Improving Task Performance in User Experience Writing

— a validation of two methods in digital and in-person user contexts

LINNEA HOLM
Abstract

User Experience (UX) Writing is part of the product’s overall user experience design. It can be a part of a product content strategy as much as the visual design. The base for any UX centered profession is the understanding of usability — the practice of making products or services easy to use for the intended target group. The difference between the UX designer and UX writer is that the writer contributes with an in-depth understanding that textual content contributes to an overall better user experience along with the product brand.

This study includes two UX writing usability evaluation methods for a defined functional service application. By performing the UX writing methods through an online form as well as in an in-person user interview, the validity of transforming an established method into an online tool was investigated. The liabilities of both contexts were evaluated and future developments are suggested. To accomplish this, a literature review was conducted and an online tool was developed. This was used to gather data from 20 users, 10 in the online context and 10 in the in-person context. The datasets were then compared for validation, which in turn served as a foundation for further discussing the possibilities of using an online tool in this situation.

The conclusion is multifaceted. The in-person user tests require resources in terms of time for the UX writer, finances in terms of compensation for the users and result in a smaller data set. Digital user tests also require time and resources, but can generate a much larger data set, seeing as it provides a data set of the same quality level as the in-person user tests. Considering the advantages of an online tool, it can provide a valid replacement for in-person user tests, if one considers that the assets outweigh the minor liabilities. However, the potential of a future collaborative dataset and all the further developments suggested in this paper are what is truly worthwhile investigating further.
Sammanfattning


Improving Task Performance in User Experience Writing
— a validation of two methods in digital and in-person user contexts

Linnea Holm
KTH Royal Institute of Technology
EECS School of Electrical Engineering and Computer Science
Stockholm, Sweden
lihol@kth.se

ABSTRACT
User Experience (UX) Writing is part of the product’s overall user experience design. It can be a part of a product content strategy as much as the visual design. The base for any UX centered profession is the understanding of usability — the practice of making products or services easy to use for the intended target group. The difference between the UX designer and UX writer is that the writer contributes with an in-depth understanding that textual content contributes to an overall better user experience along with the product brand.

This study includes two UX writing usability evaluation methods for a defined functional service application. By performing the UX writing methods through an online form as well as in an in-person user interview, the validity of transforming an established method into an online tool was investigated. The liabilities of both contexts were evaluated and future developments are suggested. To accomplish this, a literature review was conducted and an online tool was developed. This was used to gather data from 20 users, 10 in the online context and 10 in the in-person context. The datasets were then compared for validation, which in turn served as a foundation for further discussing the possibilities of using an online tool in this situation.

The conclusion is multifaceted. The in-person user tests require resources in terms of time for the UX writer, finances in terms of compensation for the users and result in a smaller data set. Digital user tests also require time and resources, but can generate a much larger data set, seeing as it provides a data set of the same quality level as the in-person user tests. Considering the advantages of an online tool, it can provide a valid replacement for in-person user tests, if one considers that the assets outweigh the minor liabilities. However, the potential of a future collaborative dataset and all the further developments suggested in this paper are what is truly worthwhile investigating further.

Keywords
User Experience Writing; UX; Usability; Usability Testing; Cloze Deletion Method; Highlighting Method

1. INTRODUCTION
User Experience (UX) design as an established practice is becoming more of a standard than an exception in development processes today [16]. User experience writing, or UX writing, is a part of the product’s overall user experience design and can be a part of its content strategy as much as the visual design. UX Writers produce textual content and evaluate its impact on the overall usability of a product — they recognize language as an essential part of a user’s experience with a product. Copywriting, the written communication for the purpose of conveying a brand or message, is a key part of creating content. As opposed to UX writers, the Copywriter writes for advertisements or publicity materials [30]. The Copywriters write to persuade. UX writing, however, is different — the creation process might be similar, but while the goal of copywriting is persuasion, a UX writer creates content for the purpose of enhancing the user experience, and not solely for marketing purposes [4]. A UX Writer looks at the entire user experience, from the problems to solve and the goals to reach, through collecting and analyzing data, creating ideas, building prototypes, testing them, and then adjusting the words and produce text. They aim to engage the user of a digital product with intuitive language that focuses on the user experience.

A target user group’s experience of a digital product should be developed around said user group, helping them solve problems and reach their potential within the framing of the product [24]. The user experience can be elevated by improving physical, mental or mechanical interactions, like accurate visual design, use of symbols or intuitive flows. However, when it is necessary for the user to not only be able to see, but also to specifically comprehend the written content, the text needs to be specifically adapted to fit that user’s needs [24].

There are different methods to measure the quality of textual content based on a purpose and a target user group. While some of them rely on testing for basic comprehension or readability, there are some methods that evaluate subjective user values connected to the content — how users feel regarding expressions, certain language or specific phrases. Even specific words can convey a sense of trust or distrust for a product, and can be evaluated accordingly [30]. All aspects of a design contribute to how well an interface is being perceived and comprehended by users, which is why it is important to test textual content, as well as the design, iteratively throughout the design process, even in its prototyping stages.

A significant aspect concerning implementing UX writing methods into a design process, is time management. Depending on the project management’s approach, for instance if the team work with agile workflows, the amount of time that can be spent on evaluating the content can vary significantly [26]. Ideally, all evaluation methods, including broad scope user tests with user interviews and large test groups, are executed for each segment of
the product. The usability testing is preferably performed in several iterations, with large data sets and a generous time frame to sufficiently analyze the test data. Realistically, this might require resources beyond capacity. To save time and resources, performing the usability testing online, can help to acquire more data in a shorter time frame.

