23-06-2017

Analysis of Sequential & Single Page Forms

Is there a measurable difference in use speed between a sequentially ordered pages in the payment form of a web store and a single-page form of the same web store?

Bor Greve 5488621, Hasse van Maarle 3480399, Petar Kostic 4075897

Utrecht University, Usability Engineering & User eXperience (INFOUE)

Supervised by: dr.dr. E.L. van den Broek, dr. C. van Nimwegen and SAs

Introduction

While web shops may be less of a threat to traditional brick and mortar stores than the hype would have you believe, they are still valuable. Kacen, Hess and Kevin Chiang found that web shops are viewed less positively by consumers than brick and mortar stores, but that webshops have their own advantages regarding brand variety and ease of browsing (Kacen, Hess & Kevin Chiang, 2013). Something web shops always wrestle with is the conversion rate. The conversion rate is the amount of customers that visit a webshop that then actually end up purchasing one or more products. Moe and Fader say that conversion rates rarely exceeded 5% in 2004 (Moe & Fader, 2004). Something that is important for conversion rates is the checkout phase of the webshop. You don't want to lose customers because your checkout is too unwieldy or slow. A lot of webshops use a multi-page checkout system and research does not seem to be able to agree on whether that is a good idea or not. Belk et al. say that people with different cognitive styles may perceive and perform differently on multi-page vs. single-page checkouts. They found that the effect of checkout design on time needed to complete the checkout process is not significant. In contrast there was a statistically significant interaction between cognitive styles and checkout design on the time to complete the checkout process (Belk et al, 2015). Jakobsen says that using single-page checkout is better because it is less prone to bugs since it uses ajax instead of javascript and that it is faster to use and has a higher conversion rate. Jakobsen also mentions however that the conversion rate on single-page versus multi-page checkout depends on the individual customer (Jakobsen, 2015). Van Haperen focusses on conversion rates and they found that no significant results could be found between single-page and multi-page checkout. Their descriptive data somewhat points towards a positive effect of single-page checkout design, they suggest that this may be because single-page checkout is newer and thus better designed.

As can be seen from the above literature, there is no clear consensus on whether single- or multi-page checkout design is better. This research aims to add to the discussion about checkout design by measuring the difference in checkout speed between single-page and multi-page checkout design. Reasoning for this area of research is that, the longer checkout takes. the more chances distracting factors have of interrupting the checkout. Whether checkout speed actually influences conversion rates is outside the scope of this research.

The above literature was found using the following search terms in Google Scholar: bricks to clicks history, multi page single page checkout, "single page checkout", e commerce conversion rates.

Research Design

This leads to the following research question: *Is there a measurable difference in use speed between a sequentially ordered pages in the payment form of a web store and a single-page form of the same web store?* To test this research question, an A/B test was performed on 32 research participants. An A/B test is a test where you present a user two versions of a webpage, in this case a checkout form, that you only vary on one aspect. You then measure the differences between both webpages, such as the time it takes to complete the form. To avoid outliers in such a small dataset the participants were chosen as close in age to each other as possible. The youngest test subject was 18 years old, the oldest 26. Nationalities of the test subjects were Dutch, German and American, and the test subjects were found using convenience sampling.

The test subjects were asked to participate in a research but were not told anything about the type of research except that it required a computer and would take five minutes. Then the test subjects were sent an e-mail, two versions of the email were made, varying only on the link that lead them to the website they were supposed to test. The emails can be viewed in appendix A. In these emails they were asked to fill out a form on a website at a for them normal speed and to answer a short questionnaire afterwards, the questionnaire questions can be found in appendix B. On the website 16 of the research participants were lead to website version A, that consisted of one long form that they had to fill out (for more information see the Material section below). The other 16 participants were lead to website version B that consisted of the same form, spread out over 3 pages. Next to the form an image was placed with example information. The reason for using example information was twofold, first it was anticipated that not all research participants might want to fill in their private information. The second reason was to avoid the speed being influenced by things such as the length of someone's name. The reason for presenting the information as a picture was to avoid some people copy pasting the information, which would skew the research data.

The research was conducted over the course of two days. There were some problems with the date field not properly creating a dropdown on some browser versions, leading in a time discrepancy, due to this fact some participants have been substituted for new participants to avoid the data being skewed. Various methods were used to collect participant data. Statcounter, a tool to gather data about website statistics, was used to collect traffic data and time data. Time and user platform data was also collected using php. Lastly qualitative data was collected using Hotjar, a tool that records all actions taken on a webpage into a heatmap and video. An independent sample t-test was used to compare the, in php collected, data from both forms in order to calculate which of the two forms was filled faster on average. Additional analysis using the questionnaire data was carried out, resulting in a prediction model using time taken as a basis to predict the chance of a participant having used the tab key to navigate the form.

Material Explanation

Several iterations have been used on small set of test subjects to achieve a final experiment. First, it turned out that within the small set the majority quickly pressed the start button without having properly read the text in either the mail or the front page of the test, resulting in them pressing the back button to reread the instructions after realizing they were missing relevant information. During the first version of the test, the time kept counting throughout all of the back-and-forth navigating, resulting in inflated test times due to varying reading speeds rather than form filling speeds. The latest version of the experiment restarts the timer every time the participant opens the first (or only) form page.

Second, some participants were found accidentally pressing enter, resulting in low time signatures at first. Form validation was employed to ensure that information is entered in all fields. PHP form validation might influence the time measurements of a participant due to page refreshes. To ensure that all fields were filled and the task is completed, a jQuery validation plugin was used, which can be found using the following URL: https://jqueryvalidation.org/. The decision was made to not require the data to be a carbon copy of the test data so that the test wouldn't be measuring typing accuracy rather than form-filling speed. Considering most internet forms would have the participants filling in data they are

familiar with, most typing errors were assumed to have stemmed from lack of familiarity with Emma Vink's example information, which is present on a post-it note on each page that contains forms.

