Influence of Annotation Media on Proof-Reading Tasks

Andreas Schmid
University of Regensburg
Regensburg, Germany
andreas.schmid@ur.de

Marie Sautmann
University of Regensburg
Regensburg, Germany
marie.sautmann@stud.uni-regensburg.de

Vera Wittmann
University of Regensburg
Regensburg, Germany

Florian Kaindl
University of Regensburg
Regensburg, Germany

Philipp Schauhuber
University of Regensburg
Regensburg, Germany

Philipp Gottschalk
University of Regensburg
Regensburg, Germany

Raphael Wimmer
University of Regensburg
Regensburg, Germany
raphael.wimmer@ur.de

ABSTRACT

Annotating and proof-reading documents are common tasks. Digital annotation tools provide easily searchable annotations and facilitate sharing documents and remote collaboration with others. On the other hand, advantages of paper, such as creative freedom and intuitive use, can get lost when annotating digitally. There is a large amount of research indicating that paper outperforms digital annotation tools in task time, error recall and task load. However, most research in this field is rather old and does not take into consideration increasing screen resolution and performance, as well as better input techniques in modern devices. We present three user studies comparing different annotation media in the context of proof-reading tasks. We found that annotating on paper is still faster and less stressful than with a PC or tablet computer, but the difference is significantly smaller with a state-of-the-art device. We did not find a difference in error recall, but the used medium has a strong influence on how users annotate.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; • **Applied computing** → **Annotation**.

KEYWORDS

annotaion, proof-reading, digitalization

ACM Reference Format:
Andreas Schmid, Marie Sautmann, Vera Wittmann, Florian Kaindl, Philipp Schauhuber, Philipp Gottschalk, and Raphael Wimmer. 2023. Influence of Annotation Media on Proof-Reading Tasks. In *Mensch und Computer 2023 (MuC '23), September 03–06, 2023, Rapperswil, Switzerland.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3603555.3603572

MuC '23, September 03–06, 2023, Rapperswil, Switzerland

ACM ISBN 979-8-4007-0771-1/23/09.
https://doi.org/10.1145/3603555.3603572

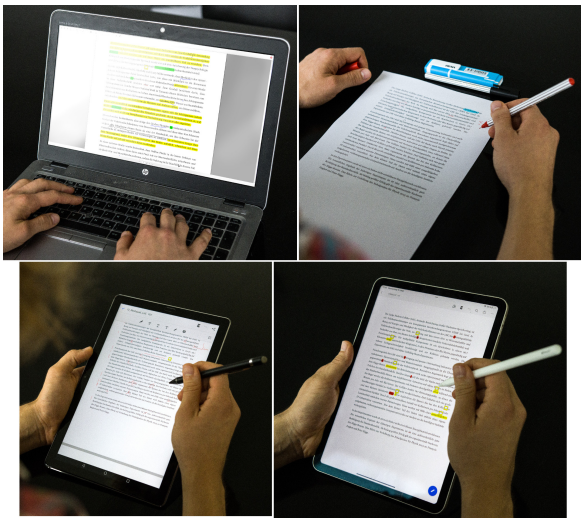

Figure 1: Proof-reading a text on different media: laptop, paper, and two tablets.

1 INTRODUCTION

Annotating documents is a common task, which today is frequently done digitally. That has some advantages, such as making annotations easily searchable or enabling collaboration on annotation efforts for studying or research. However, little research has been done on the interaction between annotation tool and user. It is unclear how digital annotation tools change users' annotation behavior or influence their performance on the task at hand. There is also evidence in literature that users perform better when proof-reading on paper, opposed to digital devices [45]. Annotating documents can support different reading goals, so it is plausible that the medium in which annotating takes place will also affect reading goals.

When writing about personal document annotation, researchers rarely give a formal definition. However, Ovsiannikov et al. [35] provide a useful working definition which has been expanded by Desmoulins and Mille [9]:

> *"Annotations consist of an anchor (the specific part of the text they relate to, not necessarily where an annotation is made), their visual form (markers, underlining, different colors etc.) and their content (e.g. notes, scribblings). They are additions to a document rather than edits of it."*

Marshall [30] studied annotation behavior on paper to build a framework that can help to develop digital annotation tools. Their observations have shown how difficult it is to transfer the expressiveness and variety of paper annotations to the digital world [29]. While the form of an annotation usually follows its function, annotations on paper can be "endlessly inventive" [28]. Researchers recognized early on that reading and annotating on paper and on screens both had advantages, as shown by efforts to combine the best of both worlds with projects such as the *Xlibris e-reader* or paper augmented digital documents [17, 37].

In a literature review on annotation technology, Wolfe noted that in order to make annotation technology a pervasive part of reading, stylus-based input for handwriting and collaboration features would need to be improved [47]. Much of the research efforts on digital annotating since then has been focused on supporting learning efforts and enabling collaboration, for example by sharing annotations between students and teachers [13, 25]. Even though stylus-based handwriting has certainly improved in the last two decades, it is still inferior to writing on paper. The stylus-enabled touchscreen device, while increasingly affordable and popular, has not yet replaced desktops and conventional laptops in office settings.

With this paper, our goal is to find out how digital media compare to paper during annotation tasks. To investigate annotation behavior and proof-reading performance with modern devices, we conducted three within-group user studies (one remote study and two lab studies). With our studies, we address the following research questions:

- ***RQ₁*** *"How does the medium used influence proof-reading performance?"*
- ***RQ₂*** *"How does the medium used influence annotation behavior?"*
- ***RQ₃*** *"Do remotely conducted proof-reading studies yield valid results?"*

We found that paper outperforms digital media in terms of task completion time and perceived task load, but modern tablets come in as a close second. Error recall and total number of annotations was similar between conditions, but on each device, different types of annotations were preferred. Our work contributes to the existing body of research on annotation behavior and proof-reading performance with state-of-the-art software and hardware. Furthermore, it is a first step towards systematically evaluating the feasibility of remote proof-reading studies.

This paper is structured as follows: We first summarize findings from former proof-reading and annotation studies. Then, we describe methodology and results of each of our studies separately. Finally, we consolidate our findings and discuss implications on future studies and practical application.

2 RELATED WORK

Sellen and Harper [39] argue that, because of its unique affordances, paper will continue to be used in office contexts. With paper, users are not restricted by anything but its physical properties and their own creativity. It can be used flexibly – one can add sketches or notes, it can be used with different pens, and folded or ripped apart if a smaller sheet is needed. Documents can be arranged in physical space to work with multiple documents simultaneously [34]. Steimle et al. [43] investigated the differences of physical and digital media in note-taking tasks during a university course. They found that students preferred paper for reading and the option to quickly take notes. On the other hand, digital media have the advantage of quick and easy document sharing (for example via e-mail), context search and easy modification of content. Additionally, there is no need to print documents and all information can be on one device [43].

2.1 Reading on Different Media

Many past studies investigated how reading differs between paper and digital screens. In their literature review from 1992, Dillon [10] claims that such reading studies have to distinguish between *reading process* (e.g. gaze, navigation, manipulation of the medium) and *reading results* (e.g. reading time, fatigue, content understanding). They found that the trend in such studies went towards investigating only the *results* and were oftentimes limited to only few dependent variables [10]. Overall, studies about reading on different media yielded mixed results. Muter et al. [33] compared two hour reading sessions from book and TV in 1982. In both conditions, their participants experienced fatigue and dizzyness, but there was no significant difference between media. There was also no difference in content understanding. Reading was generally slower when reading off the screen, but authors claim that this effect could be due to participants being unfamiliar with the medium.

More recently, reading studies with modern devices, such as tablets or e-readers, have been published. Baker et al. [2] and Schugar et al. [38] could not find a difference in content understanding between paper and e-readers. In contrast, Jeong [21] found that reading on paper leads to better content understanding and less exhaustion in comparison to e-readers. In an eye-tracking study, Siegenthaler et al. [41] found that reading speed was similar on paper and e-readers. However, shorter fixation duration indicated better reading performance on e-readers in certain situations.