There are several aspects to take into consideration when transforming an in-person method to an online method. If one can validate the level of dataset quality in the two contexts, the possibilities for further development can be explored.

1.1 Research Question
To perform an evaluation of the text in a functional service application, two UX writing methods in two different contexts were performed — a developed digital context and an in-person context. The UX writing methods can together create a basis for improvement for the existing text in a user interface. The main purpose of this study is to perform a validation of the data acquired in the digital tool. The method will result in the gathered user data from the two contexts where a comparative assessment will conclude the study. The research question investigated is:

When evaluating the usability of textual content, what are the assets and liabilities with using an online testing tool compared to in-person user tests?

1.2 Principal: Mobiento/Deloitte
The principal, a digital consulting agency, Mobiento/Deloitte1 is supporting this thesis. The context, a functional service application, is part of a project where they have been involved developing the user experience.

2. THEORY AND RELATED RESEARCH
This chapter will establish related work and background as to which methods were chosen for the developed digital and in-person user context — The Cloze deletion as well as the Highlighting method. It will also cover the bases of UX writing and human-information interaction, as well as the importance of having a visual structure and visual design elements present when evaluating the text in a user interface.

2.1 UX Writing
The base for any UX centered profession is the understanding of usability. Usability is the practice of making products or services easy to use for the intended users [7], which makes it the foundation for UX writing as well. UX in itself has appeared within the design development community to cover the components of users’ interactions, more than solely effectiveness and efficiency [25].

Similar to User-Centered Design (UCD), UX studies focus primarily on the intended user, which can be broad or narrow in its definition, and the associated context of use for that product or service. The focus remains throughout the iterative design and evaluation process of a product. This requires an in-depth understanding of the user and the context of use [17]. While the UX designer can cover a multitude of areas that concerns the user experience, the UX writer focus on the research, development and evaluation of textual content for the user in the intended context. This in turn can imply several things, including content strategy, micro copywriting, copywriting, but also an overall understanding of the entire product design and development process. Hence, the difference between the UX designer and UX writer is that the UX writer contributes with an in-depth understanding on textual content, well written text, contributes to an overall better user experience but also the product or company brand. The UX writer role, as described on Google Career [31], the company’s own page for job postings:

“As a User Experience writer, you are an advocate for Google design, working to shape product experiences by creating useful, meaningful text that helps users complete the task at hand. You help set the vision for content and drive cohesive product narratives across multiple platforms and touch points.”

This suggests that not only should a UX writer produce textual content, but rather a product wide narrative and helpful content, which still has the intended user or users at its focus. This includes readability and comprehension as well as consistency [12]. There are several aspects in producing UX text that are important for the user, but equally important to the brand. They operate in symbiosis. Expressing a brand image as an inherent part of text requires the writer to produce a tone and a voice that caters to the user, and in extension to the brand. Both are important, and a well performed text design can lead to both goals being achieved. Producing the text for an interface needs to be in the right style, tone and voice for the brand. This will help the user feeling like they are at the center of the experience [32].

The UX writer can do user research, produce concepts, actual text, as well as testing and evaluating to make sure the text is consistent, helpful, authentic and useful [12]. Producing text that is helping the user requires research, and the goal is to design an overall well produced user experience that supports the user becoming an expert in using the interface while they use it [21]. The UX writer produces the right words, at the right time, in the right context, with an accurate process in the right team.

2.2 Human-Information Interaction and Visual structure
In psychology and information science, researchers look at the information-processing model of cognition [6] as to how information is acquired and managed by human consciousnesses. These researchers were some of the first to apply thinkalouds during information-processing to, for instance, be able to evaluate information. They argue that the human mind can be represented as three different stages: computational, representational and implementational. Instead of looking at information as something that communicates one way, their approach was to consider how the users, the humans, seek and use the information, in order to understand how to better improve it in terms of comprehension, as well as understanding that the cognitive approach has important theoretical and practical associations. The information-processing model of cognition [6] provided a foundation for Human-Computer Interaction (HCI), which has inspired information scientists to evaluate information based on usability [2].

Human-information interaction (HII), the interaction between humans and information, should consider both sides as entities, and not just the human part of the interaction. This is important, as the nature of the human actions are both mental and physical, but the actions can be initiated by either the human entity or the information entity. The fact that the information itself can act as an entity in the interaction needs to be taken into consideration when

---

1 https://www.mobiento.com/
producing user friendly information. While HCI addresses interaction between humans and technology, HII acts more on the internal human interactions. This is key in understanding how to evaluate information based on usability [18].

Theoretically, a human uses information with little to no reciprocity, but in HII, a human interacts with the information, which can evolve based on the interaction. This must be taken into consideration when both producing the content and evaluating it. Both of these are the responsibilities of the UX writer.

In regard to visual structures, it is vital to take visual design aspects into account when testing for usability. Visual structures affect human comprehension, not only in terms of visual design elements, but also in terms of other types of visual structures that the written content is placed in. When presented in a terse, structured way, people can scan and comprehend the information more easily [14]. Visual structures that generally helps increase the level of comprehension for users include, but are not limited to, visual hierarchy — titles, headings, boldness, bullets, and indenting and placement in the visual design [15].