Statcounter and Hotjar were employed to obtain additional data regarding the participants while they filled in the form. Hotjar was used to record the specific mouse and keyboard actions used, while statcounter was used to obtain timing, page view count and general traffic. Due to some participants using uBlock Origin, AdBlock or other script blocking extensions, some participants whose data was recorded by the PHP code was not recorded by either service. This lead to a significant difference in the participant sets. It was decided to primarily focus on the data gathered by the PHP code, however, there is still a section describing observations made from the Hotjar recordings.

In Figure 1 you can see the start screen for versions A and B, the start screen was the same for both versions. In Figure 2 you can see the test screen for version A. In Figure 3-5 you can see the test screens for version B. Finally, in Figure 6 you can see the thank you screen for versions A and B, the thank you screen was again the same for both versions.

Website version A can be found using the following URL:
http://www.students.science.uu.nl/~4075897/uxopdr3a/
Website version B can be found using the following URL:
http://www.students.science.uu.nl/~4075897/uxopdr3b/

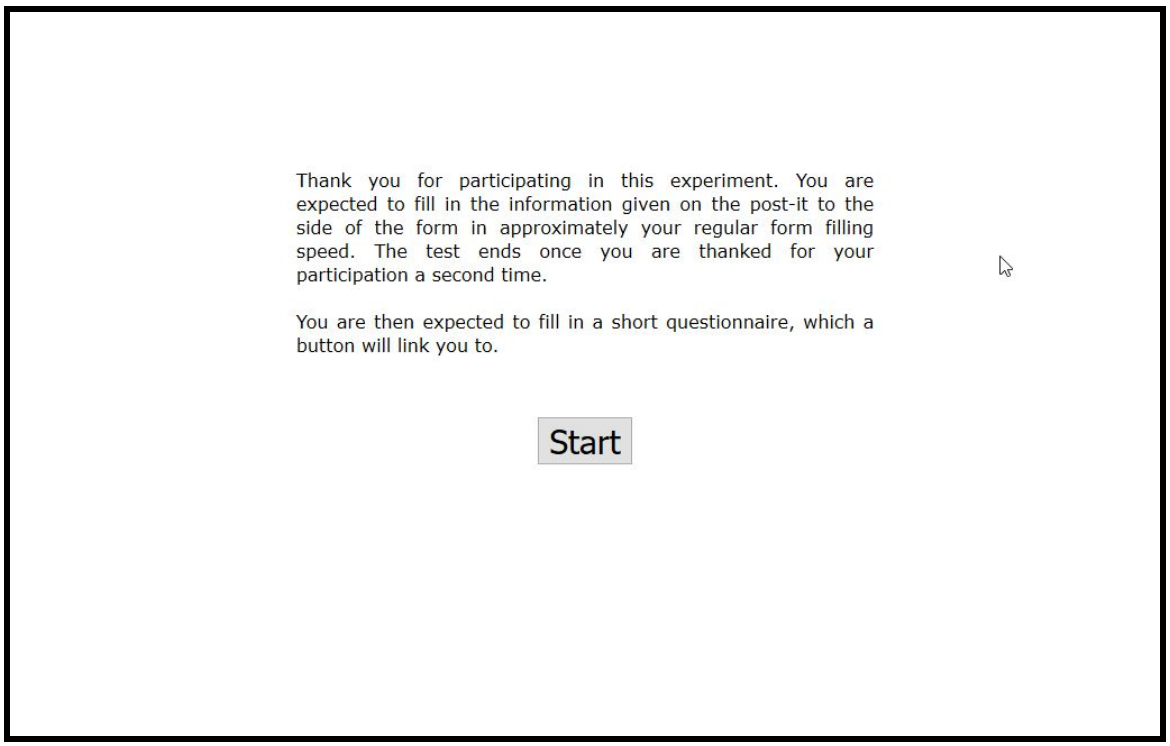

Fig. 1 - Start screen for versions A and B

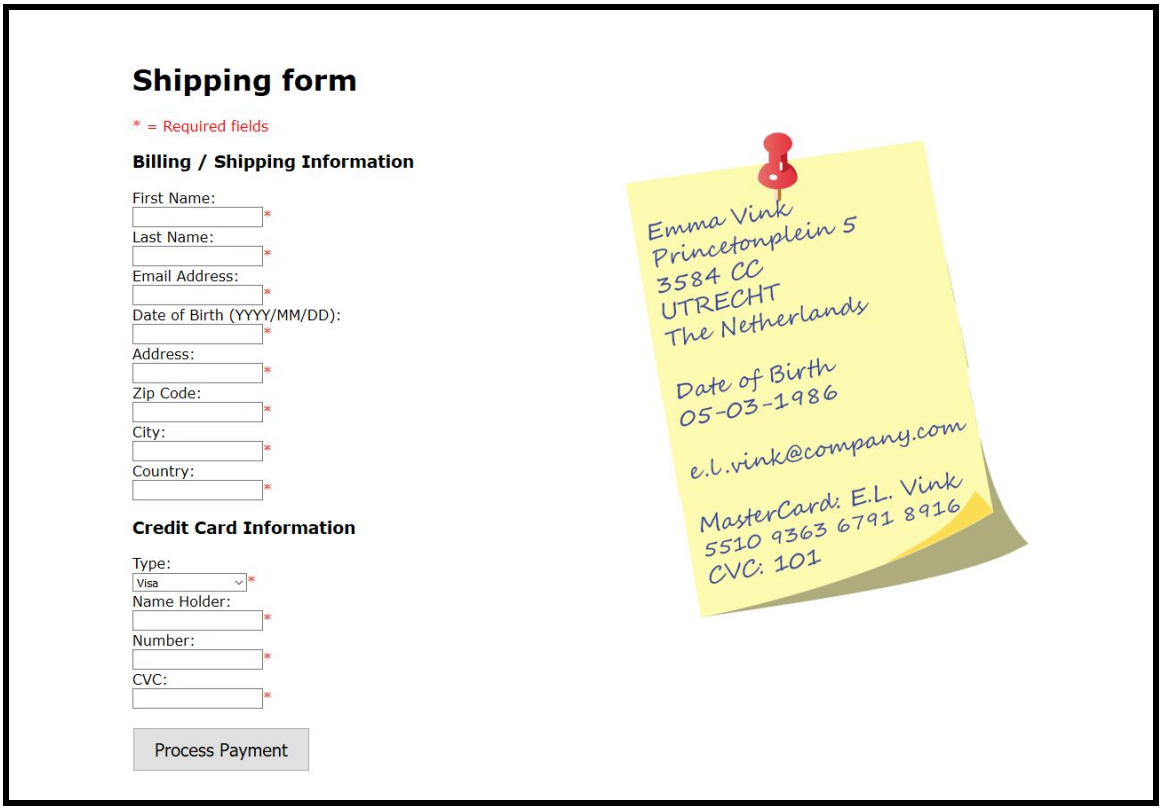

Fig. 2 - Main screen for version A

Fig. 3 - First main screen for version B

Fig. 4 - Second main screen for version B

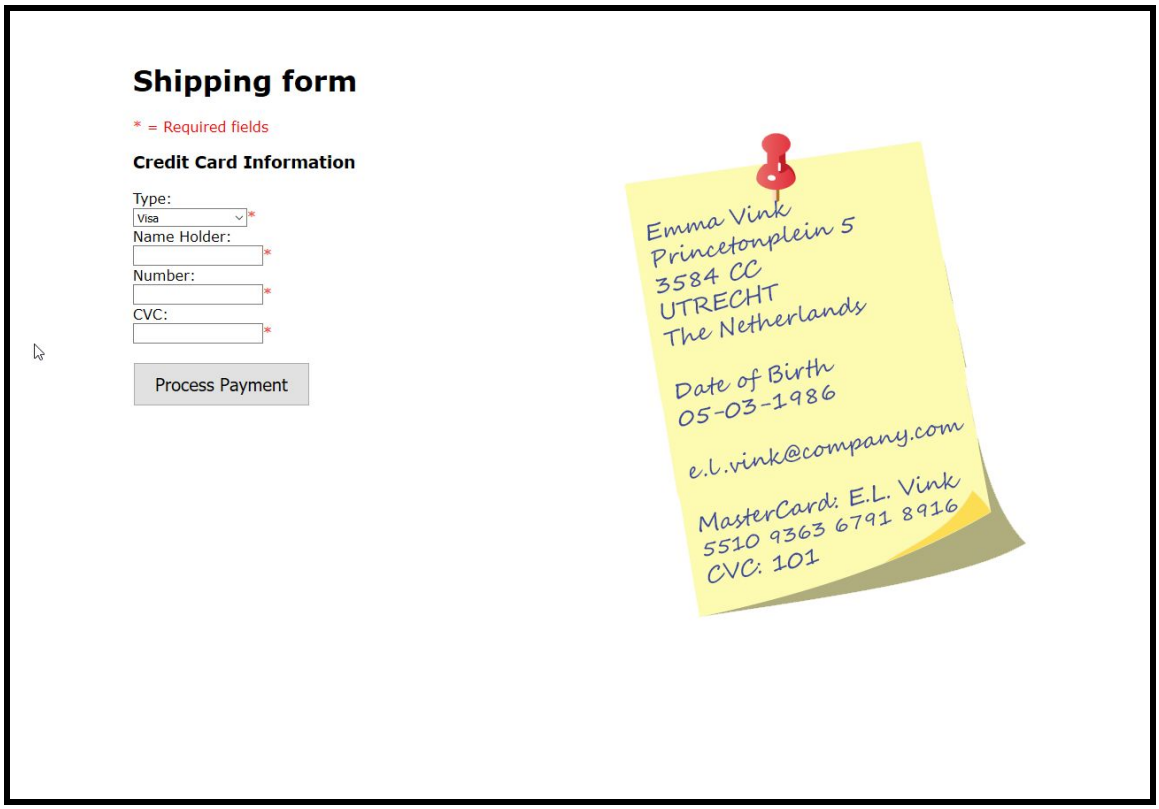

Fig. 5 - Last main screen for version B

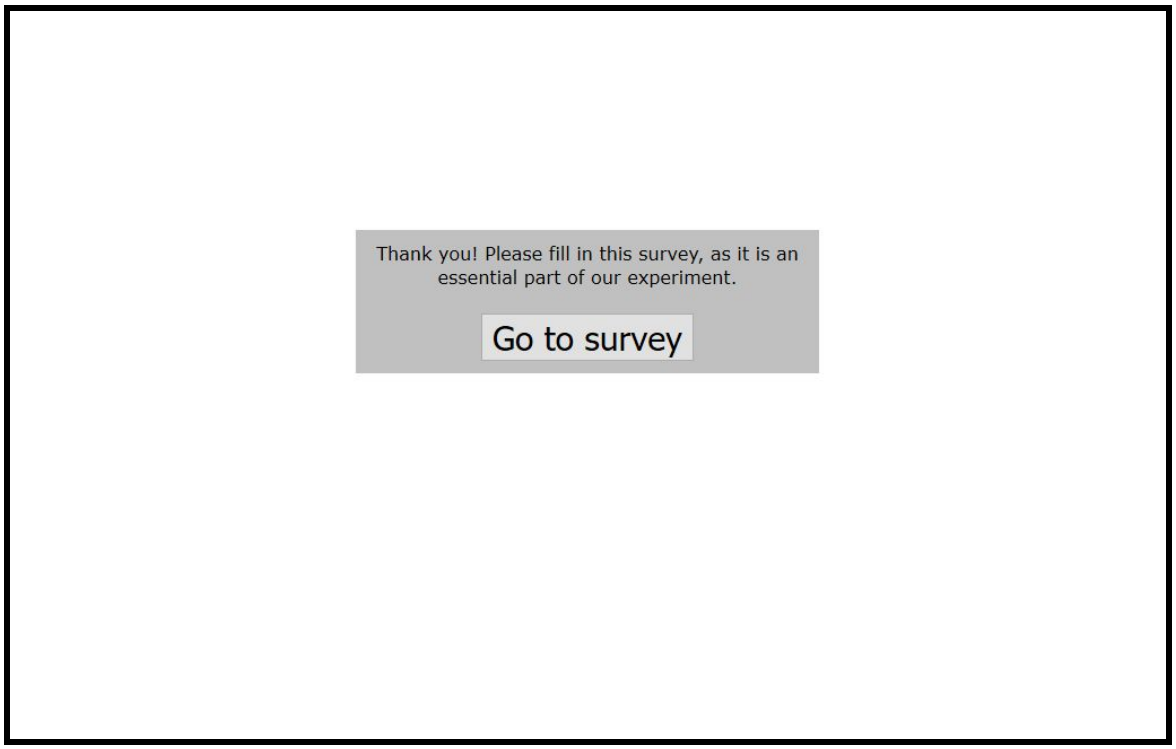

Fig. 6 - Thank you screen for versions A and B

Quantitative Analysis

Dataset Adjustment

To bring the data to a more representative state, a few outliers resulting from confusion and idling on the part of the participants were removed. One participant used their mobile phone rather than a PC. Due to scaling web pages and the necessity of having the post-it note to the side of the form this difference of input device falls outside the scope of this research.