In 2014, Korwisi [24] compared annotation tasks on paper and tablet. The iPad performed better for writing-intensive annotations. However, reading on paper was rated less exhausting by participants.

Chen et al. [5] investigated the influence of familiarity with a medium on digital reading. They compared content understanding between paper, tablet, and PC in a between-subject study. Even though they could not find an influence of the familiarity with the medium, the paper group performed better overall. In a similar study, Singer and Alexander (2017) [42] investigated whether there

is a difference in performance when reading digital or paper articles. They also compared the study's results with participant's self-assessment on which medium they would perform better. Participants expected to perform better with digital media, but some relevant information could be reproduced better in paper condition. In a study with 1st to 6th grad pupils, Lenhard et al. (2017) [27] compared reading on paper with a monitor. They found that their participants were reading faster but understood content less accurately in the digital condition. Støle et al. [44] confirm those findings in a study in 2020. Children performed significantly better in a reading comprehension test when reading on paper regardless of previous reading competence. Most mentioned studies found that reading on paper leads to less physical strain and better content understanding than reading on digital media. However, many of those studies focused on qualitative aspects or tried to use abstract measures such as content understanding.

2.2 Proof-Reading Performance

From the 80s on, with the proliferation of computers in the workspace, many researchers compared reading on digital media to reading on paper [6, 15, 16, 33, 46, 48]. In several studies, proofreading performance was used as an indicator of general reading performance. Early studies have found significant differences in reading speed, with participants reading printed text faster than text on CRT displays [16, 33]. Gould and Grischkowsky's study from 1984 featured proof-reading without annotation, participants pointed out errors via a light-pen instead. They found that participants proofread 20-30% faster on paper than on CRT monitors [16]. Wright and Lickorish (1983) found that both speed and accuracy (number of errors found out of all) were significantly lower when texts for proof-reading were presented on a screen [48].

To understand which factors influence proof-reading performance, later studies isolated individual variables. For example, Creed et al. [6] controlled for lighting conditions, latency, and screen resolution. Wilkinson and Robinshaw (1987) [46] asked participants to verbally signify errors in both screen and paper conditions, to avoid a confounding effect caused by the annotation process. They found that proof-reading accuracy and speed was significantly lower and fatigue increased faster when participants read on CRT monitors [46].

As those studies show, hardware from the 1980s was far inferior to using pen and paper for proof-reading. Acoording to Köpper et al. [26], these apparent drawbacks of reading texts on screens may be related to restrictions of technology used in those studies in the 1980s. Modern LCD and OLED displays feature higher resolution, better contrast, and less flickering than CRT monitors. Studies comparing the technologies report an improved performance in visual tasks done on LCD [26, 32]. Furthermore, people spend much more time working with computers and looking at monitors today than they did in the 1980s. However, research has shifted towards studying the differences in reading comprehension and deeper understanding of texts when comparing physical and digital media. Major apparent factors in this shift are the proliferation of digital learning materials in schools and universities, as well as the popularization of e-readers.

One of the few comparative studies in the LCD-era featuring proof-reading was conducted by Wharton-Michael in 2008 [45]. Eighty-four students, most often communication majors, proofread two short journalistic articles, one on paper and one digitally. They were instructed to make "any changes needed" on paper and used the "track changes" tool in Microsoft Word to correct errors in the digital condition. Each task was limited to eight minutes. Participants found significantly fewer errors in the paper condition, however, a significant effect of the read article and the proofreading condition was observed. Wharton-Michael suggests the participants' familiarity with the topic of one of the articles as a possible cause [45]. In 2016, Köpper et al. conducted a study designed to test whether the CRT-era findings about reading on screens could be reproduced. In an between-group experiment, 136 participants proof-read texts on a TFT-LCD screen or on paper by vocally signaling found errors and their line number. They could not find significant differences in proofreading speed and performance between the conditions. However, participants showed more symptoms of eyestrain in the screen-condition and a subgroup of participants that experienced both conditions strongly preferred proofreading on paper [26].

2.3 Annotation Behavior

There are few recent studies which compare physical and digital annotating. Agosti et al. looked at the usage of text margins for annotations, finding that margins play an important role when annotating documents, improving legibility of annotations and the overall speed of the process. They suggest that adding virtual margins to PDFs could improve the quality of digital annotations [36]. In a field study with researchers, Kawase et al. analyzed annotated papers and classified them by their reading goal [22]. They found that papers annotated for review and feedback featured more written notes and less highlighted text than papers annotated for learning or other goals. By comparing online annotations to paper annotations they observed that online annotations were often short and in the form of notes, while paper annotations where more often in the form of text highlighting. Despite this difference, these annotations often served the same function of signaling for future attention. They inferred that highlighting on paper was a way of reducing cognitive load and keeping focused on the task, while the interaction with keyboard and mouse when making digital annotations impeded that function [22]. Blustein et al. [3] investigated annotation behavior in a long term study over the course of 3 years. They collected annotations of diploma and bachelor seminars and classified them using Marshall and Brush's taxonomy [31]. They found that most annotations form compounds and almost all contained at least one text element. Therefore, authors claim that text is an essential part of annotations. Kim et al. [23] compared note-taking behavior on tablet PC and PDA with paper. Participants were asked to watch a 30 minute video and take notes to answer questions later. PDA notes were shorter when created free-hand, and, when using the keyboard, no symbols such as arrows or circles around words were added. Similar results were found for notes created with a tablet. On paper, participants used symbols and different forms of annotation to connect or separate elements. Cunningham and Knowles [8] conducted a medial comparison of annotation and

note-taking behavior of researches at an IT conference. They observed people taking notes, conducted semi-structured interviews, and investigated notes. Author's focus was to find reasons why and how notes were taken to derive recommendations for better note-taking. Among their subjects, paper was the preferred medium, but PDAs, tablets, and phones were also used. Text was short on all media, but with paper, many non-text notes, such as arrows, circles, stars, were used in the notes. Hastreiter et al. [20] investigated annotation behavior when digitally editing scientific text. They compared acquired annotations with the results by Marshall and Brush [31]. Furthermore, they compared 26 annotation tools for the *Apple iPad* and found that most of them use analog annotation metaphors or adapt workflows from desktop annotation tools to (multi-)touch input. In a user study with participants reading and annotating a text to answer questions later, they found that more highlighting was done with digital media, more underlining on paper. The number of notes was almost identical.

In conclusion, researchers studying reading speed and performance on paper vs. screen often used proof-reading as a performance indicator. However, applied annotation methods that are either outdated, or tried to eliminate annotation overhead for all conditions. Differences in annotating between paper and screen are not well-investigated, although there is some evidence that people will create different annotations depending on the tools at hand. To our knowledge, there are no quantitative comparisons of proof-reading performance between paper and state-of-the-art digital devices.

3 METHODOLOGY

To investigate the differences in proof-reading performance and annotation behavior on paper and digital media, we conducted three within-group user studies. In all three studies, participants were asked to proof-read texts and correct errors as if they would help a friend with an assignment. We used German Wikipedia articles with an "excellent" rating and added syntactic, semantic, and formal errors to them. We used different texts for every study, so participant overlap would not be a problem. For each study, we performed a pre-study to check whether used texts were similar enough to each other in terms of difficulty and readability.

The first study was an asynchronous remote user study comparing annotation behavior and proof-reading performance between paper and digital annotation tools (Section 4). Even though this study has high ecological validity due to participants using their own devices, internal validity is limited because of the remote design. Therefore, we conducted a follow-up study in a controlled laboratory environment, comparing annotation behavior and proof-reading performance between paper, PC, and an affordable tablet (Section 5). As many participants reported problems with the tablet used in the lab study, we replicated this study, comparing annotation behavior and proof-reading performance between paper and two tablets (Section 6).