This suggests that usability evaluation of text requires the visual design aspect of the user interface to be present, as opposed to merely evaluating the text itself without context. In addition to the importance of having the accurate visual structures, the realistic scenario, adds to the supporting of users staying in their role and reduce hesitation during their use of the product. The closer that the test context represent reality, the more reliable the test results and the resulting insights into product usability [7].

2.3 Qualitative and Quantitative Methods for Improving Textual Content

Evaluating content for usability in a digital product can both include methods that result in quantitative metrics and methods that include subjectively based results in a more qualitative manner. There are established methods like readability and comprehension testing [23], that can provide insights for an already existing text — these are generally dependent on large data sets and will tell the UX writer if the user understand the content of the text, but not necessarily if the goals that are set for the content are met.

The subjective methods might not necessarily be as measurable as the quantitative methods, but provide a base together with the quantitative method for improvement regarding the usability of the text. By using one method to generate quantitative data and one method to generate qualitative data, both approaches are covered. One important aspect in setting up the evaluation, both in online format and for the in-person tests, is to entertain the advantages of each context. Structured in-person tests require interviewers to behave consistently when reading questions and recording answers, something that online forms already have as an advantage [28].

2.4 The Cloze Deletion Method

The Cloze deletion method, which can be described as a ‘fill in the gap’ type of test, is a commonly used test to assess the understanding or evaluate specific words in a text [19]. The method is to blank out every nth word, where n varies depending on context. Depending on what words or phrases one wants to look for, the blanked out word pace can vary.

During the test, the users are asked to read the text, silently, and fill in the blank spaces to the best of their ability. In some cases the method can be used to test the users, as is sometimes the case during second language education. In this case the words filled in by the users are compared to the pre-existing content, to evaluate specific words the users feel confident about or already have in their vocabulary. The percentage of correctly restored words form a metric to measure success, where 60% or above are considered successful [33]. Synonyms and misspellings count, and a recurring synonym might indicate that this particular word should be exchanged. Iteration based on the Cloze deletion method result do not only include replacing words, but also rephrasing of sentences to clarify some of the potential confusion.

Cloze methods are generally used on students in learning a first or secondary language, but also on adults, to measure comprehension. Other studies have explored the possibility of digitizing Cloze methods, but for this purpose rather than testing for usability. There has been some automation of Cloze methods in learning grammar, where educators can turn online texts into Cloze methods [5], [13].

The Cloze deletion method can be in either a maze or an open-ended format, where the maze format offers the user three options for each blanked out word, and the open-ended format just leaves a blank for the user to fill out [27]. The validity of the Cloze method for assessing comprehension has been accepted based on a correlation between Cloze scores and scores on other forms of assessment [9], [11]. The method is used within the UX writing community, where the assessment addresses the user’s comprehension of the textual content. In extension, that also measures the level of comprehension regarding the service or product itself. It works as a part in the process of evaluating the entire design and is an important step in content strategy.

The maze format of the Cloze method is considered to be a global measure of reading that requires decoding, fluency and comprehension skills [8], which is suitable in measuring adequacy in language. The open-ended Cloze format, on the other hand, has been suggested to measure the ability to access and search background or word knowledge [8], which is more accurate in testing for usability.

2.5 The Highlighting Method

The Highlighting method [3] contains qualitative value based testing, and caters more to user values. Performing an evaluation of the target users perceived values can help the content creator to understand the impact of what they are writing, and adjust accordingly. A target user will act differently depending on the level of, for example, trust for a product or a brand. With a higher level of perceived brand image, a product can generate greater customer loyalty and less negative attitude to price fluctuations and less vulnerability compared to competitors, among several aspects [1].

The Highlighter test [10] consists of users highlighting sections in written content that confirm or does not confirm their perception of the content and the overall product. The users are asked to read a text and highlight, in two different colors, parts of the text that confirms or denies their values. The chosen values are set up in pairs, and should be opposites against one another. One task includes the user being asked to highlight everything in the text that makes them trust the service in green, and highlight everything that does not make them trust the service in red. These value pairs can be exchanged, depending on what is relevant and interesting to look for.

This makes this method a valuable asset in testing for usability — something that might be perceived as distrustful or difficult for some users can decrease the usability, significantly for products or
services connected to an online banking context [3]. If a lot of the words or expressions correlate in one or the other color, the content creator or UX writer can get an overview of the values or feeling the user has towards that phrasing, and adjust these in a desired direction.

2.6 Hypothesis
UX writing is an important part of the design process and content strategy. The resources required in implementing UX writing into the development process of a digital product, might not always be prioritized in an agency setting. With this in mind, two UX writing methods can be adapted into a design development process, by using an online tool.

3. METHOD
This section will introduce the scientific methodology, which includes the initial research and literature review regarding UX writing and usability, along with the in-person user tests and the development of the online tool for validation of the online setting.

To investigate the research question, an explorative study in two different contexts were set up. The first part of the study included a literature review and the development of an online tool to perform the two UX writing methods online. This tool was then tried out on users in the context of a fund savings page in a functional service application. To be able to validate the quality of the user data acquired by using this tool, the second part of the study included a control study with in-person user tests. This method is of appropriate character based on the lack of comparative research within the subject, taking into account the recent nature of the methods in a digital context.