This pruning had the added benefit of bringing the participant counts of both tests to the same number. After comparing the p-values of the t-tests (unpruned $p = 0.64$, pruned $p = 0.1393$) on both the pruned and unpruned data set it is clear that this led to an increase in statistical significance.

Results

Table 1 - Number of participants (N), Mean, Standard Deviation (SD) and Standard Error (SE) by version

Version	N	Mean	SD	SE	Variance
A	16	106.90	42.85	10.71	1836.251
B	16	130.93	46.59	11.65	2170.820

The means of the data sets can be seen in Table 1 along with the standard deviation and standard error values. The difference between the two means is 24.03 seconds, with B being the mean slower form.

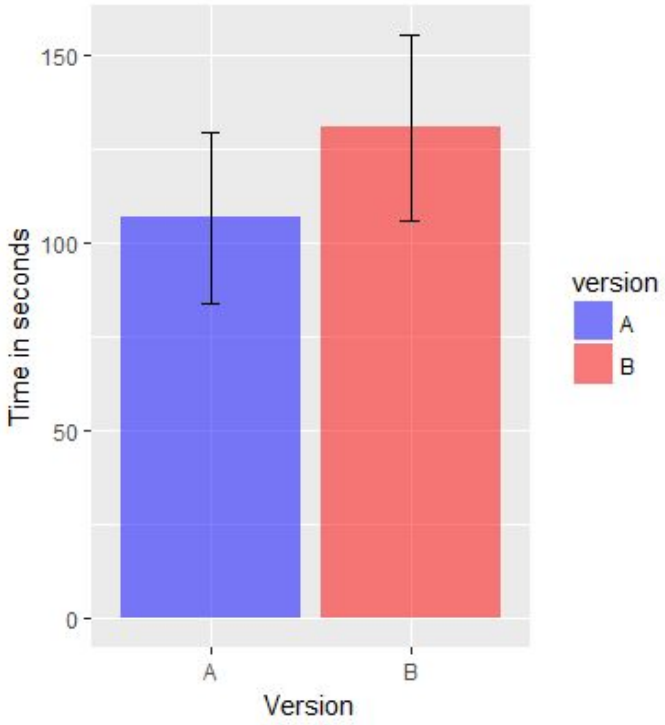

Fig. 7 - Mean and error of test versions A and B

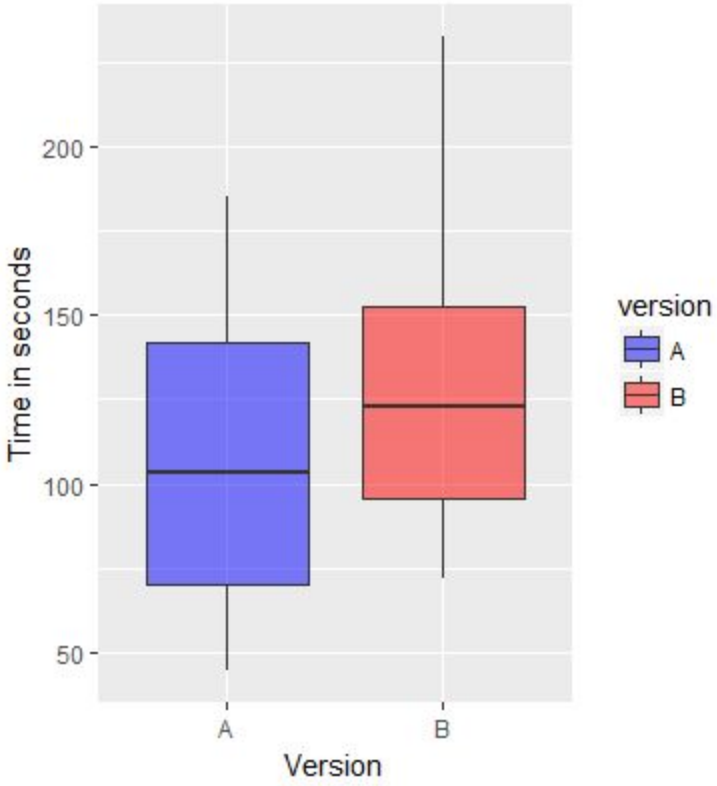

Fig. 8 - Range of time values within the data sets

The full range of time values within the different data sets can be seen in Figure 8. Both the upper and lower bound of the B set are higher than the A set. The whiskers showing variability are long despite pruning, showing that there are still some outliers present in the data. The boxes show the spread of the majority of the data.

