Choose Popovers over Buttons for iPad Questionnaires

Kevin Gaunt[1], Felix M. Schmitz[2] and Markus Stolze[1]

[1] Institute for Software, University of Applied Science Rapperswil, CH-8640 Rapperswil-Jona, Switzerland
[2] Institute of Medical Education, Medical Faculty of the University of Bern, CH-3010 Bern, Switzerland

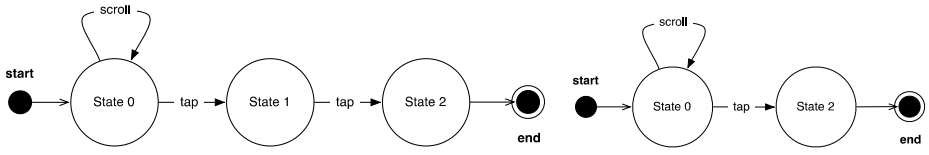
{kevin.gaunt, markus.stolze}@hsr.ch
felix.schmitz@iml.unibe.ch

Abstract. When designing questionnaires for iPad an important design decision is whether to use popover listings or button listings for representing single-choice selections. In this paper we examined effects of each listing method on performance and subjective preferences when performing a non-linear selection task. A quantitative experiment (N = 39) with the two within-factors (1) listing method (popover versus button) and (2) task completion time (15s versus 7s versus 5s) was conducted. Results show subjects performing significantly better when using popovers, which they also strongly preferred. We attribute this to lower extraneous cognitive load and shorter forms, ultimately requiring less scrolling. Results also show the expected effect of task completion time on performance: the longer the allotted time, the higher the test scores.

Keywords: popover listing, button listing, single-choice questionnaires, cognitive load, performance, iPad.

1 Introduction

Cherished or despised, questionnaires are an accepted assessment instrument across the globe. Increasingly paper-based evaluation forms will be replaced by their electronic equivalents. With its low weight, long battery life, and high interactivity iPad offers an important platform for these electronic questionnaires. It also offers a new type of user interface widget: the popover [1]. The popover (cf. Section 3.3 for an example) takes after other commonly used widgets, such as drop-down menus, dialog boxes and inspector palettes. Historically, when designing electronic questionnaires, either drop-down menus or multiple radio buttons were used to represent the available choices [2, 3]. With iPad the question shifts to whether these choices should be contained within a popover rather than being represented by multiple button controls listed directly within the form. Henceforth, we will label the use of a single popover as "Popover Listing" (PL) and the use of multiple buttons as "Button Listing" (BL).



(a) "Popover Listing" (PL) (b) "Button Listing" (BL)

Fig. 1. Steps needed to select a choice depending on listing type. State 0: No Selection; State 1: Displaying Popover; State 2: Choice Selected.

Figure 1 identifies the difference in interaction between the two designs. Figure 1a shows the steps needed to make a choice within a questionnaire when the PL design is applied. Figure 1b shows the same, this time for listing type BL. It can be observed, that by choosing BL one step can be eliminated from the task. Thus, by using iPad as a survey instrument (i.e. single-choice questionnaires), PL will require more "tap"-interactions when choosing a value. Additionally, by using BL all of the questions' choices are visible at one glance. This might indicate BL to be the better performing design. However, BL also increases the chance of spurious selections to occur while interacting with the device: buttons might be pressed when the user only intended to scroll the form. Further, by displaying fewer choices, PL obviously decreases the screen's information density at state 0 (see Figure 1). In this context, mental processes like searching for the right item by performing a non-linear selection task could be fostered in cause of lower extraneous cognitive load [4]. This is an indicator that PL's design might prove to be more accurate. The question of which listing type offers a better fit for iPad questionnaires still remains: to the best of our knowledge, no empirical research has been done on this topic so far. However, related work in the field of web surveys shows radio buttons outperforming drop-down menus [2,3]. In this paper, we investigate how these findings translate to touch-screens when questionnaires are completed in a non-linear fashion.

2 Hypotheses

We have already established that "Popover Listing" (PL) requires more "tap"-interactions than "Button Listing" (BL). Thus BL could help the user to perform better. On the other hand, PL might make spurious selections less likely and could reduce cognitive load. As both listing types have their advantages, it is unknown which of them offers better performance while conducting a non-linear selection task (see H1).

H1: Performance differs depending on listing type.

Performance can be regarded as the number of successfully completed tasks within a given period of time. Therefore, we are also interested how performance will vary depending on how much time is available for task completion. It is generally

accepted, that after a certain threshold, a shorter task completion time should lead to inferior performance (see H2).

H2: Lower task completion times reduce performance.

Consequentially, we are interested if listing type and task completion time interact. As we expect both conditions to have an effect on performance, interaction seems possible (see H3).