We quantify proof-reading performance as a combination of the variables *error recall*, *task completion time* and *perceived task load*. Furthermore, we define annotation behavior as a combination of the tools used to create annotations, the number of annotations, and qualitative properties such as the style of annotations. Additionally, we evaluate the feasibility of conducting annotation studies remotely. To this end, we exploratively compare the results from the remote study to those of the lab studies to gather first indications on whether the results align.

4 REMOTE STUDY: PAPER VS. PC

To find out how annotation behavior and proof-reading performance differs between annotating on paper and on a digital media, we first conducted an asynchronous remote user study.

We prepared two texts based on "excellent" rated German Wikipedia articles[1] for our proof-reading study. We selected topics that were easy to understand without any previous knowledge, but which participants were unlikely to be familiar with. Both texts were shortened to one page each (539 words/4073 characters; 531 words/4056 characters) and formatted to DIN A4, left-justified text, 11 pt Calibri font and single line spacing. Then, we inserted 35 errors into each text: 7 misspellings, 7 punctuation errors, 10 formatting errors, 4 syntax errors, 4 capitalization errors and 3 duplicate words. We adopted this error type distribution from Wharton-Michael [45]. We increased the error density Wharton-Michael used in their study, which resulted in a normally distributed 20-95% error recall rate in a pilot study.

For our study, we used a within-groups design and counterbalanced combinations of text and medium, as well as task order with a balanced latin square. We prepared large format letters containing instructions and a printed text to annotate during the PAPER condition, which we sent to our participants. Detailed instruction and the text for the digital condition could be found on a website. Participants were instructed to proof-read and annotate the documents as if proof-reading a friend's thesis in a way that they would be able to correct the text based on their annotations. For the PC condition, participants were allowed to use any PDF reader capable of annotating. In case they had no personal preference, we suggested the browser-based PDF annotation app *Xodo*[2]. We did not prescribe the application to use because using a different reader than their usual one would decrease ecological validity. However, we explicitly asked participants to use a laptop or desktop PC. After each of the two proof-reading tasks participants were asked to fill out a NASA TLX questionnaire [18, 19] presented on the study website. After completing both tasks they were asked to fill out the Affinity for Technology Interaction (ATI) questionnaire [12] and a custom questionnaire on demographics, proof-reading experience, and their perception of the texts. After completing the questionnaires, participants uploaded the annotated PDF and a scan or adequate photograph of the annotated print document to our website. The process was designed to take roughly one hour altogether.

We recruited 29 (14 female, 15 male) participants via convenience sampling. Prerequisites for participation were to have no uncorrected vision impairments, no dyslexia and to have German as one's first language. On average, participants were 27 years old (sd: 8.33) ranging from 19 to 59 years. 55.17 % of participants were students.

[1] https://de.wikipedia.org/wiki/Meidum-Pyramide
https://de.wikipedia.org/wiki/Wilder_Mann
[2] https://pdf.online/

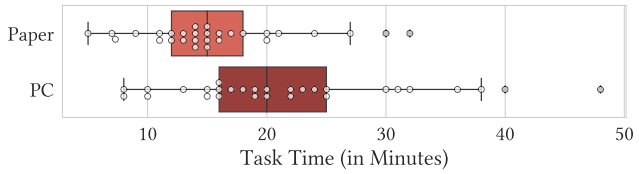

Figure 2: Task completion time in minutes for proof-reading on paper or PC. Participants were significantly faster on paper.

Table 1: Results of the NASA-TLX questionnaire for proof-reading on each medium. Task load was lower for annotating on paper on all scales. We found a significant difference for the total score, as well as multiple sub-scales.

NASA-TLX Scale	Paper Mean (std)	PC Mean (std)	t-test p-Value
Mental	56.9 (21.3)	60.9 (22.3)	0.278
Physical	**19.3** (14.4)	32.8 (28.3)	**0.011***
Temporal	**28.6** (20.1)	41.2 (26.7)	**0.002***
Performance	40.2 (26.8)	43.1 (27.3)	0.531
Effort	**48.6** (21.1)	59.0 (23.1)	**0.027***
Frustration	**31.2** (21.5)	43.4 (28.5)	**0.03***
Mean	**37.5** (13.6)	46.7 (15.8)	**0.003***

Results of the ATI questionnaire [12] indicate an above average affinity for technology (4/5) for our participants. Of our 29 participants, 13 used the *Adobe Acrobat Reader DC*, 11 used *Xodo* (the online tool we proposed), 3 used *Foxit Reader*, and *Microsoft Word* as well as the built-in PDF reader of *Microsoft Edge* was used by one participant each.

4.1 Results

Shapiro-Wilk tests have shown that error recall, task completion time and NASA-TLX scores were normally distributed. If not stated otherwise, we used a dependent-sample two-sided t-test for hypothesis testing.

Task completion time was significantly lower ($t(28) = 4.317, p = 0.0002$) for annotating on paper (mean: 15.7 minutes) than for the PC condition (mean: 21.9 minutes) (Fig. 2).

Perceived task load was lower for annotating on paper on all scales of the NASA-TLX (Table 1). We found significant differences for physical and temporal demand, effort and frustration, as well as the overall score.

We did not find a significant effect of the used medium on error recall (Fig. 3), neither in total nor for any specific type of error.

We analyzed annotated documents and counted and categorized annotations into the types available in most PDF readers (underline, strike-through, highlight, free-hand, comment). For paper documents, we classified annotations as *highlight* if a highlighter was used, as *underline* or *strike-through* if this was clearly the participant's intention, as *comment* if there was written text, and as *free-hand* for any other annotation, such as symbols. There was no significant difference ($t(28) = 1.357, p = 0.186$) between conditions

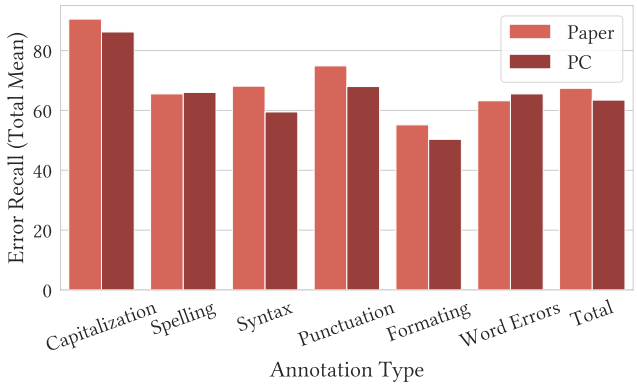

Figure 3: Percentage of errors found by error type for proof-reading on paper and PC. We did not find a significant difference for any error type.

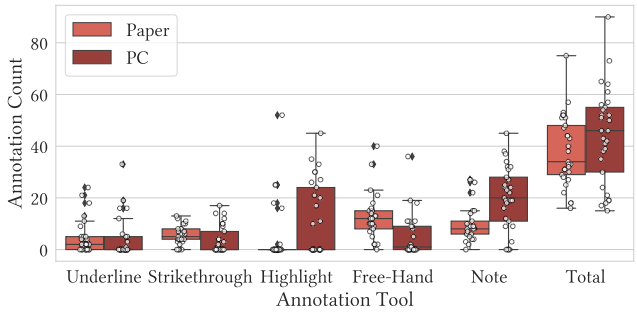

Figure 4: Amount of annotations split by annotation type while proof-reading on paper or with a PC. Depending on the medium, different types of annotations are preferred. On average, participants made more annotations with the PC. However, the difference is not significant.

for the total amount of annotations, but we found vast differences in annotation count for all individual categories except *underline* (Fig. 4).

In most cases, participants used only one annotation tool (e.g. one pen) when annotating on paper, while most digital documents were marked using three different tools. On paper, many participants used freehand symbols, such as an arrow, to indicate capitalization errors. In digital documents, similar errors were marked with the highlighter tool and a note clarifying the correction to be made, resulting in two counted annotations (Fig. 5).

As we did not have control over participants' environment due to the remote setting, we checked the collected data for potential distortion of results by confounding variables with results of our post-study questionnaire. Using independent-sample t-tests, we did not find a significant influence of task order on error recall ($t(48) = 0.477, p = 0.635$), task load ($t(48) = 0.609, p = 0.545$) or task completion time ($t(48) = 1.808, p = 0.076$).