A detailed breakdown of the spread of records among time values can be found in Figure 9. It shows that set B has a more concentrated peak in record density around the 110 seconds while set A is a little more spread out. However, set B slopes off less rapidly towards the higher time values in the data set, implying more outliers on the slow end.

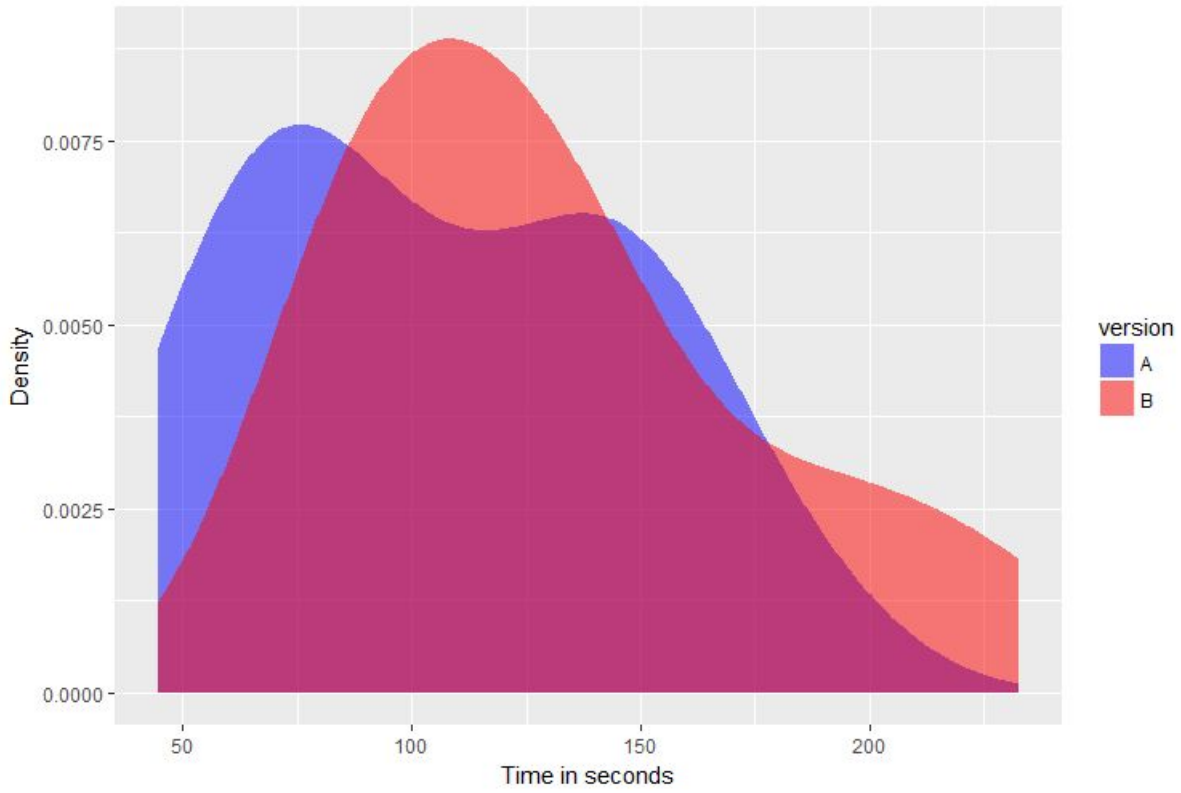

Fig. 9 - Time-Density plot showing the amount of records for any given time in seconds

t-test

An independent-samples t-test was carried out to compare the time taken to fill in forms A and B. There was a difference between the time taken for A (Mean = 106.90, SD = 42.85) and B (Mean = 130.93, SD = 46.59) with t(29.792) = -1.5189, p = 0.1393 as seen in Tables 1 and 2.

Null Hypothesis: There is no relationship between A and B.

Considering $p \geq .05$, this sample does not provide sufficient data to conclude that the null hypothesis does not hold.

Table 2 - R output from all t-tests

Paired t-test results from R					
Version	**Type**	**t**	**df**	**p-value**	**95% CI**
Unpruned dataset	Welch two sample	-0.47774	27.4	0.64	-72.96660 45.39144
Pruned dataset	Welch two sample	-1.5189	29.792	0.1393	-56.365792 8.292492
	Paired t-test	-1.4469	15	0.1685	-59.44504 11.37274

In the case of the pruned dataset, both the Welch and regular t-test are reported. This is because they had different results, unlike the unpruned dataset. Considering both the A and B sets have the same number of participants, this is likely caused by the variance within the pruned data set.

Additional Analysis - Correlation

Results from the questionnaire have been tied to the time recordings of each participant for quantitative analysis. Age and time were plotted against each other in a scatter plot, upon which Pearson's correlation coefficient was used to find statistical correlation and significance between the independent variable 'age' and dependent variable 'time'. However, the p-value on both data sets imply the statistical relevance of the data to be insignificant (p-value for A: 0.0817, p-value for B: 0.7217). A simple linear regression was used to draw a line to make Figures 10 and 11 more intuitively insignificant.

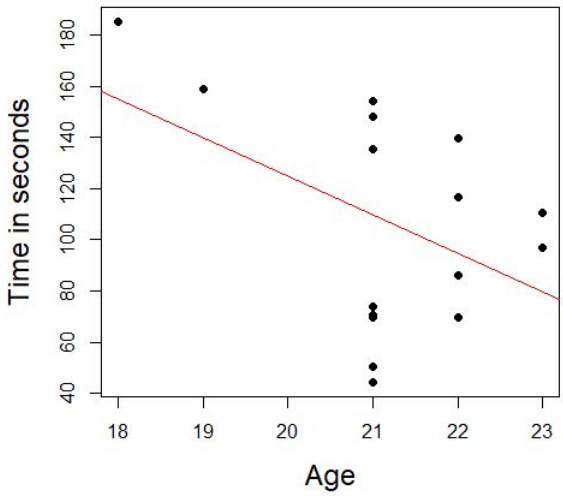

Fig. 10 - Age-Time scatter plot on version A with a linear regression line. (Pearson correlation coefficient coefficient: P = -0.45, p-value = n.s.)