The UX writing methods chosen for this study are of two characters, described in depth in the previous chapter. The Cloze deletion method is of qualitative character and the Highlighting method is of qualitative character. These methods together provide deeper insight as to what words or expressions can help the users experiencing the content in the functional service application. This qualifies the methods as appropriate for this study, as exploring both qualitative data and quantitative data sets renders a suitable foundation for validation. The main focus regarding the method is the two contexts in which these methods were performed, the in-person user context and the developed digital context.

3.1 Literature Review
The first phase of the study included a thorough literature review, in order to have a foundation for the development of the online tool and the setup of the user tests. The aim was to gain deeper insights into human-information interaction, the specific methods used, user experience design in general and user experience writing in particular. This was to both be able to perform a control study in terms of the in-person user tests and to develop the online form tool in a sufficient manner. The results of the literature review are manifested both in the description of the UX writing methods and the online tool itself.

3.2 Functional Service Application
The application was provided in a prototype format (fig 1). One page within the application was chosen, on which the user could learn more about fund investments. The client, a Swedish bank, provided fund investment and saving opportunities, and the purpose of the page was centered around this. It included, but was not limited to, information about helping the user, or customer in this case, to start a fund savings account, to transfer money into funds or learn about fund service packages provided by the bank. The page was chosen based on its characteristic in being a service that required the users to fully understand the content of the text.

The content was written in Swedish.

The choice of content to evaluate by using the UX writing methods was chosen based on discussions together with the principal, Mobiento/Deloitte. Although several contexts were suggested, the incentive for choosing this particular application was based on two main assets. First and foremost, the fact that the application had a lot of textual content that was important for the user to properly understand in order to use the service made for a well-suited context to use the UX writing methods on. Subsequently, the agency had been a part of developing the user experience for the application, were still part of the design and development process and in contact with the client. This provided useful insights from both the UX designers at the agency and to some extent contact with the client.

3.3 Online Tool Objective and Development
To answer the research question and explore the liabilities and the advantages of using an online tool for conducting user testing, the in-person methodology had to be replicated in a digital format.

The online form tool was developed with the vision to keep it as similar as possible to the already established paper and pen-methods, while at the same time taking advantage of the digital format. Using an online tool comes with many advantages — consistency, anonymity and the reduction of cognitive bias, for example. The consistency of having the same instructions delivered in the same manner and performing the test in their own time on their own preferred location, minimizes the risk of cognitive bias. Cognitive information bias is when the past experiences of a user will in one way or another affect their approach in test participation. This can be an issue when testing for usability and is an important factor to consider. [20]. This will of course be the case in both user testing contexts. Cognitive ego bias, in contrast, when emotional aspects can influence the participation, will be minimized in the anonymity that comes by using an online tool. The users can stay anonymous while participating in the study, and although they are aware of who will analyze the results, they are reassured that their participation in no way will be analyzed in terms of their performance, but rather as a part of a data set used for validation. Even though the users are reassured in the in-person context as well, the online context might provide a sense of reduced influence in this regard. In this study, it is considered an advantage to be consistent not only to reduce cognitive bias, but for the validation as well.

Since the aspect of the design being evaluated was to focus on the textual content, an option would have been to use an established form tool and have the users look at the text. To gain further insights and have control over the visual structures and flow of the online tool, the decision was made to handle the development of a tool. The assumption of having everything in one application and not having to jump between pages was important in keeping the online context experience as similar as possible to the in-person context. The option to use an established tool was looked into, but deemed not good enough in this context. Since the objective was to look at the textual context from a usability perspective, it was important to bring along the other aspects of the methods that could affect the quality of the user data, even though these aspects were not necessarily the ones being evaluated. This includes the visual design and the structural hierarchies.
The initial section of the online tool included some information about the user, including age, gender and perceived level of knowledge about fund savings, corresponding to the questions and instructions that were asked and presented in the in-person user tests, followed by the Cloze deletion test and the Highlighting method.

3.3.1 Design Markup and Prototyping

A mockup version of the application already existed, courtesy of the client and the user experience designers at the agency, thus the textual content was already produced. To translate this into a distributable digital form, markup code was produced, meaning that the visual design replicated the intended design as a static page. As previously mentioned, having the text in a visual interface was crucial to get accurate user data concerning usability. The texts on the page were in markup as well, as opposed to the whole design being just a single image. The reason for this was to be able to perform the methods and to send the user data into a database. The online tool was developed with the intention of being used in a desktop setting.

Before applying the specific tools to use in the Cloze and Highlighting methods, some sketches were made on paper to try out different layouts and get an estimate of where the tools would work without being too intrusive to the design. In order to have permission from the agency to use this design of the selected application, all the bank brand assets had to be removed. The branding was replaced in terms of color, logotypes, fonts and brand specific images. All visual structures and textual contents were kept intact.

3.3.2 Cloze Deletion Method

To replicate the open-ended Cloze deletion method, as described in previous section, 40 words were blanked out in the text on the page, with a number attached to each blank (fig 2). This number was also found in a corresponding input field at the bottom of the browser window, allowing the user to read through the text and fill in the blanks with a word that they considered to be a suitable fit for the context.