The participant's assessment of text difficulty in the post-study questionnaire indicates that both texts were similarly difficult to read and understand. Independent-sample t-tests did not show an

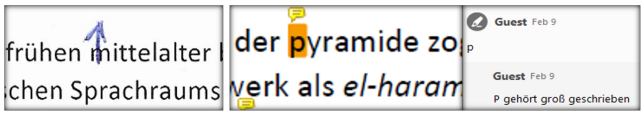

Figure 5: Example of a capitalization error marked on different media. On paper (left) a symbol is used to indicate the error. The digital annotation (center) is further clarified with a note (right).

effect of the text on which the proofreading was performed on any of our dependent variables (error recall: $t(48) = 0.100, p = 0.921$, task completion time: $t(48) = 1.048, p = 0.299$, task load: $t(48) = 1.239, p = 0.220$). Even though this result indicates that differences between texts did not confound our results, participant's assessment alone is not sufficient to be certain there were no influence. However, possible remaining effects should be counteracted by counterbalancing of text/tool combinations and order of conditions.

Overall, we found an influence of the proof-reading medium on task completion time, perceived task load, and annotation style. Annotating on paper was faster and less exhausting for our participants. We did not find a significant difference between media in terms of error recall and total annotation count.

5 LAB STUDY: PAPER VS. PC VS. TABLET

As the proof-reading study described in section 4 was conducted remotely, we did not have control over the participant's hardware and environment. Therefore, we conducted a similar study in a controlled lab environment to see if results were consistent with findings from the remote study. An additional *Tablet* condition was added to the existing set of *Paper* and *PC* as tablets are oftentimes used to annotate documents.

In contrast to the remote study, we compare annotation behavior and proof-reading performance with different annotation media in a controlled lab setting. Participants were asked to proof-read a text with each of the following media: PC with the *Adobe Acrobat Reader DC*[3] PDF reader, Tablet with the *Adobe Acrobat Reader Mobile App*[4] for Android, and Paper. For the texts, we used "excellent" rated German Wikipedia articles[5]. Those texts were shortened to about the same lengths (538 words/3944 characters; 534 words/3974 characters; 533 words/4133 characters), formatted according to our university's template for theses (11 pt roman font, justified text, 1.5× line spacing). Similar to other proof-reading studies [7, 16, 40, 45, 46, 49], we introduced 15 errors into each text: 4 misspellings, 2 capitalization errors, 4 syntax errors, 2 semantic errors, 3 duplicate or missing words. As there was one more condition than in the remote study, we reduced text length and number of errors to keep the total time for the study approximately the same. We used a within-groups study and counterbalanced conditions with a balanced Latin square.

[3]https://www.adobe.com/acrobat/pdf-reader.html
[4]https://www.adobe.com/acrobat/mobile/acrobat-reader.html
[5]https://de.wikipedia.org/wiki/Mungo_Man
https://de.wikipedia.org/wiki/Knickpyramide
https://de.wikipedia.org/wiki/Holzschnitt

The study was conducted in a controlled environment in a lab for user studies at our university. For the Paper conditions, participants were provided with a printed text and a set of writing tools (highlighters and sharpies in different colors, ball head pens, pencils with eraser head, ruler). For the PC condition, we used a lab computer with Windows 10 Education, a 24" display with $2560 \times 1440\,px$ resolution, and a wireless mouse and wired keyboard. For annotating, participants used *Adobe Acrobat Reader DC*. For the Tablet condition, we used a *HUAWEI MediaPad M5 10.8"*[6] (released in 2018) with Android 8 and an off-brand stylus. The tablet as a 10.8" screen with $2560 \times 1600\,px$ resolution.

Before the study, we asked participants for their demographic data, about their media usage, and their experience in annotating. Additionally, we used the ATI questionnaire [11] to find out about participant's affinity for technology. Before each condition, participants were allowed to familiarize themselves with each medium by testing its annotation functions on an unrelated text. We asked participants to tell us when they started and stopped annotating each text, so we could measure task time. Afterward, participants filled out the NASA-TLX questionnaire. After all conditions were completed, we asked participants about their preferred medium in a brief post-study interview. In each annotated document, we counted found errors and classified annotations based on Marshall's categories [28, 31]. Additionally, we documented which tools were used for each annotation. This evaluation was conducted by a single person to counteract inter-rater effects.

We recruited 36 participants (17 female, 19 male; aged 19 – 32, mean age: 24.5) for our study. All of them were native German speakers and had a university background. Their mean ATI score was 4.25 (range: 2.56 – 6), which is slightly higher than for an average population (3.5).

5.1 Results

If not explicitly stated otherwise, we used repeated-measures ANOVA to test for main effects and Bonferroni-Holm-corrected dependent-sample t-tests for post-hoc analysis.

Mean task time was 11 minutes for the *paper* condition, 15 minutes for *PC*, and 16.5 minutes for the *tablet* (Fig. 6). However, task time varied strongly in all conditions:

- paper: 5.0 – 19.3 minutes, mean: 11.4, sd: 3.67
- PC: 4.9 – 46.1 minutes, mean: 15.1, sd: 8.26
- tablet: 5.6 – 29.4 minutes, mean: 16.5, sd: 6.54

Normal distribution of task completion time was violated in the PC condition due to strong outliers. As the participants related to those outliers performed normally in both other conditions, we argue that this effect is caused by the annotation medium and could also occur in the wild. Therefore, we decided against removing those outliers from the further analysis and instead used the non-parametric Friedman test with post-hoc Wilcoxon signed-rank tests with Bonferroni-Holm correction. We found a main effect of the used annotation medium on task completion time ($\chi^2 = 30.7, p < 0.001$). A post-hoc analysis shows significant differences between Paper and PC ($p < 0.001$), Paper and Tablet ($p < 0.001$), as well as PC and Tablet ($p = 0.023$).

[6]https://en.wikipedia.org/wiki/Huawei_Mediapad_M5

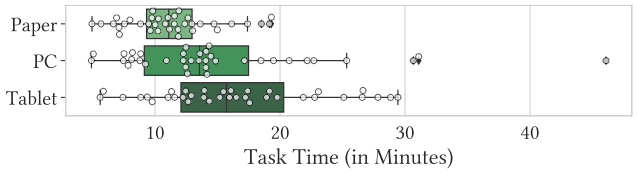

Figure 6: Task completion time for proof-reading on different media. Participants were significantly faster on paper.

Table 2: Result of the NASA-TLX questionnaire for proof-reading on paper, PC, and tablet. Task load was significantly higher for annotating with the tablet than for annotating on paper or the PC.

NASA-TLX Scale	Paper Mean (std)	PC Mean (std)	Tablet Mean (std)
Mental	51.2 (29.5)	51.5 (28.8)	60.8 (25.5)
Physical	15.4 (13.3)	21.4 (22.4)	46.9 (32.3)
Temporal	26.1 (22.4)	25.8 (22.9)	42.5 (27.1)
Performance	40.0 (21.2)	37.6 (23.3)	46.4 (25.9)
Effort	43.9 (26.8)	47.5 (25.3)	66.0 (20.4)
Frustration	26.2 (25.3)	29.6 (24.9)	62.2 (29.3)
Mean	33.8 (16.9)	35.6 (16.0)	54.1 (17.8)

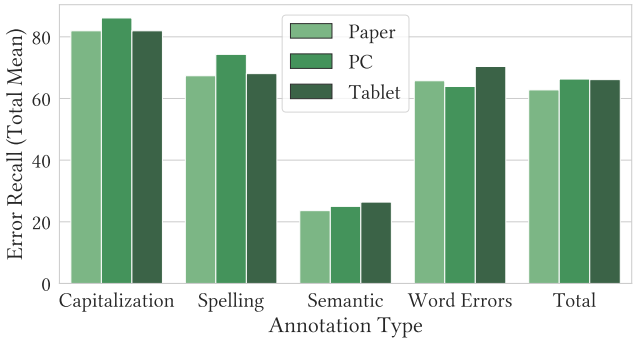

Figure 7: Error recall for proof-reading on different media in percent, split by error type. We did not find a significant difference between media for finding any type of error.

