Social Sciences & Humanities Open 4 (2021) 100226

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

Interacting with academic readings — A comparison of paper and laptop

Nanna Inie *, Louise Barkhuus, Claus Brabrand

The IT University of Copenhagen, Center for Computing Education Research, Rued Langgaards Vej 7, 2300, København S, Denmark

ARTICLE INFO

Keywords:
Digital reading
Text annotation
Academic reading
Active reading support

ABSTRACT

Academic reading, a form of active reading, often involves interaction with the text. Highlighting and annotating are some of the most common strategies of interacting with academic texts, yet we have limited understanding of exactly how such interactions affect reading comprehension in digital versus analog reading environments. In this paper, we present an exploratory study, comparing how university students (n = 50) interact with a digital and a physical text, focusing on highlights and annotations. We compare reading time, interaction with, and subsequent memory of the texts. We make nine observations about reading time, highlighting and annotation frequency and memory scores between paper and laptop. We find that students annotate significantly less on laptop than on paper, but that neither highlighting nor annotations influence subsequent memory of the text in either condition. Our broader contribution is to augment understanding of how different interaction features aid academic reading in a natural environment.

1. Introduction

University students read increasing amounts of course literature in digital formats. Although digital academic texts have not replaced physical textbooks and printed sheets of paper (Abuloum, Farah, Kaskaloglu, & Yaakub, 2019; Mizrachi, Salaz, Kurbanoglu, Boustany, & ARFIS Research Group, 2018; Pálsdóttir, 2019; Vincent, 2016; Zhang & Niu, 2016, pp. 207–222), especially for longer texts, most university students currently read in a combination of digital and analog media (Abuloum et al., 2019; Mizrachi et al., 2018; Pálsdóttir, 2019). Students' choice of medium is especially influenced by factors such as material cost, availability of texts, type of reading (e.g. selective or extended reading), and possibilities of interaction with the text, such as highlighting, annotation and note-taking (Pálsdóttir, 2019). In order to design appropriate digital reading technologies, we need grounded knowledge about how digital reading technologies are used, and how their interaction design compares to reading on paper.

The differences in reading performance and learning outcomes between reading digital and physical texts have been studied extensively, particularly with a focus on reading speed and comprehension of text, however, producing divergent conclusions in terms of whether there is evidence for a consistent performance deficit when reading digital texts, e.g. (Dillon, McKnight, & Richardson, 1988; Kara, Augustine, Shand, Bakner, & Rayne, 2019; Mangen, Walgermo, & Brønnick, 2013; Rockinson-Szapkiw, Courduff, Carter, & Bennett, 2013; Singer & Alexander, 2017a, 2017b). The detailed ways in which students interact with different reading support tools and milieus have been investigated to a lesser degree, yet with noteworthy exceptions (Brady, Cho, Narasimham, Fisher, & Goodwin, 2018; Freund, Kopak, & O'Brien, 2016; Johnston & Ferguson, 2020; Kol & Schcolnik, 2000; Wolfe, 2008). Research in the area is still attempting to identify robust moderating factors for why studies of digital versus analog performance seem to yield conflicting results (Delgado, Vargas, Ackerman, & Salmerón, 2018). A 2018 meta-analysis of reading comprehension studies by Kong and colleagues observed that the magnitude in differences in performance between paper and screen follows a diminishing trajectory in newer research (defined as post-2013 studies), however at p = .10, declared that the trajectory was not statistically significant. Another meta-study by Delgado and colleagues, also from 2018, concluded the opposite: that digital reading inferiority *increases* over the years (Delgado et al., 2018). The survey studies were based on different samples of previous research, which may explain this discrepancy.

Delgado and colleagues suggest *the shallowing hypothesis* (Annisette & Lafreniere, 2017) as a possible explanation of why screen-based reading may impede comprehension; because the use of most digital media consists of quick interactions driven by immediate rewards (e.g. number of "likes" of a post), readers using digital devices may find it difficult to engage in sustained, challenging tasks, such as deep reading and comprehension. A main challenge in the field of reading research thus appears to be to identify which specific factors consistently influence

* Corresponding author.
E-mail addresses: nans@itu.dk, nans@itu.dk (N. Inie), barkhuus@itu.dk (L. Barkhuus), brabrand@itu.dk (C. Brabrand).

https://doi.org/10.1016/j.ssaho.2021.100226
Received 3 June 2021; Received in revised form 22 October 2021; Accepted 30 October 2021
Available online 13 November 2021

learning outcomes and preferences, and how we may improve these factors in the design of digital tools for reading (Delgado et al., 2018; Pearson, Buchanan, & Thimbleby, 2013).

One promising avenue is further studies of the *interaction features* of digital tools, in other words, whether digital tools for reading afford and support the interaction necessary for students to engage in *active reading* (Adler & Van Doren, 2014; Pálsdóttir, 2019; Pearson et al., 2013). Active reading of a text is often accompanied by activities such as highlighting and annotating the text (O'Hara, 1996). Previous studies have indicated that software developed specifically to mimic the lightweight interactions of paper reading have the potential to reduce the divide between digital and analog reading in both performance and experience measures (Pearson et al., 2013). This paper demonstrates a detailed investigation into academic interactive reading and examines in further detail how different interaction design features can influence active reading comprehension.

The paper presents a preliminary study with 50 university students, where we compare reading an academic text on paper and on a laptop (which is the most commonly used hardware for reading among university students (Mizrachi et al., 2018)). Although designed as a controlled experiment, we took several measures to prioritize ecological and external validity, which we describe in further detail in the *Experiment* section (3). We measured reading time, interaction with a scholarly text, and subsequent memory of the text in both media.

The novelty of the research falls in three categories: 1. Few, if any, previous studies have focused specifically on interaction features and their relation to reading time and comprehension for individual students. We investigate the relation between different types of interaction with a text on participants' memory of the text, and we study the effects *within-subjects*. 2. Few previous studies specify reading, and characteristics of reading in relation to the specific tool or medium being used to read in the study (Singer & Alexander, 2017a, 2017b). This paper presents a detailed description of the software used in the experiment and its interaction design, hoping to make our findings relevant to designers and developers of digital reading software, and 3. Where previous research has largely consisted of either rigid, controlled experiments or small-scale investigative studies, our experiment is meant to mimic students' regular study setups as thoroughly as possible.

Interaction with digital texts is particularly interesting to explore, because out of the factors previously identified as central for reading experience and reading preferences when choosing between analog and digital formats (for instance; ability to concentrate, ability to remember what was read, convenience and expenses, and technological limitations/possibilities (Pálsdóttir, 2019)), *interaction design* (both possibilities and affordances of the technological environment) is the main factor that researchers in reading support software have a an essential opportunity to improve (Pearson et al., 2013).

2. Background and related work

Research in the area of digital versus analog reading is so wide-reaching that it would be impossible to include a comprehensive overview over all the studies that have investigated the area. We therefore focus on summarizing the research which illuminates the relevance of the three focal areas of this paper: reading speed on laptop versus paper, reading comprehension on laptop versus paper, and interaction with text on laptop versus paper.

