validity vary as we move across different devices and questionnaire domains?*

7 CONCLUSION

This work experimentally evaluated online questionnaire response widgets, namely radio buttons, discrete sliders, and dropdowns, with 30 participants for their impact on the complete questionnaire's reliability and construct validity. We evaluated the widgets with three scale questionnaires (short, medium, and long-length) using a 7-point Likert scale. The results show that all questionnaires of varied lengths can achieve optimum reliability using radio buttons and dropdowns. Sliders should be handled with utmost care as they produce the lowest reliability score compared to other widgets. Valid results can be produced with dropdowns and sliders for any questionnaire size. All widgets consume equal time to complete the questionnaire. However, dropdowns require most click counts, which may cause fatigue and tiredness in users. Complimenting this finding, dropdown was also seen as the lowest metric of ease of selection and quick completion. Radio buttons were seen as offering the least format satisfaction. However, they outperform other user experience metrics and can be recommended for use. Last but not least, sliders outperformed all five user experience metrics. Based on the results presented in this paper, this paper raises awareness of possible methodological concerns with scale widgets that motivate further consideration. Overall, this paper recommends that the selection of the response widget should be based on the researcher's study aim, as every widget comes with its own merits and demerits.

ACKNOWLEDGMENTS

This publication was supported by an Excellence Bursary Award from the University of Glasgow and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA - 390781972.

REFERENCES

[1] Yasmeen Abdrabou, Mariam Mostafa, Mohamed Khamis, and Amr Elmougy. 2019. Calibration-free text entry using smooth pursuit eye movements. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–5.

[2] Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68, 3 (2013), 255–278.

[3] Oriol J Bosch, Melanie Revilla, Anna DeCastellarnau, and Wiebke Weber. 2019. Measurement reliability, validity, and quality of slider versus radio button scales in an online probability-based panel in Norway. *Social Science Computer Review* 37, 1 (2019), 119–132.

[4] Jeanne H Brockmyer, Christine M Fox, Kathleen A Curtiss, Evan McBroom, Kimberly M Burkhart, and Jacquelyn N Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video gameplaying. *Journal of Experimental Social Psychology* 45, 4 (2009), 624–634.

[5] Vesna Bucevska. 2000. Designing a web versus a paper questionnaire-some general and special issues. *Public Opin Q* (2000), 1–8.

[6] Paul Cairns. 2019. *Doing better statistics in human-computer interaction.* Cambridge University Press.

[7] Paul Cairns, M Soegaard, and RF Dam. 2016. Experimental methods in human-computer interaction. *Encyclopedia of Human-Computer Interaction* (2016).

[8] Eduardo H Calvillo-Gámez, Paul Cairns, and Anna L Cox. 2015. Assessing the core elements of the gaming experience. *Game user experience evaluation* (2015), 37–62.

[9] Ashley Colley, Sven Mayer, and Niels Henze. 2019. Investigating the Effect of Orientation and Visual Style on Touchscreen Slider Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–9.

[10] Lee J Cronbach. 1950. Further evidence on response sets and test design. *Educational and psychological measurement* 10, 1 (1950), 3–31.

[11] Paul N Dixon, Mackie Bobo, and Richard A Stevick. 1984. Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychological Measurement* 44, 1 (1984), 61–66.

[12] Habiba Farzand, Karola Marky, and Mohamed Khamis. 2024. Out-of-Device Privacy Unveiled: Designing and Validating the Out-of-Device Privacy Scale

(ODPS). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642623

[13] Euan Freeman. 2022. Ultrasound Haptic Feedback for Touchless User Interfaces: Design Patterns. In *Ultrasound Mid-Air Haptics for Touchless Interfaces*. Springer, 71–98.

[14] Miriam Greis, Hendrik Schuff, Marius Kleiner, Niels Henze, and Albrecht Schmidt. 2017. Input controls for entering uncertain data: Probability distribution sliders. *Proceedings of the ACM on Human-Computer Interaction* 1, EICS (2017), 1–17.

[15] iDev.Games. 2023. *Fruit Salad Slide HTML5 Game - on iDev Games*. Retrieved March 30, 2023 from https://idev.games/embed/fruit-salad-slice

[16] Jacob Jacoby and Michael S Matell. 1971. Three-point Likert scales are good enough.

[17] Rob Johns. 2010. Likert Items and Scales. *Survey Question Bank: Methods Fact Sheet* 1 (2010).

[18] Maurits Clemens Kaptein, Clifford Nass, and Panos Markopoulos. 2010. Powerful and consistent analysis of likert-type ratingscales. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2391–2394.

[19] Paul Kline. 2000. *A psychometrics primer*. free Assn books.

[20] James R Lewis. 2018. Comparison of item formats: agreement vs. item-specific endpoints. *Journal of Usability Studies* 14, 1 (2018), 48–60.

[21] Hotaka Maeda. 2015. Response option configuration of online administered Likert scales. *International Journal of Social Research Methodology* 18, 1 (2015), 15–26.

[22] Angelica M Maineri, Ivano Bison, and Ruud Luijkx. 2021. Slider bars in multi-device web surveys. *Social Science Computer Review* 39, 4 (2021), 573–591.

[23] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5421–5432.

[24] Duncan D Nulty. 2008. The adequacy of response rates to online and paper surveys: what can be done? *Assessment & evaluation in higher education* 33, 3 (2008), 301–314.

[25] Mark James Parnell, N Berthouze, and D Brumby. 2009. Playing with scales: Creating a measurement scale to assess the experience of video games. *University College London, London, UK* (2009), 35.

[26] Sheikh Rivu, Yasmeen Abdrabou, Thomas Mayer, Ken Pfeuffer, and Florian Alt. 2019. GazeButton: enhancing buttons with eye gaze interactions. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–7.

[27] Vera Toepoel and Frederik Funke. 2018. Sliders, visual analogue scales, or buttons: Influence of formats and scales in mobile and desktop surveys. *Mathematical Population Studies* 25, 2 (2018), 112–122.

[28] Vera Toepoel, Brenda Vermeeren, and Baran Metin. 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 (2019), 57–74.

[29] Muhsin Ugur, Dvijesh Shastri, Panagiotis Tsiamyrtzis, Malcolm Dcosta, Allison Kalpakci, Carla Sharp, and Ioannis Pavlidis. 2015. Evaluating smartphone-based user interface designs for a 2d psychological questionnaire. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 275–282.

A RESPONSE WIDGET QUESTIONNAIRE (RWQ)

The following questions were presented to participants after filling out each scale questionnaire. Participants were asked to rate their experience on the following statements on a 7-point scale ranging from "strongly disagree" to "strongly agree".

(1) *The response options were clear to me.*
(2) *It was easy to select a response for each item.*
(3) *I am satisfied with the format of the scale used in the questionnaire.*
(4) *It was easy to understand the response options of the questions.*
(5) *I was able to complete the questionnaire quickly.*