H3: Depending on task completion time, the listing type influences performance.

Finally, we are interested in the users' perception of both listing types. Good indicators for this are: the users' speed perception, confidence perception, and their overall preference. We expect subjective perception to differ (see H4).

H4: Subjective perception differs between listing types.

3 Method

3.1 Subjects

We performed the experiment with 40 subjects (31 male, 9 female). All subjects were third year computer science students (mean age was $M = 24.5$ ($SD = 3.27$)) attending a course on human computer interaction. 20% of the subjects indicated having no prior experience with touch-screen devices (iOS or Android), 15% indicated having some experience, and 65% indicated being daily users of iOS or Android devices. All subjects were recruited during a lesson on advanced user interfaces and were fluent in German.

3.2 Design

To test our hypotheses, we performed a quantitative experiment. The subjects' task was to identify an object of a given colour and register its occurrence on an iPad questionnaire (see Material). The experiment was a 2 x 3 within subject design. The independent variables are listed in Table 1. The dependent variables were test score (as an indicator of performance) and 3 survey ratings (as indicators of subjective preference). The test score was determined by counting every correct choice per trial. For this to be possible, the test software recorded every one of the subjects' choices. Per trial each subject could reach a test score between 0 and 10 points. In order to avoid a sequence effect, we counterbalanced the listing type order (see Procedure). The surveys' results were compiled from a survey administered after the experiment. Besides questions regarding the subjects age, gender, and iOS/Android experience the survey contained questions to determine which application the subjects perceived as faster, which application the subjects felt more confident with, and which application

they preferred overall. The choices for the latter three questions were all either "Popover Listing" (PL) or "Button Listing" (BL).

Table 1. Overview of the independent variables.

Independent Variable	Value
Listing Type	1) Popover Listing (PL)
	2) Button Listing (BL)
Task Completion Time	1) 15 seconds
	2) 7 seconds
	3) 5 seconds

3.3 Material

The task-related stimulus material was selected from a pool of 500 objects of varying colours. Ten qualifying objects made up a trial. For each trial the objects were drawn at random ahead of time. This means that all subjects were presented the exact same series of objects. To improve discoverability within the questionnaire, objects were classified into five groups (tools, transportation, animals, sports, fruits). Every object was coded to be of one of five colours (red, purple, yellow, blue, black). The language used to express all categories, objects and colours was German. All trials (see Procedure) were assembled in a single PowerPoint presentation, which then was displayed on a 2.5m screen using a classroom LCD projector. An introductory slide was shown to inform the candidates how much time they had to select the given object before the next one would be displayed. The subsequent ten slides were the randomized objects that form a single trial.

Subjects performed the experiment on iPad (2010) devices for which two separate applications were developed. They differed only in listing type. Figure 2 shows the application developed for "Popover Listing" (PL) and for "Button Listing" (BL). Neither application provided any long-term feedback when choosing a value to ensure equal conditions for each trial's task. The size of the buttons or table cells representing the available colours for each object had to be equal and determined by the largest label. Although iPads can be used in both portrait and landscape orientation, we made sure that the developed applications supported only portrait orientation. The devices were sufficiently charged and had all power saving features disabled.

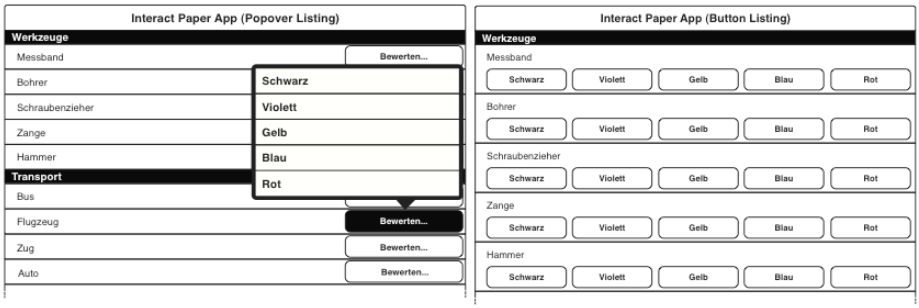


Fig. 2. Comparison of both iPad applications, "Popover Listing" (left) and "Button Listing" (right) differing only by listing type.

3.4 Procedure

Participants were divided into groups of four (totalling 10 groups). Each group was separately led to the prepared room. There they were instructed to seat themselves at any of four test beds positioned three meters from a 2.5-meter screen (angular diameter 45.24°). Before starting the experiment, the test supervisor gave a short introduction. Both the experiment procedure and its tasks were explained. Additionally, participants were told not to expect any visual feedback for their selections to occur. To ensure all subjects had understood the instructions, a quick training session was held. Participants were informed the first trial would allow them 15 seconds per object to find and select the correct choice.