Physical books and printed paper still seems to still be primary, preferred media for academic reading, while the most commonly used hardware for students to read digital texts on is the laptop (Mizrachi et al., 2018). Therefore, these formats have been logical departures for numerous comparative studies, and a lot of academic work has been dedicated to developing software with the aim of making it as *lightweight* as physical reading, e.g. (Adler, Gujar, Harrison, O'hara, & Sellen, 1998; Chen, Guimbretiere, & Sellen, 2012; O'Hara, 1996; Pearson et al., 2012, 2013; Schilit, Golovchinsky, & Price, 1998; Tashman & Edwards, 2011), however few of these systems have had broad impact in observed empirical reading practices.

Newer research in students' preferences emphasizes that one of the major factors in their choice of medium is the flexibility to switch between tools: "*It depends on the condition, the material and my mood*" (Icelandic university student, 2016) (Pálsdóttir, 2019). According to the study from which this quote stems, by Palsdottir (Pálsdóttir, 2019), students are, generally, not negative towards electronic reading material, provided that appropriate technological functionality is available, and that the reading material has been designed in a way that satisfies their needs. Other primary factors mentioned as determinants of choice of medium are: ability to concentrate, ability to remember what was read, convenience and expenses, and technological limitations/possibilities (Pálsdóttir, 2019).

Many researchers believe that embedding physical book metaphors such as indexes, tables of content, bookmarks etc. enhances digital versions of documents, while other researchers maintain that future designs should, in fact, avoid implementing the book metaphor in digital tools due to the limits it enforces upon conceptual models of search and non-linear navigation (Pearson et al., 2013). Overall, there are no conclusive, obvious focal areas for interaction design of digital tools for reading, because we have not yet cracked the code to identify robust, moderating factors for performance and preferences between physical and digital reading (Delgado et al., 2018).

2.1. Reading speed

Older studies (before 1990) consistently showed that reading on a screen was significantly slower than reading on paper. In fact, such studies often produced a performance deficit of between 20% and 30% when reading on screen (Dillon et al., 1988). These studies were conducted on old hardware with flickering screens with poor resolution, and performance deficits of this amplitude have not been replicable as screen technology vastly improved over the years. Reading speed now seems to be roughly similar on paper and on screen (Kong, Seo, & Zhai, 2018). One recent study looked into reading speed of different digital formats, investigating PDFs versus EPUB readings across different media - laptop, tablet, smart phone, and e-reader (ZengXue et al., 2016). Even though the study was based on a very small participant number ($n = 15$), results pointed to EPUB being the faster medium for all hardware forms except the laptop, where participants read faster in a PDF.

Several studies have shown that screen-based reading *behavior* is different than paper-based reading. In general, screen-reading is characterized as involving more time spent browsing and scanning of the document, as well as non-linear and one-time reading. Consecutively, less time is spent on in-depth and concentrated reading (Baron, 2015; Delgado & Salmerón, 2021; Hillesund, 2010; Liu, 2005; Loh & Kanai, 2016; MangenDon, 2014; Singer Trakhman et al., 2018, 2019). Newer research has, however, shown little to no difference in reading speed on computer screens in comparison to reading on paper. While reading at a specific pace is not a common performance metric for good academic reading, the time spent on a text does seem to have some correlation with the level of engagement with the text. Sage et al. (Kara et al., 2019) showed that longer reading time (on paper, desktop computer or tablet) was marginally or significantly related to better comprehension, higher satisfaction, higher perceived control, lower perceived difficulty, and higher confidence in a quiz following the reading. This may not be surprising — the students who spent longer on the text also memorized it more and felt more confident. This study was based on a between-subjects measurement, so we can not know if the relations are similar across all three media.

2.2. Reading comprehension

Reading comprehension is one of the most common performance metrics in comparison studies, although *comprehension* is not necessarily

defined or measured equally across studies. Reading comprehension is highly relevant to university students who read for academic purposes, and memory is decidedly relevant for comprehension. The *Remember-Know* learning paradigm coined by Tulving (Tulving, 1985) describes the relation of memory to learning: Knowledge which is *Remembered* is typically recollected in close association with related information pertaining to the learning episode. It is more vulnerable to fading with time. Knowledge which is *Known* is recalled, retrieved, and applied without any such additional contextual associations. By implication, it is assumed that *Known* knowledge is indicative of better learning (Conway, Gardiner, Perfect, Anderson, & Cohen, 1997). Studies of a comparative nature, such as the present study and its predecessors, are usually short-term based and can better assess if the knowledge is *Remembered*, rather than *Known* on a longer term. Different variations of memory tests performed shortly after the reading are also the most common estimate of reader comprehension in such studies.

Comparison studies of reading comprehension between paper-based and digital reading have yielded mixed results. In 2008, Noyes and Garland claimed that total equivalence of paper and screen comprehension was not possible to achieve, but that more sophisticated comparative measures and more positive user attitudes had resulted in a move towards that goal (Noyes & J Garland, 2008). Some newer studies found that students comprehended or learned equally well reading on paper versus digitally (Kara et al., 2019; Margolin, Driscoll, Toland, & Kegler, 2013; Subrahmanyam et al., 2013). One recent meta-analysis has indicated that the magnitude of the difference in reading comprehension between paper and screen follows a diminishing trajectory (Kong et al., 2018) (although at $p = .10$, this tendency was not statistically significant), while a different meta-analysis showed the opposite: that the advantage of paper-based reading increased over the years (Delgado et al., 2018).

Generally, there seems to be an argument for paper-based reading supporting better reading comprehension to at least *some* degree, especially when comprehension is measured by means of more specific questions, rather than questions about the 'main idea' of the text (Singer & Alexander, 2017a, 2017b; Singer Trakhman et al., 2018, 2019). Some studies have suggested that digital media impair metacognitive regulation and thus learning (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014), while other studies comparing paper, PC, iPad, and Kindle, have not been able to reproduce this claim and found no difference in metacognitive accuracy for text reading across study media (Norman & Furnes, 2016).

Reading comprehension is a valuable indicator of performance in academic reading studies. It would therefore be a significant achievement to identify the moderating factors of the digital interface which have a significant impact on comprehension. Because digital devices are, unavoidably, increasingly used in our educational systems, it is pertinent that we understand how to design, deploy and use them in a way that preferably fosters and, as a minimum, does not hinder reading comprehension and learning (Delgado et al., 2018).

2.3. Interaction: highlights, annotations, and notes

Contrary to reading speed and comprehension, *interaction* behaviour patterns have been shown to vary consistently when comparing digital to physical text reading, with higher levels of interactivity for physical formats. Similarly, interaction possibilities have been repeatedly mentioned as a determining factor for why students prefer physical over digital texts (Kara et al., 2019; Pálsdóttir, 2019). Newer PDF reading software allows interactive features that are very similar to the traditional physical interaction features (e.g. highlighting, adding comments and notes), so what is wrong with the interaction design of digital software that students claim not to use its features?