We found a significant main effect of the used annotation medium on perceived task load, represented by the total NASA TLX score ($F(2.0, 70.0) = 28.626, p < 0.001$). This effect is highly significant ($p < 0.001$) between Tablet and both, Paper and PC. We did not find a significant difference in task load between Paper and PC. Detailed results of the NASA TLX and its subscales can be seen in Table 2.

On average, participants found 10 of 15 errors with the PC and on paper, and 9 errors with the tablet (Fig. 7). At least one error was found every time, and all 15 errors were found only once. We did not find a significant main effect of the annotation medium on participants' error recall.

We found a significant effect of the used annotation medium on the number of created annotations ($F(2.0, 70.0) = 7.930, p = 0.001$).

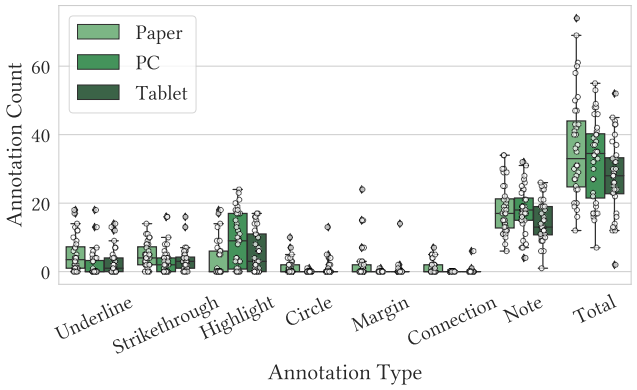

Figure 8: Number of annotation made while proof-reading on different media, split by annotation types. Some annotation types were preferred on different media. There was no significant difference in the total amount of annotations.

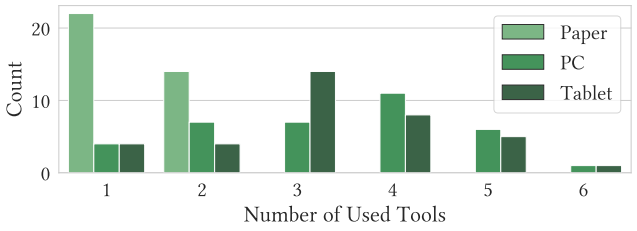

Figure 9: Number of different tools used with each medium. On paper, participants used only one or two pens. In both digital conditions, participants switched between tools provided by the system.

During the Paper condition, participants created more annotations than with the digital devices:

- paper: 12 – 74, mean: 36, sd: 15.33
- PC: 7 – 55, mean: 33, sd: 12.25
- tablet: 2 – 52, mean: 28, sd: 11.03

Post-hoc analysis has shown that the difference is significant between Paper and Tablet ($p = 0.026$). A detailed depiction of the number and type of annotations per condition can be seen in Fig. 8.

We observed that on paper, users stick with one pen and use it for different annotation types. On the PC and tablet, they switch between tools depending on the type of annotation they want to create (Fig. 9).

Among our participants, 61.1% preferred the PC for annotating (reasons: more space, undo), 44.4% preferred paper for its simplicity, 2.8% preferred the tablet.

Even though we found a significant difference in error recall between texts, we assume that this does not influence our other findings because of the counterbalanced conditions.

Several technical problems occurred with the tablet used in this condition, for example occasionally malfunctioning palm detection and unreliable tracking of the pen. Therefore, the tablet performing significantly worse than paper and PC regarding task completion

time and perceived task load could have been caused by this specific device and not the medium *tablet* in general. We further investigated this by comparing the *MediaPad* and paper to a state-of-the-art tablet in a follow-up study, which is described in section 6.

6 LAB STUDY: PAPER VS. MEDIAPAD VS. IPAD

The *Huawei MediaPad M5* used during the *tablet* condition in the previous lab study was prone to technical problems. Therefore, we conducted an informal exploratory study with four participants who regularly used a tablet for reading and annotating texts. The procedure of this pilot study was similar to the studies presented earlier in this paper: Participants were asked to proof-read two texts – once on paper and once with their own tablet and the PDF reader they were used to. The results of this exploratory study indicated a much smaller difference in task time between paper and tablet compared to the previous lab study. This finding reinforced our assumption that proof-reading performance was deteriorated by technical limitations of the *Huawei MediaPad* in our previous lab study. Therefore, we replicated this study with a state-of-the art *Apple iPad air 2022* with the corresponding stylus. In this replication study, we compared annotation behavior on paper, the *iPad*, and the *MediaPad* to explicitly measure the difference between the two tablets.

Again, we based the texts for our study on three "excellent" rated Wikipedia articles[7]. We shortened those texts to one page (297 words/2449 characters; 356 words/2620 characters; 344 words/2446 characters) and introduced the same types and amount of errors as in the original study. Like in the previous study, we asked participants to find and annotate errors in the three texts we prepared with errors. Each participant annotated one text on paper, one text with the iPad, and one text with the *MediaPad*. On both digital devices, the *Adobe Acrobat Reader* for the corresponding operating system was used. Text/tool combination and order of conditions was fully counterbalanced with a Latin square.

Before the study, we asked participants about demographic data, experience in annotating, and their affinity for technology using the ATI questionnaire. For both tablet conditions, we allowed participants to try annotating an unrelated document to familiarize themselves with the medium. After each condition, participants filled out the NASA-TLX questionnaire. After the third condition, we asked participants which of the three media they preferred for annotating.

For the replication study, we recruited 18 participants (5 female, 13 male; aged 22 – 36, mean 27.4). Our participants had a mean ATI score of 3.43, which indicates a slightly below average affinity for technology. None of them had prior experience in editing or annotating text on tablets.

6.1 Results

Participants completed the task fastest on paper (mean: 9.5 minutes), about two minutes slower on the iPad (mean: 11.5 minutes) and slowest on the MediaPad (mean: 14 minutes) (Fig. 10). As task times in the MediaPad condition were not normally distributed, we used

[7]https://de.wikipedia.org/wiki/Large_Hadron_Collider
https://de.wikipedia.org/wiki/Osterinsel
https://de.wikipedia.org/wiki/Werkzeuggebrauch_bei_Tieren

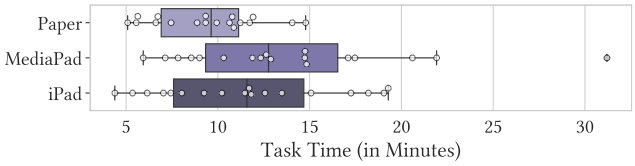

Figure 10: Task completion time for proof-reading on paper or two different tablets. There is a significant difference between all three media.

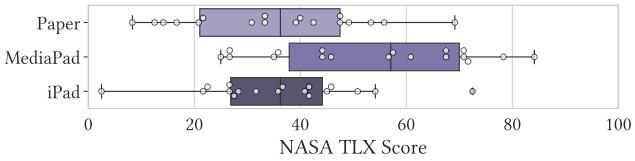

Figure 11: Perceived task load (mean NASA-TLX score) for proof-reading on paper or two different tablets. Task load was similar for paper and the Apple iPad, but significantly higher for the Huawei MediaPad M5.

Table 3: Results of the NASA-TLX questionnaire for proof-reading on paper or two different tablets. The MediaPad performed worse on all sub-scales of the questionnaire.