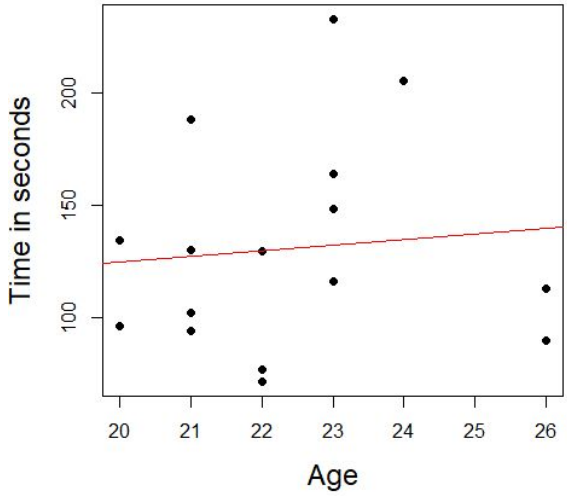

Fig. 11 - Age-Time version scatter plot on B with a linear regression line. (Pearson correlation P = 0.1, p-value = n.s.)

As such this research is unable to provide sufficient evidence with the given sample in favor of a correlation of statistical significance between age and form-filling speed between the ages of 18 and 26.

Extra (unrelated) Analysis - Exploring models to predict the use of Tabs when navigating shipping forms

The use of the tab key to set the typing cursor to the next form element was tracked using the questionnaire (TRUE of FALSE) and subsequently tied to the time recordings of each participant. Logistic regression was used to model the chance of having used the tab key based on the time spent to fill out the forms. The training data used the pruned (adjusted) set of version A and B.

Table 3 - Results and analysis of the logistic regression model.

Version	**Predictor**	**Coef. estimate**	**Std. Error**	**z-value**	**p-value**
A	Intercept	4.72	2.26	2.076	0.0379
	Time	-0.037	0.01	-2.014	0.0440
	Pseudo R^2 - McFadden: 0.29 - Cox & Snell: 0.31 - Nagelkerke: 0.43		**Likelihood Ratio Test** $\chi^2 = 6.04$, $p = 0.014$ Df difference = -1 Log likelihood diff = -3.02		
B	**Predictor**	**Coef. estimate**	**Std. Error**	**z-value**	**p-value**
	Intercept	2.66738	1.87750	1.421	0.16
	Time	0.01150	0.01261	-0.912	0.36
	Pseudo R^2 - McFadden: 0.047 - Cox & Snell: 0.05 - Nagelkerke: 0.076		**Likelihood Ratio Test** $\chi^2 = 0.847$, $p = 0.357$ Df difference = -1 Log likelihood diff = -0.424		

Note: There is some discussion around the significance of pseudo-R^2 values, so they were only used as a rough indicator of for the predictive power. The likelihood ratio test is omitted as the models are not compared.

The Coef. estimate in the second column of version A in Table 3 is the coefficient associated with the variable listed to the left. It is the estimated amount wy which the log odds of "Probability of using Tabs" would increase if time would increase by one second. In third column, the standard error associated with these estimates can be seen. The z-value in the column to the right is obtained by dividing the coefficient estimate by the standard error. However, this quotient cannot be assumed to be normally distributed with our sample size of N = 16. Next to the z-values are the two-tailed p-values that correspond to those z-values.

A p-value of 0.0379 for time indicates significance at the 5% level. The null hypothesis that the predictor's coefficient is equal to zero (Intercept) can thus be rejected. A plot of the regression model can be found in Figure 12.

The low pseudo-R^2 values in Table 3 indicate low to medium predictive power of the model.

Contrary to version A, the data of version B showed no predictive power nor significance in our analysis and as such was not plotted and no performance was measured.

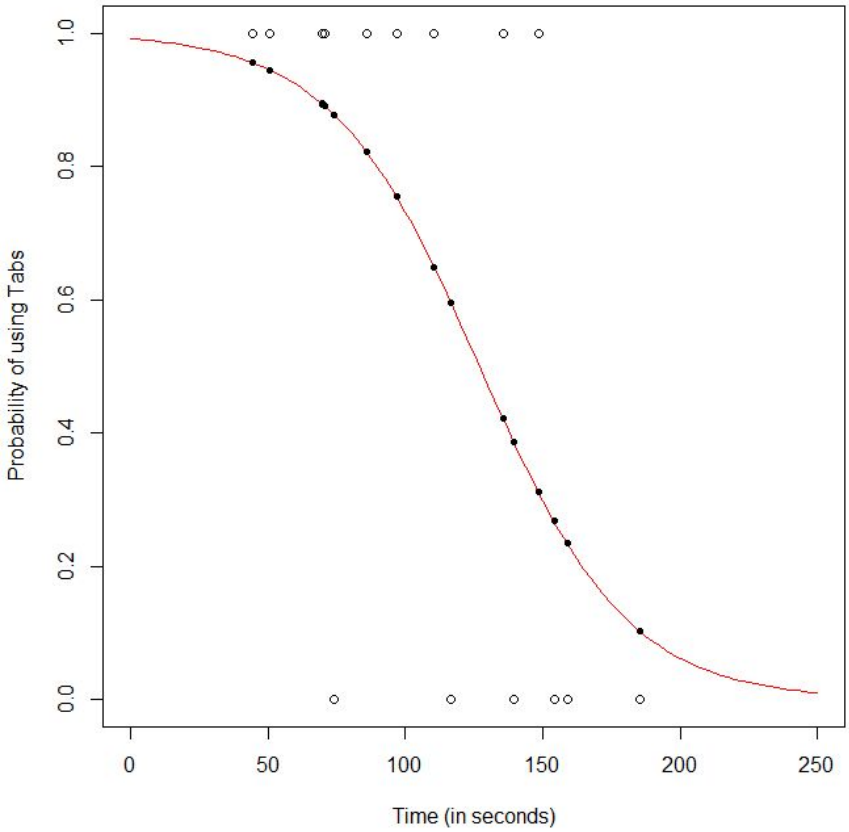

Fig. 12 - Probability of using tabs as a function of time trained on the data of version A. Note the overlap of of the unfilled data points on the top and bottom. These are individual observations of recordings in version A used to fit the model indicating the use of spaces (TRUE = 1.0 and FALSE = 0.0)

To validate the model that was trained using the data of version, A test run was ran against the data of version B with a decision boundary at a probability of 0.5. Afterwards, a ROC curve was plotted. Prior to analysis, the data was shuffled to prevent overfitting.