Annotations are made up of several elements and can take many different forms (Jennifer Sarah Pearson, 2012). Highlights, underlining, notes, and doodles in the text are all categories of annotation. Academic students most often create annotations for themselves, and annotating serves purposes such as memory aiding, bookmarking, signaling attention, working through problems, and interpretations of text (Pearson et al., 2012, 2013). Highlights and annotations known as *high-value annotations* have experimentally been used to reverse-identify important keywords and passages in information-heavy books, and they seem to be a reliable indicator of salient bits and passages (Frank, Price, Marshall, & Golovchinsky, 2003), but whether annotations help the reader *learn*, seems to be highly dependent on the type of annotations.

We adopt the following definitions by Marshall (Marshall, 1997) and Pearson and colleagues (Pearson et al., 2013) for distinguish between common forms of interaction with text while reading:

Highlights or underlining, which are made with a highlighter pen that colors on top of the text, or with a pen or pencil, by drawing a line underneath the text.

Annotations, which are doodles, arrows, markings of interest and scribbles, often made in the margin of the text (marginalia).

Notes, which are usually written in separate documents of the read text, and make sense in and of themselves without their immediate connection to the reading.[1]

How we measure highlights and annotations in our study is explained in further detail in the *Data analysis* section (section 3.6).

2.3.1. Highlights and underlining

In 2013, Dunlosky and colleagues (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013a, 2013b) reviewed an extensive amount of scientific papers about various study techniques, and concluded that highlights and underlining are ineffective, time-consuming, and may, at worst, cause adverse effects to learning: *"It may be that underlining draws attention to individual items rather than to connections across items"* (Dunlosky et al., 2013a, 2013b). Much more effective is post-engagement with the text, such as self-quizzing, distributed studying, and self-explanation (Dunlosky et al., 2013a, 2013b). Interestingly, even though this research is almost a decade old, it seems that students continue to use highlighting actively – at least in print formats. In studies that looked at reading and text interaction during controlled experiments, most conclude that annotation and highlighting behavior has not migrated in a useful way from physical to digital environments (Ben-Yehudah Eshet-Alkalai, 2018; Jennifer Sarah Pearson, 2012; Liu, 2005). For instance, Johnston and Ferguson observed that study participants had "some awareness" of digital features, but that they did not use them at all when reading in an e-book (Johnston & Ferguson, 2020). Sage and colleagues let the participants in their study interact with the digital platforms in "whatever way seemed natural", but also observed that most students did not take advantage of the digital features (Kara et al., 2019). In a 2016 survey study with nearly 10.000 student participants spread over 19 countries, over 80% of students reported that they highlight and annotate in print, and barely a third of students claimed to do so in electronic documents (Mizrachi et al., 2018). This survey is built on self-report, and it is not clear whether the authors provided the respondents with a definition of annotation, e.g. the difference between underlining and taking notes.

2.3.2. Annotations and notes

According to Adler (Adler & Van Doren, 2014), reading can be classified as elementary, inspectional, analytical, or syntopical, depending on how thorough an overview over or engagement with the

[1] In this paper, for clarity, we distinguish between annotations as: notes taken in or on the document that is being read, and notes as: notes made in a different medium from the text, such as a notebook, a notepad, or any note-taking software such as OneNote, Evernote, Notion etc. We do this to avoid confusing the properties of reading support software with those of note-taking support software, as they are usually different media on both paper (a book versus a notebook) and digitally (Adobe Reader versus OneNote).

text, the reader has or acquires. It is likely that a university student will at least strive to read literature on the analytical and syntopical levels, and these types of reading are usually highly facilitated by and dependent on the creation of annotations, both in-text and externally (Adler & Van Doren, 2014; Pearson et al., 2013).

Note-taking practices will undoubtedly vary depending on numerous factors, e.g. the format of what is being read, the type of content, ownership of the book or material, and location in which the student is reading. Kara and colleagues (Kara et al., 2019) found that out of 120 American university students, 59% reported to take notes on separate paper, 23% take notes on the reading itself, 7% take both handwritten and typed notes, 5% only type their notes, and 6% take no notes. These numbers were based on questions about how the students take notes *outside* of the classroom, i.e. when reading or preparing for class. The study does not clarify whether this note-taking behavior was based on reading digital or print media. According to another survey of several hundred American students, students explain that they take notes to enhance encoding and external storage, and they report high confidence in their note-taking abilities (Witherby & Tauber, 2019).

A meta-study from 2005 of 57 other studies in note-taking by Kobayashi (Kobayashi, 2005) found a "positive but modest (mean weighted ES = 0.22, and mean unweighted ES = 0.29)" encoding effect of taking notes. The review also found that moderating factors of this effect seem to be mechanical demands of note-taking, type of learning outcome measure, and publication characteristics, rather than students' spontaneous note-taking procedures. Newer research suggests that students still rely heavily on taking notes, and that they are flexible in their note-taking behavior (e.g. choosing between taking notes on paper or on a laptop, but that students often do not know how or when to take notes most efficiently (Morehead, Dunlosky, Rawson, Blasiman, & Hollis, 2019).

Most studies of note-taking practices are based on self-report, rather than studies of students' actual notes and note-taking practices. In the current study, we therefore found it relevant to study the relationship between annotations, reading time, and comprehension, in the interest of shedding more light on whether digital note-taking influence reading differently from paper-based note-taking.

2.3.3. Developing reading software to support interaction

Several researchers in Human-Computer Interaction (HCI) have suggested that annotation on a computer is an activity that competes with the reading itself, due to the lack of direct manipulation (Kawase, Herder, & Nejdl, 2009). The cognitive workload required to interact with digital devices appears to be higher than that of interaction with paper, and the creation of annotations introduces another cognitive workload (Inie & Barkhuus, 2021; Pearson et al., 2013). Therefore, it is essential that digital tools are designed to require as little cognitive effort as possible, leaving more processing power for the primary reading task (Pearson et al., 2012, 2013). Indeed, studies have indicated that the simplest textual and interactive environment is associated with the highest comprehension outcomes and better user experience (Buchanan & Pearson, 2008; Freund et al., 2016; Pearson et al., 2012).

A significant challenge for the development of good software to support annotation is that few existing studies of digital reading specify the reading environment (or software) provided to participants, and even fewer describe its interaction design. This means that it is difficult to identify the exact interaction features which invariably influence the reading experience and performance, and thus to design novel digital tools for academic reading. The main objective of our study is to explore the differences between active reading in a digital and paper-based medium, specifically with a focus on how individual academic readers interact with a specific text in the two different formats, and whether different interaction patterns have a relation to their reading time and reading comprehension.

In this study, we have focused on the interactions *highlights* and *annotations*, as have previously been identified as crucial interactions for active reading (Marshall, 1997; Pearson et al., 2013), and as they are the interaction modalities most consistently featured in existing digital reading software.

We investigate the following six research questions:

- **RQ1: How does academic reading on a laptop compare to physical paper in terms of *reading time*?**
- **RQ2: How does academic reading on a laptop compare to physical paper in terms of *interaction*[2]?**
- **RQ3: How does academic reading on a laptop compare to physical paper in terms of *memory*?**
- **RQ4: What is the relationship between *reading time* and *memory* on paper vs. laptop?**
- **RQ5: What is the relationship between *interaction* and *reading time* on paper vs. laptop?**
- **RQ6: What is the relationship between *interaction* and *memory* on paper vs. laptop?**

To investigate these questions, we designed and carried out a controlled experiment, where we could observe students' reading time, perform a memory test, as well as measure textual interactions; in particular, highlighting and annotations. We designed the experiment as a within-group study, both as a way of reducing error from natural differences and preferences between subjects, but also to reach more comparable results. In practice, this meant that every participant read under both treatments, i.e. both on paper and on laptop.