NASA-TLX Scale	Paper Mean (std)	iPad Mean (std)	MediaPad Mean (std)
Mental	50.6 (28.1)	47.8 (25.3)	60.0 (26.1)
Physical	23.3 (22.1)	28.9 (22.1)	48.6 (28.0)
Temporal	34.2 (22.8)	35.3 (24.9)	44.2 (26.2)
Performance	36.7 (24.1)	33.1 (23.1)	48.3 (29.1)
Effort	38.1 (22.0)	39.4 (18.1)	57.2 (23.2)
Frustration	29.2 (22.6)	33.1 (20.2)	64.7 (23.5)
Mean	35.3 (17.0)	36.2 (15.3)	53.8 (18.9)

the non-parametric Friedman test, which shows a significant main effect of the used annotation medium on task completion time ($\chi^2 = 10.778, p = 0.005$). Using Bonferroni-Holm corrected post-hoc Wilcoxon signed-rank tests, we found a significant difference in task time between paper and both tablets (iPad: $p = 0.018$, MediaPad: $p < 0.001$), and between the two tablets ($p = 0.018$).

A repeated-measures ANOVA shows a strong and highly significant effect of the used annotation medium on the NASA TLX score ($F(2.0, 34.0) = 26.628, p < 0.001$) Post-hoc dependent sample t-tests with Bonferroni-Holm correction have shown that annotating with the MediaPad resulted in a significantly higher task load than with both paper ($p = 0.012$) and the iPad ($p = 0.012$) (Fig. 11). Task load was almost identical for paper and iPad for the total score ($p = 0.86$), as well as all sub-scales, with a maximum difference of 1.1 points for *physical demand* (Table 3).

Error recall was very similar in all three conditions, the iPad performed slightly better than paper and the MediaPad (paper: 72.6 %, iPad: 78.5 %, MediaPad: 70.4 %) (Fig. 12). However, a repeated measures ANOVA did not show a significant effect ($F = 3.17, p =$

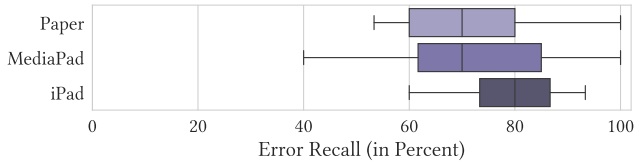

Figure 12: Error recall in percent for proof-reading on paper or two different tablets. On average, participants found 6 % more errors on the iPad than on the other two media. However, this effect is not statistically significant.

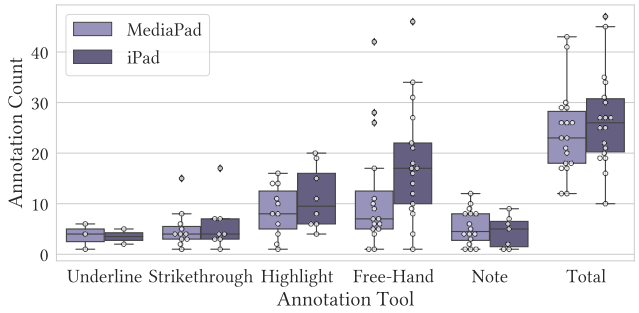

Figure 13: Number of annotations made while proof-reading on the Apple iPad and the Huawei MediaPad, split by annotation types. On the iPad, significantly more annotations were made with the free-hand tool. Data points where participants did not use a certain tool were excluded from this graph for better readability.

0.054). This goes in line with our findings from the remote study and the original lab study.

We used a Python script using the *PyPDF2* library[8] to count annotations for the digital conditions. Using a Wilcoxon signed-rank test, we found that on the iPad (mean: 17.1), participants made significantly more free-hand annotations ($Z = 2.59, p = 0.009$) than on the MediaPad (mean: 10.3) (Fig. 13).

Participants tended to use a wider variety of tools when annotating on digital media (Fig. 14). This finding also coincides with the results of the first lab study. However, the effect was smaller in this replication study. A third of participants used three different tools on paper – in contrast to none in the original study.

7 DISCUSSION

Based on our findings, we answer our research questions as follows:

RQ₁ We found a significant difference in task time between annotation media in all three of our studies. We could not find a significant difference in error recall between annotations media in any of our studies. We could only find a significant difference in perceived task load between annotation media in our remote study. If there was a significant difference, paper always outperformed digital media.

[8]https://pypi.org/project/PyPDF2/

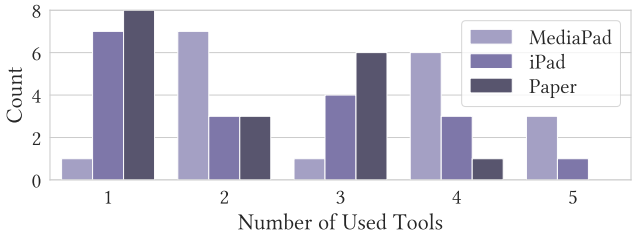

Figure 14: Number of different tools used with each medium. On average, participants used fewer different tools on paper. The effect is not as strong as in the previous study.

RQ₂ Even though total annotation count did not differ significantly between annotation media, we could observe differences in the number and types of used tools. On paper, participants used fewer different tools. Using the iPad with good stylus input, annotation style resembled paper annotations more closely than with other digital media.

RQ₃ Most of our findings from the asynchronous remote proof-reading study go in line with the results of our lab study. The only difference we found and can not yet explain is the influence of annotation media on task load, which we could not replicate in lab studies. Besides that, it seems viable to conduct proof-reading studies remotely.

In this section, describe our reasoning and discuss implications of our findings for future research and practical applications.

In all studies, we found a significant difference in task completion time depending on the annotation medium. Regardless of the other conditions, participants were faster on paper. This coincides with results of existing studies which claim that proof-reading on digital media takes longer [16, 46, 48]. On the other hand, our findings contradict the those of Köpper et al. [26] who could not find a significant difference in task time, as well as Wharton-Michael et al. [45] who observed a significantly lower error recall during a fixed task time for proof-reading on paper. One reason for faster annotating on paper could be that using a free-hand tool like a physical pen, drawing symbols to correct errors, such as striking out duplicate words, adding missing punctuation, or drawing arrows to indicate wrong capitalization, is much simpler than first selecting a tool or typing a comment. Even though a free-hand tool was available on the tablets, writing on a digital device might be slower than on paper due to missing haptic feedback and limited spatial and temporal tracking resolution [1]. Additionally, as a tablet is slightly smaller than a piece of A4 paper and people tend to write bigger on tablets [14], zooming in before annotating is common on tablets. This change in context might require users to re-orient and therefore consume time. Similarly, for annotating with the PC, some researchers argue that the higher task time [5, 44] might be a result of losing time when switching between mouse and keyboard.

Error recall was very similar among all conditions in all studies, with no significant difference between annotation media in any of the studies. This goes in line with the findings of Chan et al. [4], who argue that annotation is a means to mark an error, not to find it. According to that, finding errors should depend more on

page layout and individual skill than the used tool. However, our findings contradict other studies who found a difference in error recall in favor of paper [7, 40, 45, 46, 48].

In our remote study, annotating on paper resulted in a significantly lower task load than on the PC. However, we could not replicate those findings in our lab studies. On the other hand, in both lab studies, the task load was significantly higher for the *Huawei MediaPad M5* than on paper and the other digital device (PC or iPad). At this point, we can not certainly tell what caused this effect. One explanation could be that the alternative digital devices in our lab studies (a powerful PC with a large, high resolution screen, and a high-end tablet) were likely to be better than the devices the participants of our remote study – mostly students – had at home. Remarkably, the perceived task load for annotating on paper was very similar between all three studies. We conclude that a paper condition can serve as a good baseline for proof-reading studies in terms of task load.

We did not find a significant difference in annotation count between proof-reading on paper and on digital media in any of our studies. In all studies, we found that on paper, participants used at most three tools (usually pens and highlighter). On digital media, however, users switched between all available tools depending on what they were annotating. Those findings contradict Hastreiter et al. [20], whose participants stated in a post-study interview that they switched a lot between different pen types and colors when annotating on paper. Additionally, we found that with the iPad and its well-functioning stylus at hand, participants made significantly more free-hand annotations than on the other tablet. This observation supports Wolfe's theory that digital annotation relies on good stylus-input and should allow for handwritten annotations [47].