As soon as everybody was ready, the first trial was started. An auditory signal marked the moment when a new object was displayed. This procedure repeated itself through trial 2 and 3, using task completion times of 7 and 5 seconds respectively (see Table 2). Pauses were permitted between trials if participants felt they needed them.

After completing trial 3, participants were instructed to switch their applications. Specifically, participants needed to open whatever application they hadn't used in the first three trials ("Popover Listing" versus "Button Listing"). After switching applications, another training session was held. Trials 4 through 6 followed the same procedure as for trials 1 through 3 (see Table 2). Finally, a brief survey on the participants' impressions of the two listing types was administered.

Table 2. Schedule for a subject starting with the "Popover Listing" application.

Trial	Number of Tasks	Task Completion Time	Listing Type
1	10 tasks	15 seconds	PL
2	10 tasks	7 seconds	PL
3	10 tasks	5 seconds	PL
4	10 tasks	15 seconds	BL
5	10 tasks	7 seconds	BL
6	10 tasks	5 seconds	BL

4 Results

One subject had to be excluded from the results due to failing to comprehend the experiments tasks ($N = 39$). Initial analysis of the data showed an exceptionally high number of errors related to the colours blue and violet. We attribute this to the relatively small difference in colour when displayed using a standard projector. To remove this irregularity, we treated objects of the colours blue and violet as if they were of a single colour.

The grand mean of test scores was $M = 9.67$ ($SD = 0.45$). Specifically, the "Popover Listing" (PL) type had the highest overall test score mean at $M = 9.80$ ($SD = 0.36$). This compares to a test score mean of $M = 9.55$ ($SD = 0.62$) for "Button Listing" (BL) type. Regarding task completion time (TCT) the test score mean for 15 seconds was $M = 9.81$ ($SD = 0.37$), for 7 seconds it was $M = 9.68$ ($SD = 0.72$) and for 5 seconds $M = 9.51$ ($SD = 0.75$). The largest difference in test score means was observed at an interval duration of 5 seconds: PL's test score $M = 9.75$ ($SD = 0.55$), BL's test score $M = 9.28$ ($SD = 1.15$). Figure 3 illustrates the mean test scores for each factor level.

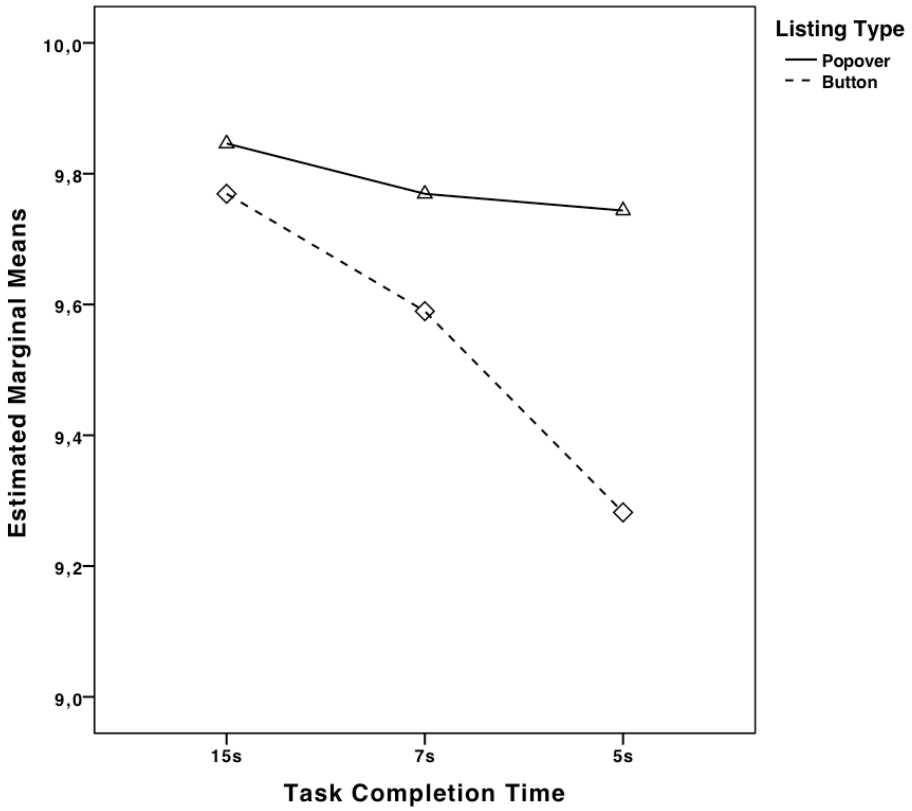


Fig. 3. Mean test scores per condition.