3. Materials and methods

3.1. Reading software: Lix

Lix is a PDF reading software which provides basic interaction functionalities similar to most digital reading software: highlights in different colors, annotations and notes anchored to a specific point in the text (see Fig. 1). The software allows for both full-page and zoomed in presentation of the text.

3.2. Treatments

In the **paper reading** treatment, students were provided with the text on printed A4 paper, two highlighter pens, one pencil, one ball pen, and sticky notes. The participants were not instructed to use any of the items specifically, but rather told that the items were available for them to use as they pleased. The text was set in a 12 pt Times New Roman with headlines in 14 pt Arial. The text had 2 cm margins on either side, to allow for annotations directly on the paper sheet. In the **digital reading** treatment, students were provided with the text on a newer laptop (from 2015) with a 15″ high resolution screen. The text was a PDF file and formatted exactly as the paper reading for comparability. The text was provided in the software Lix, as described in the section above. Digital tools can be implemented in numerous ways and often do not translate directly to their paper equivalent, but we selected a software which offered as lightweight interaction possibilities as possible (Pearson et al., 2013). To avoid distractions and the use of unintended software during the reading, the laptops were not connected to the internet.

3.3. Participants

The experiment was performed with n = 50 students at [European university, anonymized] during the fall of 2019. The students signed up by responding to in-class presentations about the study as well as flyers distributed around campus. The students were from four different study

[2] Where interaction is understood as number of highlights and annotations created during the reading.

Fig. 1. Screenshots of the Lix reading environment. Highlights (left): The reader can drag the mouse cursor over the text while holding down the mouse button, and the text is highlighted. When letting go of the button, a box appears which asks the reader to select a color for the highlight (if no color is chosen, a default color or last color used is applied). In this box, the reader can also choose to make an annotation. Annotations 1 (center): The left window of the screen can be toggled to show a list of the reader's highlights. The reader can connect a comment specifically to each highlight. Annotations 2 (right): In the top-right corner of the interface, the reader can toggle a right-side panel that allows the reader to write simple notes while looking at the text. These notes are not fixed to any point in the text. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

programs, all related to information technology. We therefore expected all students to be familiar with both technology and digital reading. The students demographics were: 27 male and 23 female. 22 students were first year-students, 18 students were second year-students, and 10 students were third year-students or older. The median age was 21 years old; 14 students were between 18 and 20 years old, 16 students were 21–23, and 12 students were 24 or older. Eight students did not wish to disclose their age. All participants gave informed consent before the beginning of the study. As compensation for their participation, all participants were gifted a free textbook (digital or physical) of their own choosing.[3]

3.4. Materials and measures

3.4.1. Texts

All students read a text of approximately ten pages. Ten pages is a typical length of a self-contained reading (such as a research article or a book chapter) from a university curriculum. The texts were obtained by writing to each of the course leaders for the students' courses, asking for 3–5 examples of "A text from the course curriculum of approximately 10 pages, which you believe corresponds to the course level 2–3 months in the future". We divided each text in two halves of equal length (but so each half made sense in itself, i.e. ended and began with a paragraph break), and each student read half the text in digital, and the other half of the text in physical form, so as to achieve the most comparable results between the two media possible.

Rather than giving all participants the same text to read, we selected a text from the individual student's curriculum, a reading that was planned two to three months in their future. This was a total of nine different texts, since the experiment had participants from nine different programs or semesters. Our reasoning was to maximize the students' internal motivation for, not only reading the text, but also *learning* from the text. We believe that the results have greater external and ecological validity when each student reads a text corresponding to their skill level and program, due, primarily, to two factors: internal motivation, and skill/specialization:

Internal motivation: Previous research has shown that internal motivation significantly impacts reading compre-hension performance, (Logan, Medford, & Hughes, 2011; Schutte & Malouff, 2007), and we assumed that students would have greater intrinsic motivation for reading the texts if they knew that they were basically doing themselves a favor by reading ahead of time in their semester (rather than a random text read only for the purpose of this experiment). During the evaluation of the experiment, several of the participants mentioned that they appreciated how the experiment was basically a structured opportunity for them to study, which we interpret as a confirmation that the internal motivation for learning from the text was high.

Skill/specialization: A student of design will likely not focus on the same elements of an academic text as a student from data science. Similarly, a student of their first year might read an academic text differently than a student from their fourth year. We believed the best strategy for obtaining useful insights for research in reading support software was investigating how students approached texts from their *normal* curriculum, rather than a text that might be familiar to s subgroup of participants and less familiar to the rest.

We took several measures to ensure the comparability of the results for students reading different texts. These measures are described in Table 1.

3.4.2. Reading speed/time

Because the students were asked to self-assess when they had finished studying their text (we wanted to avoid any time pressure), we

[3] As a side Observation; the vast majority of the participants chose a physical textbook when their book was available in both formats.

Table 1
Characterization of texts.

Subject matter	Previous knowledge of subject generally means better performance on memory and summary tasks (Recht & Leslie, 1988). All texts fell within computer science and closely related fields, and individual texts were relevant to individual participants' field of study.
Type and genre	All texts were from academic curricula, as "readers differentially allocate their processing resources according to their expectations about the genre of a text." (Zwaan, 1994). All texts were selected from the students' core course of their current semester.
Readability	Lix readability score (Björnsson, 1968): (scale is from 0 to 56+, where higher means lower readability): 46–56, one outlier of 39, mean 49.3, median 49), corresponding to the level "difficult". Flesch-Kincaid readability score (Kincaid, Fishburne, Rogers, & Chissom, 1975): Flesch reading ease (scale is from 0 to 100, where lower is lower readability): All texts were between 30 and 50 (two outliers of 28 and 57, respectively) (mean 41.6, median 43.4), corresponding to a Flesch-Kincaid grade level of college.
Length	All texts were between 4800 and 6900 words (9.6–13.8 pages), where the texts with lower readability were also the shortest. To mitigate any effects of the varying lengths, we looked at reading time as words-per-minute, rather than total reading time.
Layout	All texts were reproduced in pdfs using the same fonts and size, and same margins.

are reporting the reading speed as reading *time* in this paper. Participants were free to study the text several times if they wanted, or to re-read segments. We believe that reading time is a more relevant metric than reading speed to assess the time needed for a student understand and memorize a text.

3.4.3. Memory tests

We created a memory test for each individual text; nine different memory texts in total. Each test had 16 questions; eight questions relating to the first half of the text, and eight questions relating to the second half of the individual text. The sequence of questions was randomly permuted and it was not revealed which part of the text the question referred to. The questions were generally literal, and inferred about the most important points of the text (as defined by the authors collaboratively). The questions were of a *prompted recall* type, because we wanted the tests to reflect the participants' comprehension and memory of the text as accurately as possible (rather than their general knowledge of the subject of the text). A characterization of the memory tests is provided in Table 2. For all students, the memory test was presented on physical A4 paper and filled in with pen or pencil. Correct answers for the memory tests had been defined pre-experiment, and points were awarded when responses corresponded with the pre-defined answers. All tests were manually scored.