On the actual test page the value words were also placed in connection to each color choice. The word “Misstro” (Mistrust) under the red color and the word “Tillit” (Trust) placed under the green color (fig 4). Upon choosing a color for a highlighted section, a small red and green circle appeared next to the marked text, giving the user a choice to switch between green and red. To remove the highlight, you had to reload the entire page, removing everything and starting over, as one would have to do with a real pen and paper. The choice to switch between colors was added as this is something you can do with a highlighter in real life, adding the right color on top of the other. The Highlighting method was presented as the second user test, so that it would not produce any bias before performing the Cloze deletion method.
3.3.4 Tool Development

To be easily accessible to all users the digital tool was built using web standards — HTML for markup, styling with CSS and JavaScript both for dynamically rendering the input fields, controlling the highlighting feature and other interactive elements as well as sending data to the backend. The data was stored in a MySQL database controlled by a layer of PHP which also handled the user sessions. To make the tool easy adaptable it was built on top of an interactive web mockup exported from Adobe Experience Designer.

3.4 Pilot Study

In order to assess the UX writing methods and to ensure unnecessary flaws in the execution, a pilot study was performed with one user. The user did the in-person test, which included looking through the texts and layout both on a display and printed out on paper. This was also a way of confirming that the instructions were clear and that the user knew what to do in each step. Since the hypothesis included an assumption was that the performance of the digital tool would closely replicate the in-person user tests, it was only the content and instructions that was evaluated in the pilot study, as opposed to the functionality of the digital tool itself.

3.5 Collecting User Data in Two Contexts

The methods, replicating each other, consisted of a number of steps. Initially, the user was provided with an overview of the two methods they were about to participate in. They were assured that they were never the ones being tested, but rather the content in the design. They were informed that they were going to look into a page about fund savings, in a fictional bank called “the Linnea Bank”.

Before instructions of the two methods, the users were asked to provide their age, their gender, their occupation and their perceived level of experience of fund saving services, on a set scale of none, some level, medium level or high level experience. The first part of the test consisted of the Cloze deletion method. Here, the text had 40 spaces where the original words had been blanked out, at an approximate level of every 10th word. The users were provided with a corresponding answering sheet with 40 blanks, where they could write their proposed word.

The instructions were to read the text and suggest one word that they considered would fit. They were instructed that only one word was expected, but if they felt that two words were needed, they were allowed to provide both. If they were ambivalent between several words, they were encouraged to pick only one.

The second part of the test consisted of the Highlighting method, where the same content was displayed, with all the words present. The online context presented everything on the webpage, the in-person context was pen and paper-based for this method. Here the users were instructed to highlight words and sentences in either green or red. Highlighting text in green indicated feelings of trust towards the bank and the services presented on the page. Highlighting text in red, on the other hand, indicated feelings of mistrust. The users were encouraged to highlight all textual content that contributed to their perceived feelings of trust or mistrust, including texts on buttons and menu items.

3.6 In-person and Online User Study

To validate the quality of the user data, a control study was set up with in-person user tests. The two methods, the Cloze deletion method and the Highlighting method, were performed in-person in the same manner and as similar as possible to how the online tool was set up. The users received the same instructions and were asked the same questions, as described in the previous subsection. The participants who took part in the in-person context received the instructions verbally, carefully rehearsed to be of the same character as the instructions provided in the digital tool.

The Cloze method displayed the texts and blanked out words on a computer screen. The users wrote down their proposed words on paper, for the situation to mimic the online context, as opposed to filling in the gaps with words directly into the text. This was also for the users to get a sense of how the design looked on a display. For the Highlighting method, the texts were printed out, still in the intended visual design structure, and the highlighting was performed directly on the paper. Since the content did not change between the methods, it was deemed reasonable for one of the methods to be presented on screen and one to be presented on paper. As mentioned in Theory and Related Research, displaying the content in its intended design was an essential aspect of the two UX writing methods.

The online tool was distributed to a user group that corresponded with the user group for the in-person user tests, in terms of number and age span, to be able to test for validity.

3.7 Participants and Target User Group

20 participants were contacted and asked to participate in a study about text as a part of evaluation the user experience for a functional service application. 10 users were scheduled for a meeting in-person and 10 users were sent a link to the digital tool.

For comparability, the target group was limited to people ages 24-32 in both user groups, with varying levels of experience of fund savings and with different occupations. The goal was to recruit participants of both genders, and of the total 20 users, 45% were women and 55% were men. In the in-person user group, 50% were women and 50% were men. The digital user group consisted of 40% women and 60% men.
4. RESULTS

This section contains the results from the user study. The results will be divided into subsections, where the results from the Cloze deletion method will show quantitative results presented with numeric data set analysis. The results from the Highlighting method will show qualitative results with a thematic analysis. There are some aspects that are important to keep in mind when studying the results. Both methods result in a measure of the textual quality in terms of usability, where the Cloze deletion method provide insights as to how readable the text is and how the contextual quality helps the reader understand the content. This is presented with a score, which measures the amount of words matching the original words. The Highlighting method provides a value based assessment of the words and expressions, in this case trust and mistrust, and points to which aspects the users feel contribute to the service’s overall impression.

The aim of this study is to validate the quality level of the data sets from the different user groups, digital and in-person, and not the method scores per se. To validate the quality level, the method results will be compared and later discussed. Therefore, it is outside the scope of this study to evaluate the results from the methods, and that the evaluation rather will be between data sets.

The results from the pilot study included minor consistency errors and visual errors connected to browser functionality. This was corrected for the actual study.