Even though our remote study and our lab studies were too different to compare systematically, our findings suggest that results are very similar between both settings. However, as shown by our third study, different models of digital devices can strongly influence task time and perceived task load. This might also be reflected by the significant difference in perceived task load between annotation media in our remote study, which we did not find in our lab studies. Therefore, it is necessary to control for influence by the used device and software by explicitly asking participants about the exact system they used during the study. Furthermore, we suggest to always include a Paper condition in future annotation studies, as it has proven to be a very solid baseline in terms of task completion time and perceived task load. This makes comparisons to the large body of existing research on annotating and proof-reading possible.

8 LIMITATIONS

In all three of our studies, we used a short proof-reading task to gain insights on annotation behavior and performance on different media. Even though proof-reading is also used as a proxy for annotation in other publications [16, 20, 46, 48], we are aware, that there are many differences to other types of annotating, for example active reading or summarizing text. During such tasks, features of digital media, such as context search, copying text, creating references, or using dedicated literature management software is probably a more important factor than being faster when annotating. Additionally, in our lab studies, participants were using hardware and software provided by us. Even though this mitigates confounding effects caused by different hardware and software, participants might have been used to different tools and therefore performed differently than with their own setup. Furthermore, because of the short tasks, we could not measure long term effects such as exhaustion in our studies. As our participants were mostly young adults with academic background, it is likely that most of them had some experience in annotating text. Therefore, our finding might not be generalizable to a heterogeneous population. We also did not specifically recruit participants with experience in annotating with tablets. The difference between paper and a good tablet might be even closer for such a group. In terms of perceived task load, we could not replicate findings from the remote study (significant difference between PC and paper) in our lab study. As we can only speculate about the reason with the data available, this is a subject to investigate further in the future.

9 CONCLUSION AND FUTURE WORK

In this paper, we presented the results of three studies comparing annotation tasks on different media: paper, PC, and two different tablets. We measured task time, error recall and task load, as well as qualitative and quantitative aspects of made annotations. As many related studies are quite old regarding the fast-paced development of new hardware, we contribute to the field of annotation and proof-reading research by re-evaluating past findings with studies using current hardware and software. Our results can serve as a foundation for future proof-reading and annotation studies, and help practitioners with design decisions regarding hardware and software used for annotating digital documents.

We have found that paper still outperforms digital media for annotating documents. Additionally, perceived task load was almost constant among all three of our studies when it comes to annotating on paper. Therefore, we suggest using paper as a baseline condition for future annotation studies. Even though a state-of-the-art *Apple iPad* with a well-functioning stylus performed only slightly worse in terms of task time and perceived task load, the older *Huawei MediaPad* performed far worse. This suggests that not only the device type, but also the quality of the specific device in terms of stylus tracking, responsiveness, and available software have a significant influence on task time and task load. While we only observed this as part of our proof-reading studies, it is likely that similar performance differences can also be found in other use-cases that require precise stylus input. We therefore recommend to *(a)* use state-of-the-art hardware and software if possible, *(b)* use tasks that are robust against such performance differences, or *(c)* include a baseline condition with a known performance. Furthermore, we recommend developers of annotation software and hardware to try to replicate the experience of annotating on paper as closely as possible with accurate and responsive stylus support.

Following the studies presented in this paper, one next step in annotation research should be to increase ecological validity by investigating annotation behavior in the field, for example through diary studies. Additionally, it would be interesting to see how people use annotation for their personal or professional work. A first step into this direction could be to acquire and analyze annotated physical and digital documents, for example from a university course.

Furthermore, our findings suggest that remote studies are a feasible way to investigate proof-reading performance and behavior. Therefore, a large-scale remote study with a heterogeneous sample would be an important step towards thoroughly understanding how people annotate.

REFERENCES

[1] Michelle Annett, Fraser Anderson, Walter F Bischof, and Anoop Gupta. 2014. The pen is mightier: understanding stylus behaviour while inking on tablets. In *Proceedings of Graphics Interface 2014*. 193–200.

[2] Rebecca Dawn Baker. 2010. *Comparing the Readability of Text Displays on Paper, E-Book Readers, and Small Screen Devices*. Ph. D. Dissertation. University of North Texas.

[3] James Blustein, David Rowe, Ann-Barbara Graff, James Blustein, David Rowe, and Ann-Barbara Graff. 2011. Making Sense in the Margins: A Field Study of Annotation. In *International Conference on Theory and Practice of Digital Libraries*, Heiko Schuldt (Ed.). Number 6966 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, 252–259. https://doi.org/10.1007/978-3-642-24469-8

[4] Alan H. S. Chan, Steve N. H. Tsang, and Annie W. Y. Ng. 2014. Effects of Line Length, Line Spacing, and Line Number on Proofreading Performance and Scrolling of Chinese Text. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56, 3 (May 2014), 521–534. https://doi.org/10.1177/0018720813499368

[5] Guang Chen, Wei Cheng, Ting-Wen Chang, Xiaoxia Zheng, and Ronghuai Huang. 2014. A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education* 1, 2-3 (Nov. 2014), 213–225. https://doi.org/10.1007/s40692-014-0012-z

[6] Anthony Creed, Ian Dennis, and Stephen Newstead. 1987. Proof-reading on VDUs. *Behaviour & Information Technology* 6, 1 (Jan. 1987), 3–13. https://doi.org/10.1080/01449298708901814

[7] Anthony Creed, Ian Dennis, and Stephen Newstead. 1988. Effects of display format on proof-reading with VDUs. *Behaviour & Information Technology* 7, 4 (Oct. 1988), 467–478. https://doi.org/10.1080/01449298808901890

[8] Sally Jo Cunningham, Chris Knowles, Sally Jo Cunningham, and Chris Knowles. 2005. Annotations in an Academic Digital Library: The Case of Conference Note-Taking and Annotation. In *International Conference on Asian Digital Libraries*, Pimrumpai Premsmit and Vilas Wuwongse (Eds.). Number 3815 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, 62–71.

[9] C. Desmoulins and D. Mille. 2002. Pattern-based annotations on E-books: from personal to shared didactic content. In *Proceedings. IEEE International Workshop on Wireless and Mobile Technologies in Education*. IEEE Comput. Soc, Vaxjo, Sweden, 82–85. https://doi.org/10.1109/WMTE.2002.1039224

[10] Andrew Dillon. 1992. Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics* 35, 10 (Oct. 1992), 1297–1326. https://doi.org/10.1080/00140139208967394

[11] Thomas Franke, Christiane Attig, and Daniel Wessel. 2017. Assessing Affinity for Technology Interaction – The Affinity for Technology Interaction (ATI) Scale. Scale Description – English and German Scale Version. (2017). https://doi.org/10.13140/RG.2.2.28679.50081 Publisher: Unpublished.

[12] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (April 2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150

[13] Fei Gao. 2013. A case study of using a social annotation tool to support collaboratively learning. *The Internet and Higher Education* 17 (April 2013), 76–83. https://doi.org/10.1016/j.iheduc.2012.11.002

[14] Sabrina Gerth, Annegret Klassert, Thomas Dolk, Michael Fliesser, Martin H. Fischer, Guido Nottbusch, and Julia Festman. 2016. Is Handwriting Performance Affected by the Writing Surface? Comparing Preschoolers', Second Graders', and Adults' Writing Performance on a Tablet vs. Paper. *Frontiers in Psychology* 7 (2016). https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01308

[15] John D. Gould, Lizette Alfaro, Vincent Barnes, Rich Finn, Nancy Grischkowsky, and Angela Minuto. 1987. Reading Is Slower from CRT Displays than from Paper: Attempts to Isolate a Single-Variable Explanation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 29, 3 (June 1987), 269–299. https://doi.org/10.1177/001872088702900303

[16] John D. Gould and Nancy Grischkowsky. 1984. Doing the Same Work with Hard Copy and with Cathode-Ray Tube (CRT) Computer Terminals. *The Journal of the Human Factors and Ergonomics Society* 26, 3 (June 1984), 323–337. https://doi.org/10.1177/001872088402600308

[17] Francois Guimbretiere. 2003. Paper augmented digital documents. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*. 51–60.