3.4.4. Measuring highlights and annotations

Because the experiment took place in a "lab setting" (albeit in familiar rooms on campus), students did not have access to their personal note taking tools or environments. We therefore look at *highlights* and *annotations* created in or on the read document (as opposed to notes in notebooks or specific note-taking software (as per the distinction presented in Section 2.3).

Table 2
Details of memory test design.

Structure	Memory tests followed the same "template", meaning they had the same type of questions in the same order. Each test had 10 "Fill in the blank(s)" for question 1, 3, 6–9, 12, 13 and 16 and 6 "According to the text, is this correct?" for question 2, 4, 10, 11, 14, and 15.
Wording	All tests had similar wording, modelled after the OECD Programme for International Student Assessment (PISA) tests (OECD).
Question design	All texts were read by researchers, who assessed the most salient concepts of the texts. These concepts were the frames of the memory test questions.

Highlighting and underlining. Highlighting is defined as the drawing over text with a colored pen that allows the complete visibility of the text underneath. Highlighting requires a different tool from other annotations, both in physical and digital reading environments - the highlighter pen (and most readers do not also make notes with their highlighter pen). Underlining (the drawing of a line with a pen or pencil underneath the text) serves the same purpose as highlighting, i.e. marking important words or sentences in the text. In the analysis, we have counted each highlighted or underlined segment which makes sense "in and of itself" as *one highlight*. That means, that if just one word is highlighted, it was counted as one highlight, and if multiple consecutive words in a sentence are highlighted, they were also counted as one highlight. If one or more words in a sentence are skipped, the marking was counted as one highlight if the highlighted segment makes sense independently. See examples of this in Fig. 2. This means that some subjective assessment has taken place in agreement between the authors in categorizing which segments are self-sufficient.

Some texts contain mathematical formulae. For these, we counted each 'element' of a formula as one word; e.g., the text $2 + 2 = 4$ would be counted as three words, if the whole formula was highlighted. We did not distinguish between the use of different colors of highlighters. Several highlighting colors were supplied, both for the paper and laptop condition. In some cases, different colors were used, but a detailed analysis and discussion of highlighting strategies is beyond the scope of this paper.

Annotations. Annotations are scribbles, notes, and doodles on the read document itself, sometimes connected to a specific place in the source text with arrows, brackets, proximity, or another mark. If written in the margins of the document, these can also be called *marginalia*. In our analysis, each self-sufficient annotation-segment was counted as one annotation, i.e. one comment or one circle around a text segment. Annotations are often spaced out from each other on a page, which allows for easier distinction. In the paper condition, if an underlined segment is accompanied by an annotation, the underline was counted as part of the annotation rather than as a highlight, as the underlining serves the purpose of marking a specific point in the text. See examples of annotation counts in Fig. 2, left. In the laptop condition, it was possible to distinguish annotations from each other both by formatting, and by temporal spacing between when the annotations were created (because we had screen captures of the reading process) (see Fig. 3.

3.5. Procedure

All participants completed both a reading on paper and one on a computer. To avoid adverse effects such as learning bias, general fatigue, or information overload, we randomly assigned participants to two conditions: A or B. Participants in condition A, performed their first reading (first half of the text) on paper and their second reading (second half of the text) on the laptop; whereas subjects assigned to condition B completed their first reading on the laptop and their second reading on paper.

After the readings, all participants completed the memory test associated with their text. The students had as much time as they wished to finish both the readings and the memory tests, as time constraints have shown to be one of the significant moderators in comparison studies between digital and physical, where a general increase in comprehension in paper reading is augmented in studies where participants are under time pressure (Delgado et al., 2018).

3.5.1. Pilot experiment

We performed a pilot study of the experiment with two independent student participants, one in condition A (paper reading first, computer reading second) and one in condition B (computer reading first, paper reading second). The pilot study mainly yielded insights about which