4.1 Comparing User Groups

There were 20 participants who took part in the study. Having user groups as equal as possible is the foundation of this study, minimizing the aspects that could affect the results based on anything other than the difference in the online and in-person context. The aspects that could affect the data were age span, occupation and the level of education — factors that could influence the vocabulary size and therefore the quality of the data. Level of perceived experience of fund savings and gender were also data points collected about the users, although the recruitment of the users were not made with these factors considered.

In the digital user group, the level of perceived experience of fund savings were spread out from no experience to a high level of experience. In the in-person user group, the level of experience ranged from none to some level of experience. The ages ranged between 24-32 in both user groups and the users had various occupations. All the participants had either several years of higher education or a finished higher education degree. The data can thus be equally validated for low level scores as high level scores, as long as it corresponds between user data sets.

4.2 The Cloze Deletion Method

To test for validity between the in-person user tests and the digital user tests, the similarity between the test scores are the sought values, which can entail both similarly low or high test scores. Normally when using the Cloze method to improve a text in terms of usability, synonyms and misspellings count as a match. The scores in this study, however, are strict. This means that only exact matches has contributed to the score since the subjective decision for what a synonym can entail might affect the possibility to compare the result. Misspellings of the original word were taken into account.

The percentage of correctly restored words form a metric to measure success, where 60% [33] or above are considered successful for synonyms and misspellings, and lower for a strict comparison.

The in-person user group, for the control study, had an average Cloze score of 44.0%, which means that their proposed words matched the exact original words on a 44.0% average. The users in the digital user group had an average Cloze score of 47.5%, matching the exact original words with a score of 47.5%.

Table 1. Strict Cloze Scores in both contexts

<table>
<thead>
<tr>
<th>In-person Cloze Score</th>
<th>Digital Cloze Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.0 %</td>
<td>47.5 %</td>
</tr>
</tbody>
</table>

Looking at one word at the time, the data sets from both user groups were objectively similar. The similarity between the sets for each word was apparent, and there were several instances where the data sets were close to equal, both for equally high and low Cloze scores.

One example of a high Cloze score with two similar datasets was one word where 9 out of 10 users in each user group had proposed the blank with the exact same word as the original, and the remaining user in each group had responded with a unique word respectively. This was likely due to the context of this particular word, where there were few options of words that would fit into the sentence. Although the unique words in the two user groups were not the same, it can still be considered same level quality for the two datasets. The word choice can vary based on the context of the sentence and the user’s interpretation of the context.

Another example of a similar data set, but with a low Cloze score, is one instance where there were 9 respectively 10 unique words in each data set. One user in the in-person user group had proposed the exact word as the original word, and the other 9 users had each proposed a unique word. In the digital user group dataset, 8 users had proposed 8 unique words, and 2 users proposed the same word, although no user proposed the word from the original text. Some of the words corresponded between data sets, but the context of the sentence made each word relevant, leading to each data set having the same quality level.

This is significant — both the data set with the similar type of data with the high Cloze score and the data set with the widely spread data and therefore a low Cloze score are to be considered equal for the intents and purposes of this study. As previously mentioned, the evaluation does not include an evaluation of the content in itself, but the validation of the dataset quality. This is why it is important to look at both low and high Cloze scores for validation.

4.3 The Highlighting Method

The nature of this study required a brief thematic comparison of the Highlighting results to be able to validate the quality level of the data sets from the two different contexts. The amount of highlighted words and what was highlighted was compared in a broader sense, since the dataset did not suffice for statistical validation. Hence the thematic nature of the validation.

4.3.1 User issues

The online tool presented some issues when distributing to users. Due to browser related malfunction, some users claimed they were having trouble using the highlighting tool in some of the paragraphs. The users who were verbal about having difficulties could be disregarded, and a new user was recruited in their place.
4.3.2 Highlighting results

The amount of highlighted words, expressions and sentences differed in quantity between the two user groups. The digital user group had overall less highlighted words, expressions and sentences than the in-person user group. The level of highlighted sections were greater at the beginning of the content. The division between trust and mistrust was also different between user groups, where the in-person user group generally highlighted more mistrustful (red) sections. The trustful (green) highlighted sections were more similar in quantity.

The sections that the in-person user group highlighted as trustful was also highlighted as trustful in the digital user group. The sections that the digital user group highlighted as mistrustful were highlighted as mistrustful by the in-person user group as well. This implies that the same sections were perceived as trustful or mistrustful in both user groups. Both user groups highlighted expressions that can be interpreted as vague or unclear in red (mistrust), along with misspellings and colloquial language, but also content that were direct calls to action. Additionally, both user groups highlighted negative connotations as mistrustful.

Example of a sentence highlighted as mistrustful (red):

“Transfer your ISK savings to our ISK savings without taxation” (translated)

“Flytta ditt ISK sparande till vårt ISK sparande utan beskattnings” (original)

Both user groups had highlighted expressions that are reassuring, concise and refers to the customers own customization of using the service in green (trust). They had also highlighted words that were a part of the educational aspect of the page, where there were points that describe how fund savings work in short concise sentences.

Example of a sentence highlighted as trustful (green):

“See how much a fund savings account can generate over time with our fund calculator” (translated)

“So hur mycket ett fondsparande kan generera över tid med vår fondkalkylator” (original)

There were also sections that were highlighted as both green and red depending on the user, which occurred in both user groups. The ambiguity is important for the Highlighting method, as it shows what might be individual for users and possibly worth clarifying for better inclusion. For comparability, the same parts of both user group should be ambiguous, which is the case in this study.