[18] Sandra G Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the human factors and ergonomics society 50th annual meeting* 50, 9 (2006), 904–908.

[19] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[20] Isabella Hastreiter, Manuel Burghardt, David Elsweiler, and Christian Wolff. 2013. Digitale Annotation im akademischen Kontext: Empirische Untersuchung zur Annotationspraxis von Studierenden auf Tablet-Computern. 63 (2013), 118–129.

[21] Hanho Jeong. 2012. A comparison of the influence of electronic books and paper books on reading comprehension, eye fatigue, and perception. *The Electronic Library* 30, 3 (June 2012), 390–408. https://doi.org/10.1108/02640471211241663

[22] Ricardo Kawase, Eelco Herder, Wolfgang Nejdl, Ricardo Kawase, Eelco Herder, and Wolfgang Nejdl. 2009. A Comparison of Paper-Based and Online Annotations in the Workplace. In *European Conference on Technology Enhanced Learning*. Springer Berlin Heidelberg, Berlin, Heidelberg, 240–253. https://doi.org/10.1007/978-3-642-04636-0_23 Series Title: Lecture Notes in Computer Science.

[23] Kibum Kim, Scott Turner, and Manuel A. Pérez-Quiñones. 2004. Comparing Classroom Note Taking across Multiplatform Devices. (2004). https://doi.org/10.1145/3044715

[24] Kristof Korwisi. 2014. Papier- und computergestützte Annotationsverfahren im Vergleich. (2014).

[25] Akrivi Krouska, Christos Troussas, and Maria Virvou. 2018. Social Annotation Tools in Digital Learning: A Literature Review. In *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, Zakynthos, Greece, 1–4. https://doi.org/10.1109/IISA.2018.8633609

[26] Maja Köpper, Susanne Mayr, and Axel Buchner. 2016. Reading from computer screen versus reading from paper: does it still make a difference? *Ergonomics* 59, 5 (2016), 615–632. Publisher: Taylor & Francis.

[27] Wolfgang Lenhard, Ulrich Schroeders, and Alexandra Lenhard. 2017. Equivalence of Screen Versus Print Reading Comprehension Depends on Task Complexity and Proficiency. *Discourse Processes* 54, 5-6 (July 2017), 427–445. https://doi.org/10.1080/0163853X.2017.1319653

[28] Catherine C. Marshall. 1997. Annotation: from paper books to the digital library. In *Proceedings of the second ACM international conference on Digital libraries - DL '97*. ACM Press, Philadelphia, Pennsylvania, United States, 131–140. https://doi.org/10.1145/263690.263806

[29] Catherine C Marshall. 1998. The future of annotation in a digital (paper) world. *Successes & Failures of Digital Libraries:[papers presented at the 1998 Clinic on Library Applications of Data Processing, March 22-24, 1998]* (1998).

[30] Catherine C. Marshall. 1998. Toward an ecology of hypertext annotation. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems links, objects, time and space—structure in hypermedia systems - HYPERTEXT '98*. ACM Press, Pittsburgh, Pennsylvania, United States, 40–49. https://doi.org/10.1145/276627.276632

[31] Catherine C. Marshall and A. J. Bernheim Brush. 2004. Exploring the relationship between personal and public annotations. In *Proceedings of the 2004 joint ACM/IEEE conference on Digital libraries - JCDL '04*. ACM Press, Tuscon, AZ, USA, 349. https://doi.org/10.1145/996350.996432

[32] M. Menozzi, F. Lang, U. Naepflin, C. Zeller, and H. Krueger. 2001. CRT versus LCD: Effects of refresh rate, display technology and background luminance in visual performance. *Displays* 22, 3 (2001), 79–85. Publisher: Elsevier.

[33] Paul Muter, Susane A. Latrémouille, William C. Treurniet, and Paul Beam. 1982. Extended Reading of Continuous Text on Television Screens. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 24, 5 (Oct. 1982), 501–508. https://doi.org/10.1177/001872088202400501

[34] Kenton O'Hara and Abigail Sellen. 1997. A comparison of reading paper and on-line documents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. ACM, Atlanta Georgia USA, 335–342. https://doi.org/10.1145/258549.258787

[35] Ilia A. Ovsiannikov, Michael A. Arbib, and Thomas H. Mcneill. 1999. Annotation technology. *International Journal of Human-Computer Studies* 50, 4 (April 1999), 329–362. https://doi.org/10.1006/ijhc.1999.0247

[36] Jennifer Pearson, George Buchanan, and Harold Thimbleby. 2009. Improving Annotations in Digital Documents. In *Research and Advanced Technology for Digital Libraries*, Maristella Agosti, José Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas (Eds.). Vol. 5714. Springer Berlin Heidelberg, Berlin, Heidelberg, 429–432. https://doi.org/10.1007/978-3-642-04346-8_51

[37] Bill N. Schilit, Gene Golovchinsky, and Morgan N. Price. 1998. Beyond Paper: Supporting Active Reading with Free Form Digital Ink Annotations. ACM Press, 249–256.

[38] Jordan T Schugar, Heather Schugar, and Christian Penny. 2011. A Nook or a Book? Comparing College Students' Reading Comprehension Levels, Critical Reading, and Study Skills. *International Journal of Technology in Tea ching and Learning* 7, 2 (2011).

[39] Abigail Sellen and Richard Harper. 1997. Paper as an analytic resource for the design of new technologies. In *Proceedings of the ACM SIGCHI Conference*

on Human factors in computing systems. ACM, Atlanta Georgia USA, 319–326. https://doi.org/10.1145/258549.258780

[40] Hirohito Shibata and Kentaro Takano. 2014. Reading from paper versus reading from a touch-based tablet device in proofreading. In *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, London, United Kingdom, 433–434. https://doi.org/10.1109/JCDL.2014.6970211

[41] Eva Siegenthaler, Pascal Wurtz, Per Bergamin, and Rudolf Groner. 2011. Comparing reading processes on e-ink displays and print. *Displays* 32, 5 (Dec. 2011), 268–273. https://doi.org/10.1016/j.displa.2011.05.005

[42] Lauren M. Singer and Patricia A. Alexander. 2017. Reading on Paper and Digitally: What the Past Decades of Empirical Research Reveal. *Review of Educational Research* 87, 6 (Dec. 2017), 1007–1041. https://doi.org/10.3102/0034654317722961

[43] Jurgen Steimle, Iryna Gurevych, and Max Muhlhauser. 2007. Notetaking in University Courses and its Implications for eLearning Systems. *DeLFI 2007: 5. e-Learning FachtagungInformatik der Gesellschaft für Informatik eV (GI)* (2007).

[44] Hildegunn Støle, Anne Mangen, and Knut Schwippert. 2020. Assessing children's reading comprehension on paper and screen: A mode-effect study. *Computers & Education* 151 (July 2020), 103861. https://doi.org/10.1016/j.compedu.2020.103861

[45] Patty Wharton-Michael. 2008. Print vs. Computer Screen: Effects of Medium on Proofreading Accuracy. *Journalism & Mass Communication Educator* 63, 1 (March 2008), 28–41. https://doi.org/10.1177/107769580806300103

[46] R. T. Wilkinson* and Helen M. Robinshaw. 1987. Proof-reading: VDU and paper text compared for speed, accuracy and fatigue. *Behaviour & Information Technology* 6, 2 (April 1987), 125–133. https://doi.org/10.1080/01449298708901822

[47] Joanna Wolfe. 2002. Annotation technologies: A software and research review. *Computers and Composition* 19, 4 (2002), 471–497. Publisher: Elsevier.

[48] P. Wright and A. Lickorish. 1983. Proof-reading texts on screen and paper. *Behaviour & Information Technology* 2, 3 (July 1983), 227–235. https://doi.org/10.1080/01449298308914479

[49] P. Wright and A. Lickorish. 1984. Ease of annotation in proof-reading tasks. *Behaviour & Information Technology* 3, 3 (July 1984), 185–194. https://doi.org/10.1080/01449298408901750