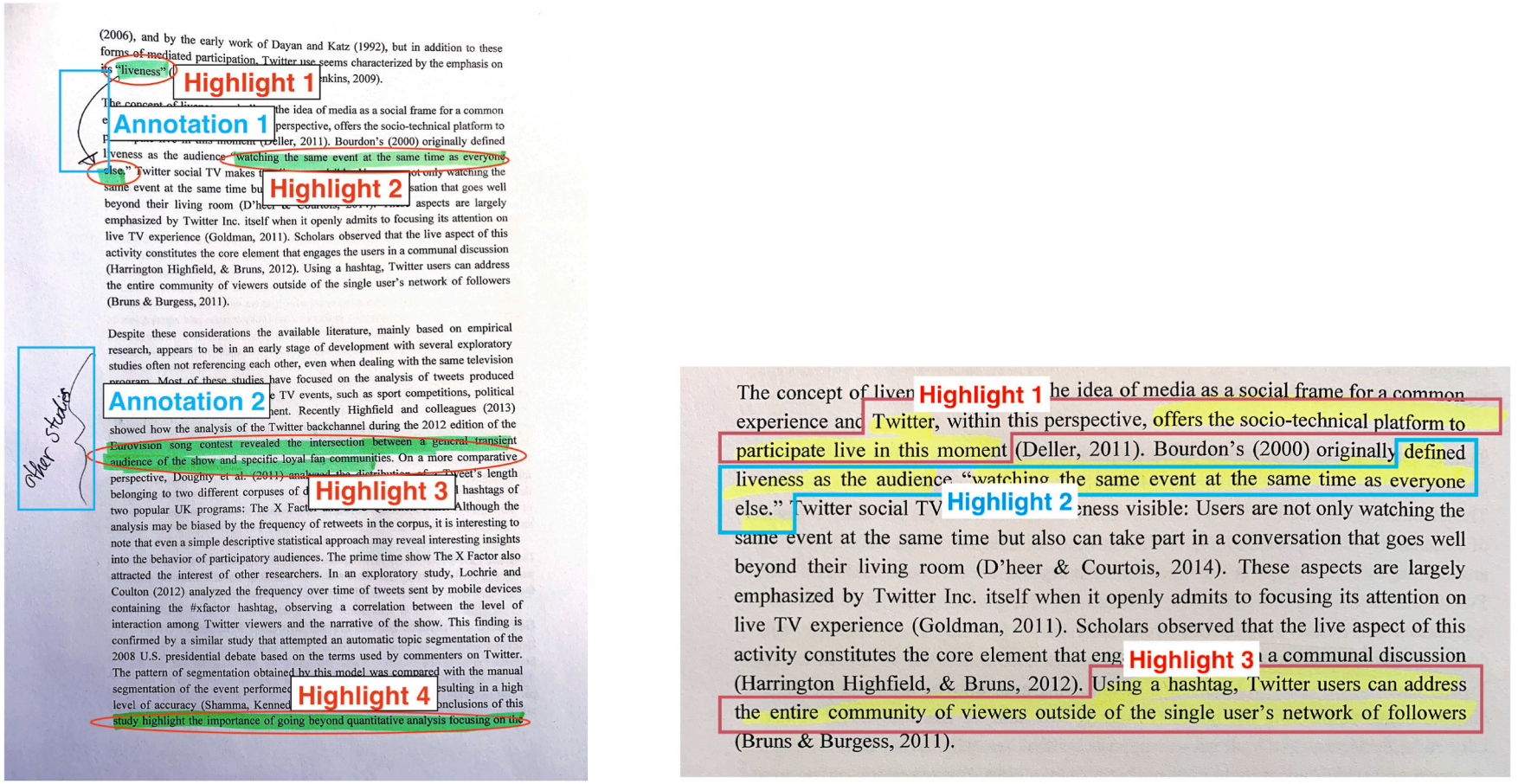

Fig. 2. Left: Both one highlighted word (Highlight 1) and one highlighted sentence (Highlight 2, 3, and 4) were counted as 'one highlight'. One self-contained annotation, such as an arrow (Annotation 1) or a mark together with a comment (Annotation 2) was counted as 'one annotation'. Right: When the highlight is interrupted over a sentence, the highlight was counted as one highlight if the highlighted part makes sense in itself (Highlight 1: "Twitter … offers the socio-technical platform to participate live in this moment").

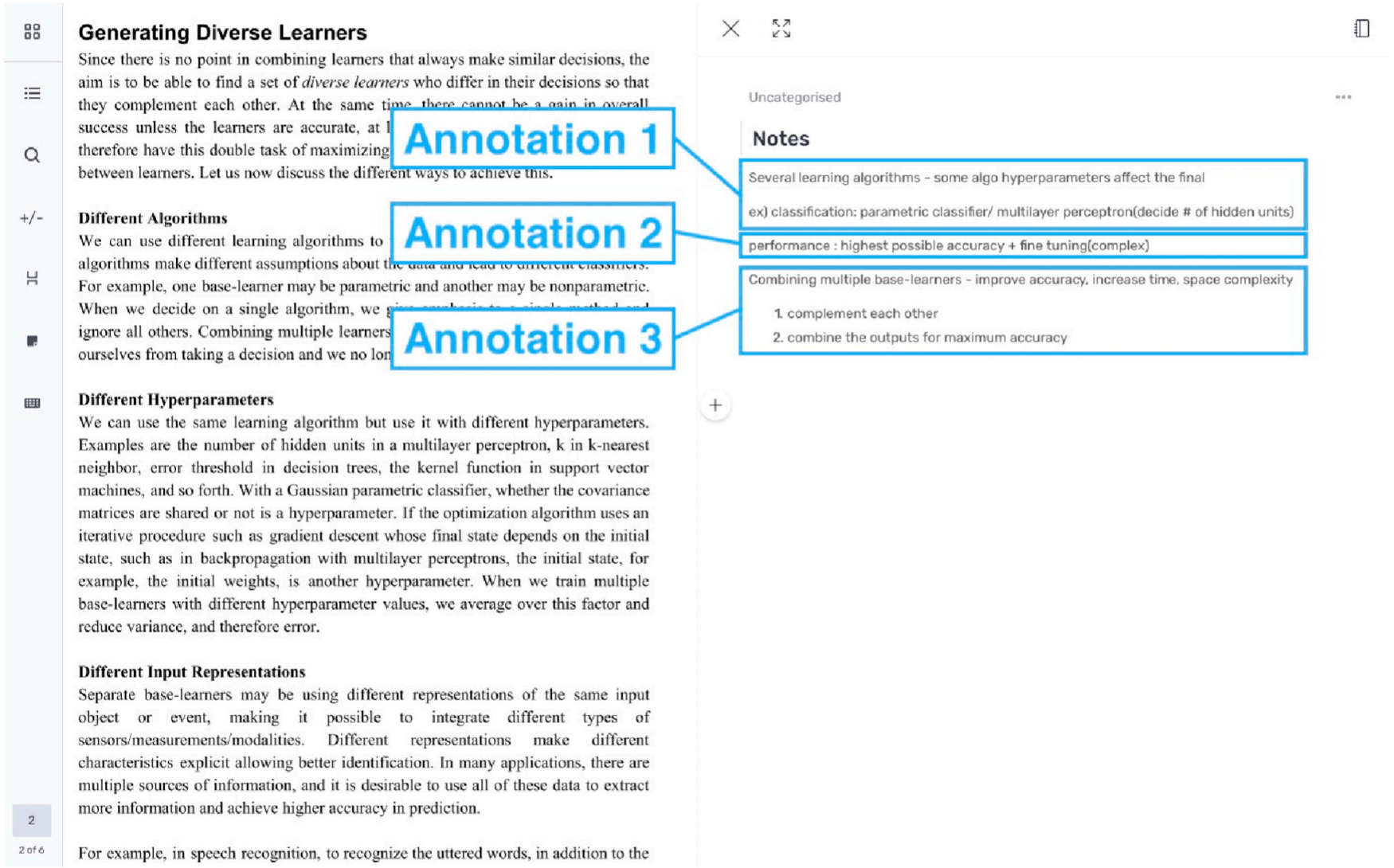

Fig. 3. Digital annotations are distinguished by either their anchoring to different parts of the text, by formatting, or by temporal spacing between when they were written.

information we should provide the study participants with, based on the questions we received during the pilot, such as "Can I use all the features of the Lix software?" and "How do I change an answer in the memory test?". We also discovered minor typesetting errors in the material, and observed that the computers should be disconnected to the internet to avoid any unwanted notifications.

3.5.2. Main experiment

The main study was carried out in a controlled, but familiar environment, in rooms at the university. Participants in condition A and condition B were reading in separate rooms, to avoid distractions from the switch between computer and paper, but also to avoid revealing information about the next task. The students were not told what would happen during the experiment, but instructed to wait for information about each task.

Participants were welcomed and thanked for their participation, and asked to turn off their phones. They knew beforehand that they would perform one or more readings relevant to their studies, but no further details.

The students were given their first text in either digital or paper format. They were instructed to read the text "as if you were preparing for class or exam, making sure to understand the major points of the text", but they were neither informed about the following memory test, nor that they had only received half a text. For the paper reading, students were presented with the tools in front of them (highlighter pen, sticky notes, and pens), and told that they could use them if they wished. For the computer reading they were given a short introduction to the Lix software and where they could find the different features and also informed that they could use them if they wished. They were not told whether they would have access to the text and notes after the reading, because we wanted them to complete their reading in as natural a way as possible, rather than worrying about studying for a test. Participants were told to indicate when they had finished their reading, and to close the computer or put the text away.

After the first reading, the second half of the text was distributed in the 'opposite' medium of before, and the participants received new introductions to the interaction tools available to them. Again, they were instructed to indicate when they had finished the reading, and to close the computer or put the text away. To provide detailed data about the students' interaction with the digital text, we set up the computers to record screen capture videos during the computer readings. However, during the execution of the experiment, two of the screen captures malfunctioned, and two data points were therefore excluded from highlight and annotation data. The data about digital highlights and annotation thus come from n = 48 students.

Post completion of both readings, the students had a short break, during which they were told not to speak to each other about the texts, and to not use their mobile phones.