Example of a sentence highlighted as both green and red from different users in both user groups:

“[The bank] is and want to be a sustainable and ethical company” (translated)

4.4 Analysis

The scope of this study does not include exploring the results in terms of which text or other content that needs to be improved in order to increase the level of usability. The purpose of this study is to explore and validate the use of a digital tool compared to in-person user tests, which is the focus of this section.

Validation includes the comparison of the data sets, its similarities and differences, and that is also why it is important to look into the potential factors that could have affected the results in either user group. Different factors could have affected the results of the two user groups, where some are related to the individual user and some might be related to the difference between the online tool and the in-person user testing. As a foundation for discussion, this section will assess what might be a consequence of the online setting and what is individual for different users.

In the Cloze deletion method, for example, both user groups had a score slightly below 50% matching words. Even though it is considered successful to score above 60%, this was measured strictly. The level of success is of less importance than the similarity between the data sets. To validate the level of similarity, the particular words must be taken into account as well. For both data sets, some of the non-matching words either included very similar synonyms, or were part of a sentence structure where the context itself could not provide a clear choice for the words, which resulted in the users providing words based on personal preference. Despite this difference, the data sets of the Cloze method can be considered to be of equivalent quality.

The Highlighting method resulted in the same type of content is labeled trustful or mistrustful, the amount of highlighted content is notably higher in the in-person user group. The difference between the user groups were to a greater extent quantity based. The number of highlighted sections in the different user groups were comparable, but not equal. The digital user group had fewer sections highlighted than the in-person user group. As suggested by Yu and Sung [29], hiding identities in an online setting may influence a lower degree of engagement, which is also indicated by the results of the Highlighting method. This suggests that the anonymity of the online participation might affect the qualitative methods, although this could be compensated for with the larger participation rate that comes with distribution opportunities of the online tool. More on this in the following section.

5. DISCUSSION

The purpose of this study was to validate the quality of the results from two UX writing evaluation methods in two different contexts — through an online user form and through in-person user tests. The two UX writing evaluation methods were of both qualitative and quantitative character, to cover two areas of evaluating usability in textual content. This section will discuss the above presented results and what this entails within the UX writing area. The research question, “When evaluating the usability of textual content, what are the assets and liabilities with using an online testing tool compared to in-person user tests?”, provides a foundation for this discussion. There are assets and liabilities to be considered for both user contexts, online and in-person, and will be accounted for in the following sections. To validate the quality of the data sets, it is important to look at the results and what aspects of each context that could have affected them. It is also of great importance to consider both good and bad scores of these two
methods. The methods and their effectiveness will be discussed in a broader sense in order to support the validation, and not necessarily to draw any conclusions regarding the methods specifically.

5.1 Liabilities and Assets
There are liabilities to be considered in using an online tool. As previously mentioned, the sense of anonymity may cause a lower degree of engagement than in-person settings, causing the data set to be smaller. The results indicate that this is the case in this study. Additionally, the results from the Highlighting method contained a higher level of engagement at the beginning of the page, suggesting that the content might have been too long for an online context, adding to the lower level of engagement [29].

The in-person user context in this study was to be considered as a control study to be able to validate the use of an online tool for the UX writing methods. Therefore, the assets and liabilities for the in-person user context will be centered around the correspondence with those of the digital context. The main liabilities with an online context consists of a lower degree of engagement for qualitative methods and the risk for confirmation bias in both qualitative and quantitative methods.

An in-person user context might add to the confirmation ego bias, based on the fact that the test leader is overseeing the participation in a more direct manner than in the online context. Despite this, the nature of the methods in this study overall adds to a reduced risk for confirmation bias as of consequence of their nature. In this case, the usability testing is more data-driven and direct than other usability testing methods, where bias is more difficult to avoid. Both the Cloze method and the Highlighting method requires the user to answer questions and provide data in order for the UX writer to draw conclusions based on the data, rather than the user directly giving feedback on the usability of the application. The tier between the user and the consequences of the data analysis makes for less bias in both an in-person and an online user context.

With these aspects in mind, it can be argued that for the intents and purposes of this study, the quality of the data sets are to be considered equal to a satisfactory degree. Some factors are important to keep in mind, as mentioned above. Certain aspects are difficult to avoid completely in any type of usability testing, like confirmation bias. This is also something that is secondary to this type of usability testing, based on the data-driven nature of these methods. While it is important to look into these aspects, some aspects can be related to the individual user as well. Further studies of these particular data sets with the user data might show correspondence between age, education or perceived level of knowledge in the area and their answers.

At the beginning of this study, it was assumed based on the literature review that it was essential that the content of a page in its entirety was presented to the users. Even though it is important to keep the content in its intended visual structures and right context [7], the results suggest that the online context made it difficult for the user to keep up the level of engagement, as previously mentioned. This could have been adjusted to reduce this liability aspect. The length of the content might have kept the user’s attention more successfully if it had been, for example, divided into several pages. Even though this might be considered part of the method criticism, it is a significant deduction of the results that is part of the discussion as well. Having identified this as the main aspect that would have to be adjusted in order to be able to use the online tool as a replacement for the methods in an in-person user context, this opens up the discussion regarding its potential.