Subjects using the application with the PL type had consistently higher test scores than when using the BL type application. This difference is strongly statistically significant ($F(1,38) = 10.065$; $p = 0.003$). Therefore, we accept our first hypothesis (H1). Data also showed that subjects with short TCTs (i.e. when allotted only 5 seconds per task) were more likely to commit errors than with longer TCTs. In order to test if this observation is significant, we computed inner subject contrasts for TCT and found a linear effect on performance ($F(1,38) = 4.798$; $p = 0.035$). Surveying the mean scores for accordant TCT levels (see above), we accept our second hypothesis (H2). Furthermore, no interaction between listing type and task completion time could be found ($F(2,76) = 2.618$; $p = 0.08$). Consequentially, we reject our third hypothesis (H3).

The survey gauged which application the participants felt more confident with (confidence perception), which application enabled them to fill in the questionnaire faster (speed perception) and which of the applications they preferred (overall preference). The results show that participants were more confident using PL type ($x^2(1) = 21.56$; $p = 0.000$) and preferred the PL type overall ($x^2(1) = 21.56$; $p = 0.000$). Participants also perceived PL to be faster. However, the difference in speed perception is not statistically significant ($x^2(1) = 3.60$; $p = 0.58$). Based on this data, we partially accept our fourth hypothesis (H4). No significant difference between answers from subjects with different levels of expertise was observed. The survey results are listed in Table 3.

Table 3. Ratings for confidence ("which listing type did you feel more confident with?"), speed ("which listing type did you perceive as being faster?") and overall preference ("which listing type did you prefer?").

	Perception Ratings		
	Confidence	*Speed*	*Overall Preference*
Popover Listing (PL)	34	26	34
Button Listing (BL)	5	14	5

5 Discussion

Our experiment provides the following four results regarding our initial hypotheses: first, the listing type does have a significant effect on the subjects' performance (H1 accepted). In particular, when using the "Popover Listing" (PL) design users perform significantly better compared to the "Button Listing" condition. Second, lower task completion times (TCT) decrease the subjects' performance significantly (H2 accepted). Third, no interaction between listing type and TCT could be observed (H3 rejected). Fourth, a statistically significant difference in the participants' subjective perception regarding their confidence and overall preference benefitting PL exists. However, the difference between participants' speed perception is not significant (H4 partially accepted).

The superior performance of PL seems to contrast prior research regarding web-based questionnaires [2,3]. But contrary to the published research, our experiment

setup did not require participants to fill in the questionnaire in linear fashion. Thus, finding the next choice was a search task that would regularly involve scrolling. We believe this to be a major reason for BL's lower performance. The total length of the questionnaire implementing PL was half that of the questionnaire implementing BL. The use of popovers causes forms to be shorter as the individual choices are presented only on demand. We conclude that a design that requires less scrolling (PL) outperforms a design that requires less tapping (BL).

We were also interested if any spurious choices occurred when interacting with the device using BL's listing type. This might happen when users intend to scroll the form and accidentally trigger a choice. We investigated this by manually analyzing all BL related error situations. This way we were able to establish three instances, which strongly suggest that those erroneous choices were the result of handling problems.

6 Conclusion

In this paper we have taken the first steps towards determining which listing type is better suited for questionnaires on iPad. Our results are limited to single choice selections in questionnaires that are completed in a non-linear fashion. For this context we are confident to report that it is strongly advisable to choose popovers over buttons for iPad questionnaires. As iPads and other tablets continue to gain traction and questionnaires remain popular, the future implications of our research are compelling.

Acknowledgments. The University of Applied Science Rapperswil (HSR) and SWITCH funded this research. We would like to acknowledge Philippe Zimmermann, Hans Rudin, Stephan Schallenberger, and Michael Graf for their contributions.

References

1. Apple, Inc. (2011, March 3). *iOS Human Interface Guidelines: iOS UI Element Usage Guidelines*. Retrieved June 10, 2011, from http://developer.apple.com/library/ios/#documentation/UserExperience/Conceptual/MobileHIG/UIElementGuidelines/UIElementGuidelines.html.
2. Johnsgard, T. J., Page, S. R., Wilson, R. D., & Zeno, R. J. (1995). A Comparison of Graphical User Interface Widgets for Various Tasks. *Proceedings of the Human Factors and Ergonomics Society, USA, 39,* 287–291.
3. Heerwegh, D. & Loosveldt, G. (2002). An evaluation of the effect of response formats on data quality in Web surveys. *Social Science Computer Review*, *20*(4), 471–484.
4. Mayer, R. E. (2001). *Multimedia Learning*. Cambridge: University Press.