Following the break, the memory tests were distributed. We gave a short instruction to the meaning of the two types of questions "Fill in the blank(s)" and "According to the text, is this correct?", and we asked if there were any questions. As this was not a timed task, they were allowed to study the test before starting. No participants had questions about the test.

3.6. Data analyses

The data collected in this experiment were: the time spent reading for each participant in two different media, the paper and digital texts with any highlights or annotations and notes produced during the readings (as well as screen captures for all digital readings[4]), and the completed memory tests for each participant.

For all our statistical analyses, we adopt a 95% confidence interval (i.e., $\alpha = 5\%$) when comparing data from paper vs. laptop (Du Prel, Hommel, Röhrig, & Blettner, 2009). We have performed Kolmogorov-Smirnov tests for all data to determine normality of distribution, t-tests and Pearson's Correlation Coefficient for data that follows normal distribution, Wilcoxon Signed-Rank tests and Spearman's rho for all data that do not follow normal distribution. Finally, we use *two tailed* tests which conservatively do not impose any assumptions on the order of any potential effects.

4. Results

4.1. How does academic reading on a laptop compare to physical paper in terms of reading time? (RQ1)

Observation 1. We observed no statistically significant difference in reading time between laptop and paper.

Fig. 4 shows box plots of the distribution of academic reading times (in Words Per Minute, WPM) on paper vs. laptop. Each box plot shows: the minimum, median, and maximum reading time along with the (25%) lower and (75%) upper quartiles (excluding subjects with reading times beyond 230 WPM). The left box plot depicts reading time on paper; the right figure plots the same for reading on a laptop. We see that the box plots largely overlap. The median reading times are comparable with 136 WPM on paper compared to 130 WPM (4% slower) on a laptop. The same goes for the averages (not shown) with 133 WPM on paper compared to 126 WPM (5% slower) on a laptop.

A Kolmogorov-Smirnov Test reveals that the data for reading time on paper are *not* normally distributed ($D = 0.23, p = .01$), but excluding one outlier data point with a reading time of 362 words per minute gives a normal distribution ($D = 0.18$, $p = .073$). The data for reading time on laptop *does* follows normal distribution ($D = 0.16, p = .15$), and thus we performed a paired *T*-test for dependent means. The *t* for these data is -1.65, $p = .11$, and the difference is *not* significant at $p < .05$.

4.2. How does academic reading on a laptop compare to physical paper in terms of interaction? (RQ2)

We answer this research question by looking first at highlights, then

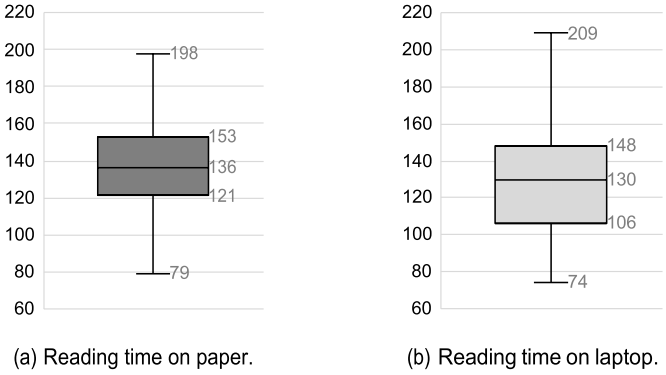

Fig. 4. Distributions of reading time in Words Per Minute (WPM).

[4] Screen captures allowed us to further investigate how students interacted with the software, such as attempting to use an interaction feature but not succeeding. We concluded, based on these recordings, that students seemed to accomplish the tasks they attempted to.

at annotations.

Observation 2a. We observed no statistically significant difference in the amount of highlighting on a laptop compared to physical paper.

Fig. 5a illustrates the percentage of readers using highlighting on paper versus laptop. We see that a third (17 out of 50) of the subjects did *not* use highlighting on paper. The same is true for laptop (16 out of 48[5]), however, these were not necessarily the same people. A fifth of the subjects (10 out of 48) never used highlighting in either medium. In general, we see a similar usage frequency of highlighting on laptop compared to paper. The average percentage[6] of words highlighted are 8.5% on paper compared to 7.2% on laptop.

These data follow a slightly skewed from normal distribution ($D = 0.21$, $p = .025$ for paper, $D = 0.22$, $p = .02$ for laptop), and thus we have performed a Wilcoxon Signed-Rank tests to reveal that the difference in amount of words highlighted per one hundred words is *not significant* when comparing laptop to paper ($z = -1.17$, p = .24).

Observation 2b. Readers created significantly fewer annotations when reading on a laptop compared to paper.

Fig. 5b shows the frequency of readers utilizing annotations on paper compared to on a laptop. We see that 40% of readers (19 out of 48) never annotate on paper. On a laptop, however, this number ascends to a staggering 83% (40 out of 48). A Fisher Exact test reveals that this difference is statistically significant at $p < .01$.

4.3. How does academic reading on a laptop compare to physical paper in terms of memory? (RQ3)

Observation 3. We found no statistically significant difference in memory of the text after reading on paper compared to reading on a laptop.

Fig. 6 shows the distribution of memory test scores on paper vs. laptop. We see that readers appear to score (slightly) better when they have read on paper than when they have read on a laptop. This is also captured by the averages which are 3.7 points on paper vs. 3.4 on the laptop (8% worse).

A Kolmogorov-Smirnov test reveals that the memory test scores for the paper part of the reading are *not* normally distributed ($D = 0.23$, $p < .001$), so we performed a Wilcoxon Signed-Rank test to compare the scores for the paper and the laptop reading. This test shows that the difference between the two conditions is *not statistically significant* at $p < .05$ ($z = -1.24$, $p = .21$).

4.4. What is the relationship between reading time and memory on paper vs. laptop? (RQ4)

Observation 4. We observed no statistically significant correlation between reading time and memory test score on paper, but a weak correlation between reading time and memory test score on laptop and on average.

A Spearman's Rho test shows no statistically significant correlation between the reading time (words per minute) and memory test scores for paper: $r_s = .13$, p = .34. Both the reading time and the memory test scores are normally distributed for the laptop condition, and interestingly, a Pearson's Correlation Coefficient shows a weak positive correlation between the two ($r(48) = 0.08$), although the *p*-value shows a non-statistically significant result ($p = .60$).

Following this result, we were interested to see whether there was a general correlation between the reading times and memory test scores, so we compared the average reading times per participant with their total memory test score (for both paper and laptop condition), and also found a weak positive correlation with $r(48)$[7] $= 0.24$, although also a statistically non-significant result ($p = .09$).

4.5. What is the relationship between interaction and reading time on paper vs. laptop? (RQ5)

We answer this research question by first looking at *highlighting*, then *annotations*.

Observation 5a. *We found a positive correlation between highlighting more of a text and reading faster on paper. We found a lower average reading time for those who used highlighting at on a laptop.*

17 readers did not use highlights at all on paper. If we divide the readers into two groups; readers "using highlighting" vs. readers "not using highlighting", the average reading time of the 33 highlighters vs. the 17 non-highlighters are 136 WPM vs. 149 WPM; i.e., those who use highlighting read 8% *slower* than non-highlighters. A *T*-test for independent means reveals that this difference is *not* statistically significant: $t(16 + 32) = 0.88$, p = .39.