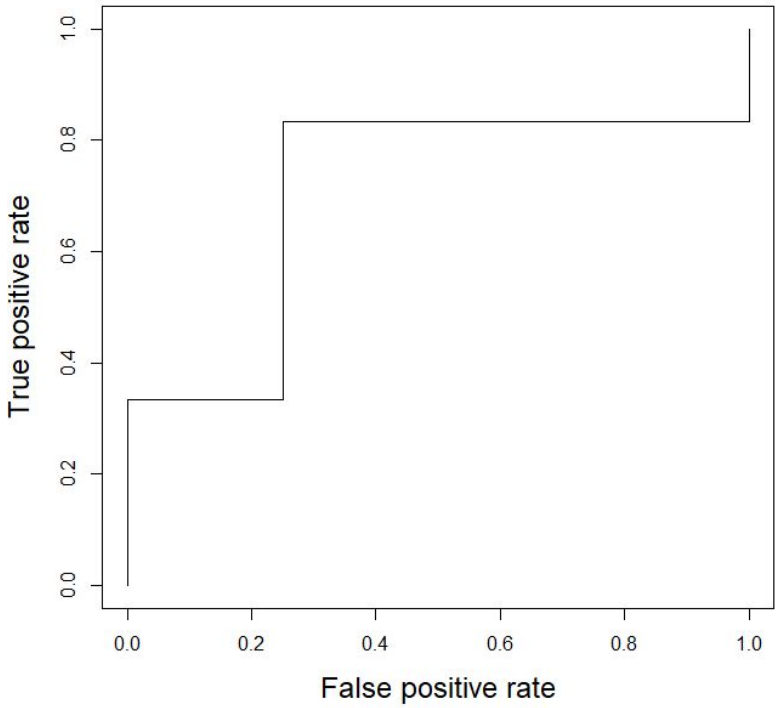

Fig. 13. - The ROC curve of the model of version A that was ran against the data of version B. AUC = 0.71

The shape of the ROC curve in Figure 13 indicates an insufficient amount of test data to measure the model's performance. The AUC of 0.71 can be assumed to random chance on such a small data set.

Given the results of our model from version A, it can be concluded that solely looking at time measurements when filling forms can be a significant predictor for using Tabs. However, it remains an open question whether more data or more covariates can be added to the model in order to make more accurate predictions. Also to be noted is that this model is tailored to our form and forms of different sizes might yield different results.

Qualitative Analysis

Based on the Hotjar videos various things that might have influenced the time spent became clear. The first thing that became clear was the fact that not all the research subjects properly read the instruction email and welcome page, they started by filling out their own data and when they came to the credit card information data they realised they didn't want to fill out their own data, went back to the welcome page, properly read what they were supposed to do and then started over. Because of this reason the statcounter data, that logged their total time on the website, have not been used. Instead, the php data that only logged their time on the actual test pages. However, they may have been faster at filling out data the second time around, because they had one practice run with their own data. Another thing that became clear is that some participants had trouble with the date field of the test pages. The date field unfortunately gave a dropdown on only some of the used browsers, resulting in some of the people who didn't get a drop down fighting with the allowed format for a bit before they managed to fill it in without getting an error that

said they needed to fill in a ‘correct date’. As mentioned before, the people to whom this happened have not been included in the research, replacements for them have been found.

Hotjar’s heat map function did not work properly during the duration of the tests, as it only recorded clicks on the front page of the test form. Had it worked it could have been used to see how many participants were able to use the html5 date picker to fill the birth date field, and how many did not get that option. It could also have provided extra certainty as to the amount of participants that used tabs to navigate the form instead of using the cursor to click the individual fields, as this was currently only recorded in Hotjar’s recordings which did not manage to capture all participants, and the highly subjective post-fact questionnaire in which participants were asked if they had used tab.