5.2 Online Tool Potential
The potential of using an online tool in testing textual content for usability is first and foremost the ability to distribute the methods to a larger user group. The available resources are used in a more efficient way, more users can be tested at the same time, making the time for the UX writer to explore other aspects parallelly.

Crowdsourcing services, like the Amazon Mechanical Turk, can be used for survey participation tasks, like the Cloze deletion method, to optimize resources [22]. The potential drawback of using Mechanical Turk or similar services, which can apply to all web-based testing, is the anonymity and lack of supervision. It can cause lower level of engagement, which is why it is important to keep the length of the content down to a certain extent. The low cost of using a crowdsourcing service can outweigh this disadvantage.

In this study, the content was very specific and presented in a minor global language, Swedish. The potential for something that is written in English adds to the benefits of using crowdsourcing for data collection. Even without using a crowdsourcing service, the time it takes to recruit users through personal networks at the workplace, or educational facilities, is advantageous compared to the required time to meet with users in-person. A common courtesy when booking appointments with users is some level of compensation, which usually entails something to make the user’s time worthwhile.

Apart from the resource optimization in terms of time and resources, the distribution factor is key for this type of data-driven usability testing, where larger data sets provide a better foundation and more precise insights for the UX writer. In addition, some level of subjectivity must be taken into account considering a smaller data set, which can be difficult to decipher. In smaller datasets, there is always the possibility of individual user divergence. With a larger dataset, multiple instances of a highlight or a suggested word can more likely mean that this is something to act upon in order to increase the usability.

5.4 Future Work
The results of this study bring out potential for future developments and possibilities. Knowing what we know now, there are further steps that can be taken to contribute to the UX design community in general and to the UX writing community in particular. This section will propose some examples of future work within this area.

Since the distribution factor is so interesting, a natural direction to take the online tool is to rebuild it so that it can be adapted to other contexts, rather than this particular page about fund savings. Other UX writers could export their mockups, with Adobe Experience Designer for instance, into the tool and use the Cloze and Highlighting methods on their content. Further, the tool could be adapted into a content management system (CMS) interface where users would not necessarily have to have knowledge of coding to be able to use it. With the CMS and adjustability in place, any UX writer could start evaluating service or product content on a larger scale by using a crowdsourcing service.

Having a large data set for each product is significant for the UX writer responsible for the usability in that particular product or service. This requires ambition and energy for each piece of user experience, and this is where the another possibility of the larger dataset comes into mind for the future — sharing it with others. Collecting the data in a database with relevant user and service
context information, age, gender, education, occupation and other generally important user data points together with information about in which context they tested for usability, would be considered useful for other UX writers as well. This database could be accessible within the community.

Taking it an additional step further, having a dataset spread over multiple context for different users and all the information needed about them, a machine learning model could be trained to learn patterns of users in the contexts. This might be counterproductive for an aspiring UX writer, but it is exciting enough to mention as a future development regardless.

5.4 Method Discussion

The purpose of this study was to validate data quality between two contexts, an in-person and a digital. The discussion in this paper is mainly based upon the assumption that the data is of sufficient quantity to validate the use of an online tool, but the study would have benefited greatly from having larger user groups in both contexts. The validation could have been validated statistically as well, which would have contributed to a necessary layer of accuracy.

The results and discussion are also to an extent based on the assumption that none of the users had significant issues with the digital tool and were able to provide all the input they desired and that was asked for. The users who did have difficulties with the highlighting tool were disregarded, however, the fact that there were issues must be regarded in the method discussion. Consequently, there could have been cases where the users did have functionality issues yet neglected to mention and settled for highlighting parts of the texts. Even though research point towards a tendency for lower engagement in online situations, this could have played a part in this matter as well. If the study were to be executed once more, the digital tool could have been tested and evaluated to a greater extent. There is also the question of controlling the environment where the users participated in the study. The in-person user tests were scheduled for a certain time, but the online tool came with no instructions regarding a time frame. Users could start the participation and finish at a later time. This could, in addition to previously mentioned disadvantages, be one of the reasons for lower level of engagement in the qualitative Highlighting method.

Having the content of the functional service application debranded might also have affected the user input in both contexts, but since this was outside the scope, it was not considered to be of importance.

6. CONCLUSION

This study includes two UX writing methods for a defined functional service application. By performing the UX writing methods through an online form as well as in a user interview setting, the validity of transforming an established method into an online tool was investigated. The liabilities of both contexts were evaluated and future developments are suggested. To summarize, in-person user tests require resources in terms of time for the UX writer, finances in terms of compensation for the users and result in a smaller data set. Digital user tests still require time and resources, but can generate a much larger data set, seeing as it provides a data set of the same quality level as the in-person user tests.

The research question, “When evaluating the usability of textual content, what are the assets and liabilities with using an online testing tool compared to in-person user tests?” has a multifaceted answer. Considering the advantages of an online tool, it can provide a valid replacement for in-person user tests, if one considers that the assets outweigh the minor liabilities. However, the potential of future collaborative datasets and all the further developments suggested in this paper are worthwhile investigating further.

ACKNOWLEDGMENTS

A special thank you to my supervisor at the institution, Björn Thuresson, for his insightful, constructive feedback and continuous support. Another thank you to my supervisors at Deloitte/Mobiento for their support and for believing in my ideas. Finally, my sincerest gratitude to all of those who participated in my study, both in person and online, and to those who read through this paper not only one, but several times during its process.

REFERENCES