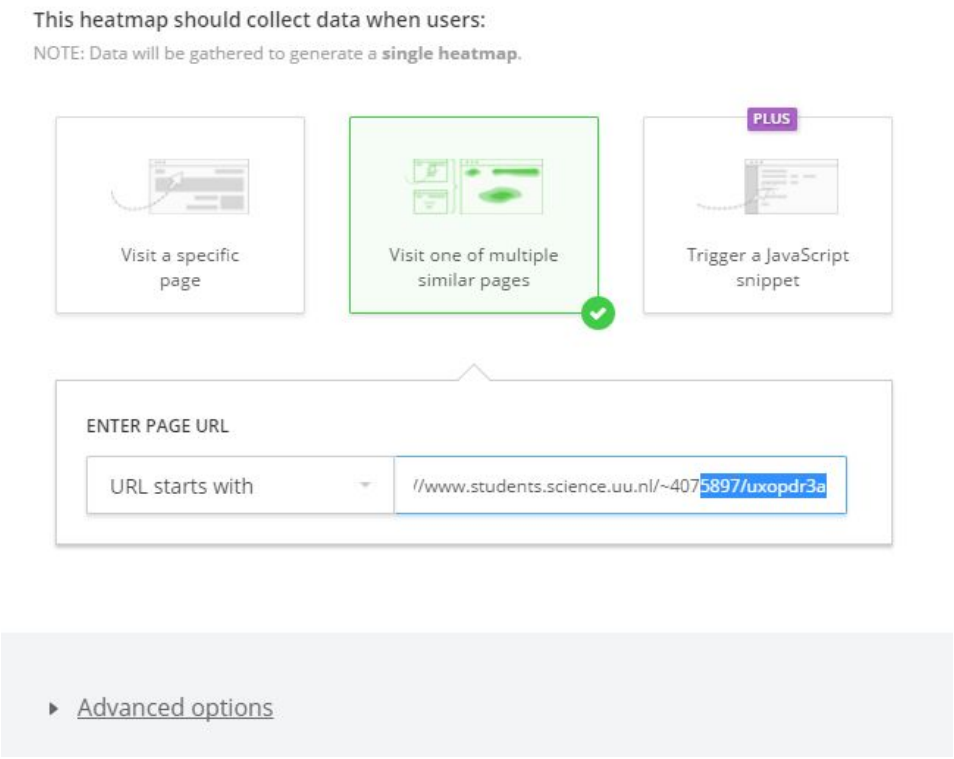

Fig. 15 - Our settings in Hotjar for both version A and B. However, only a heatmap of the start page was able to be viewed after the experiment

Conclusion

An A/B test was performed to determine the difference in time taken to fill in forms in either a sequential or single page styled online shopping form. The conducted research does not show sufficient proof that the A and B form styles are of statistically significant difference in terms of the time it takes to fill in the respective forms. This means the research hypothesis has been disproven, and that this research has been unable to disprove the null hypothesis. Future research should focus on increasing the test sample size and reducing differences in test results caused by participants not reading the instruction email properly. Research that builds upon this research could focus on whether the speed of checkout influences the conversion rates of consumers.

References

Belk, M., Germanakos, P., Constantinides, A. and Samaras, G. (2015). A Human Cognitive Processing Perspective in Designing E-Commerce Checkout Processes. *Human-Computer Interaction – INTERACT* 2015, pp.523-530.

van Haperen, M. (2012). *Validating the e-servicescape*. Master. Eindhoven University of Technology.

Jakobsen, T. (2015). *Redesigning a Single Page Checkout*. Master. Aalborg University.

Kacen, J., Hess, J., & Kevin Chiang, W. (2013). Bricks or Clicks? Consumer Attitudes toward Traditional Stores and Online Stores. *Global Economics And Management Review*, *18*(1), 12-21.

Moe, W., & Fader, P. (2004). Dynamic Conversion Behavior at E-Commerce Sites. *Management Science*, *50*(3), 326-335.

Appendix A - Emails

For version A:

Dear participant,

Thank you for participating in this research. We'd like you to go to a webpage (link below) and fill out the form you find there. Please do this on a desktop PC or laptop, not on a phone or other small screen. We understand that you might not want to fill in your own personal information on the website so next to every field of the form we have provided a small picture with example information. Please fill in the example information instead of your own! Fill in the form at your normal speed, not extra fast, not extra slow or precise, just the way you normally would. When you're done click on the "Process Payment" button, this will take you to a page with a short post-research questionnaire we'd like you to fill out. Only fill out the form and the questionnaire once. Once you are done make sure you close the browser tabs of both the research website and the questionnaire.

Thank you so much for participating!

Greetings,
Petar, Bor and Hasse

The URL for the research website is: http://www.students.science.uu.nl/~4075897/uxopdr3a

---

For version B:

Dear participant,

Thank you for participating in this research. We'd like you to go to a webpage (link below) and fill out the form you find there. Please do this on a desktop PC or laptop, not on a phone or other small screen. We understand that you might not want to fill in your own personal information on the website so next to every field of the form we have provided a small picture with example information. Please fill in the example information instead of your own! Fill in the form at your normal speed, not extra fast, not extra slow or precise, just the way you normally would. When you're done click on the "Process Payment" button, this will take you to a page with a short post-research questionnaire we'd like you to fill out. Only fill out the form and the questionnaire once. Once you are done make sure you close the browser tabs of both the research website and the questionnaire.

Thank you so much for participating!

Greetings,
Petar, Bor and Hasse

The URL for the research website is: http://www.students.science.uu.nl/~4075897/uxopdr3b

Appendix B - Questionnaire

The questionaire for version A can be found here:
https://docs.google.com/forms/d/e/1FAIpQLSeA7WGxJ5ffEq_O1lkLJcCk78aZExk-3xM3qUhCTTctlKHn_g/viewform

The questionnaire for version B is an exact copy of the questionnaire for version A. Below the questionnaire questions and a summary of the answers can be found.

Question: *What is your name?*
Answers: Names are known to the researchers but will not be published to guarantee the privacy of the test subjects

Question: *What is your age?*
Answers: See Table 4.

Table 4 - Age distribution for test groups A and B

Age	Amount in A	Amount in B
18	1	0
19	1	0
20	0	2
21	7	4
22	4	4
23	2	3
24	0	1
25	1	0
26	0	2
Total	**16**	**16**

Question: *What is your gender?*
Answers:
In group A: 3 female, 13 male
In group B: 1 female, 15 male

Question: *What occupation or occupations do you hold currently?*
Answers: answers varied widely, but most participants were students.
In group A: 12 out of 16 participants were students
In group B: 13 out of 16 participants were students

Question: *Did you do anything to speed up the filling in of the previous form?*
Answers: answers varied widely but the most common answer was either the use of tab, or ‘no’.
In group A: 4 participants stated they did nothing to speed up the filling out of the form. 11 participants mentioned using tab to speed up the filling out of the form. 1 participant said they used another method to speed up the filling out of the form.
In group B: 4 participants stated they did nothing to speed up the filling out of the form. 11 participants mentioned using tab to speed up the filling out of the form. 1 participant said they used another method to speed up the filling out of the form.