If we look at readers using highlighting and their reading time, a Pearson's correlation coefficient test shows a statistically *significant* positive correlation with $r(31) = 0.40$, $p = .02$, meaning that of the readers who highlight, those who highlight more also seem to read at a faster speed.

The amount of readers who did not highlight at all is also high for the **laptop** condition (16 out of 48). Dividing the readers into "highlighters" and "non-highlighters" shows that highlighters read, on average, (121 WPM versus 148 WPM) 18% slower than non-highlighters on laptop, which is statistically *significant* $p = .01$ ($t(15 + 31) = 2.57$).

A Pearson's Correlation Coefficient test shows a weak *negative* correlation between reading time and amount of document highlighted for those who used highlights in the laptop condition: $r(30) = -0.11$, however the result is not statistically significant at $p = .55$. It seems that highlighting more on a laptop correlates with slower reading times, but those who do not highlight at all on a laptop generally read fewer words per minute.

Observation 5b. On paper, we observed a negative correlation between creating more annotations and reading time. On laptop, the same correlation was found, but not statistically significant.

21 of 50 readers created no annotations on **paper**. The average reading time was 138 WPM for those who did not annotate at all versus 142 WPM for those who created at least one annotation. Unsurprisingly, this difference is not significant at $t(20 + 28) = -0.29$, $p = .77$.

A Spearman's Rho test reveals that there is a statistically significant *negative* correlation between reading time and number and annotations created ($r_s = -0.40$, $p = .03$), meaning that those who create more annotations also read at lower speeds.

Only 8 of 48 readers created annotations in the **laptop** condition. The average reading for those was 124 WPM versus 131 WPM for the readers not creating annotations. This difference is *not* statistically significant: $t(7 + 39) = 0.50$, $p = .62$. A Pearson's Correlation Coefficient test shows a weak *negative* correlation between amount of annotations and reading time ($r(6) = -0.40$), but one which is not statistically significant ($p = .33$) - which is unsurprising, given the low n.

[5] Recall that the interaction data collection malfunctioned for two subjects; hence N = 48 (cf. Section 3.5.2).

[6] Since the students read texts of slightly different lengths, we counted the number of highlighted words per hundred words in the text, rather than absolute number of words highlighted.

[7] The average reading time and memeory score per participant followed normal distribution, according to a Kolmogorov-Smirnov-test.

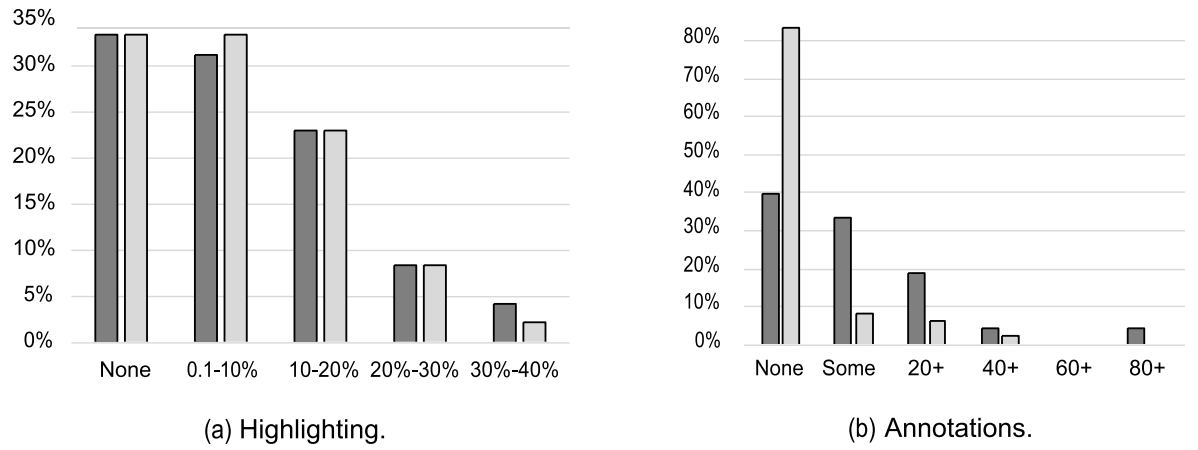

Fig. 5. Histogram with percentage of readers using what frequencies of textual interactions on paper (dark gray) vs. laptop (light gray). To the left: The (relative) percentage of words highlighted. To the right: The (absolute) number of annotations made.

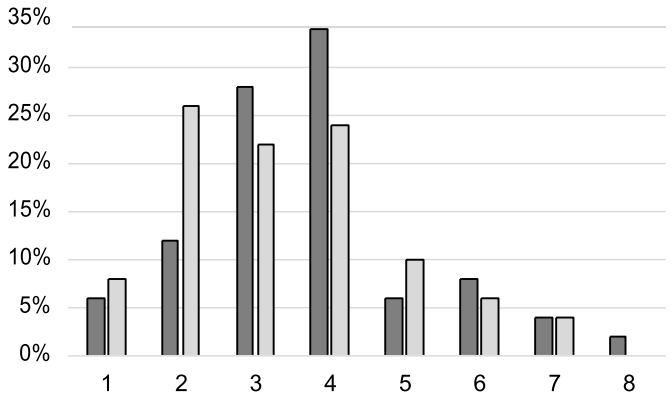

Fig. 6. Histogram with percentage of readers achieving which memory scores on paper (dark gray) vs. laptop (light gray).

4.6. What is the relationship between interaction and memory on paper vs. laptop? (RQ6)

Again, we begin by looking at *highlights*, and then *annotations*.

Observation 6a. We observed no relationship between highlighting and memory test score, neither on paper, nor on laptop.

On **paper**, both the 33 students who used highlights and the 17 students who did not use highlights scored an average of 3.7 points on the memory test.

A Spearman's Rho test shows *no* statistically significant correlation between proportion of document highlighted and memory test score: $r_s = -0.33$, $p = .06$.

On **laptop**, the 16 students who did not use highlighting scored an average of 2.9 on the memory test, while the 32 students who did use highlighting scored 3.4 on average. This difference is *not* statistically significant at $t(15 + 31) = -1.29$, $p = .20$.

A Pearson's Correlation Coefficient test shows a weak, but *not* statistically significant positive correlation between proportion of document highlighted and memory test score: $r(30) = 0.04$, $p = .83$.

Observation 6b. We observed no correlation between annotations and memory test score, neither on paper, nor on laptop.

The 21 readers who created no annotations on **paper**, scored 3.8 in the memory test on average, versus 3.7 for the 29 students who created at least one annotation. Unsurprisingly, this is *not* statistically significant at $t(20 + 28)0.16$, $p = .87$.

For the readers who did create annotations, a Spearman's Rho test shows *no* correlation between the amount of annotations created and the memory test score: $r = 0.16$, $p = .41$.

The vast majority of readers who created no annotations on **laptop** scored, on average, 3.1 on the memory test, while the 8 readers who created annotations scored 4 points on average. This difference is *not* statistically significant at $(7 + 39) = 1.65$, $p = .11$.

A Pearson's Correlation Coefficient shows a weak negative correlation between memory test score and amount of annotations created, but again, the result is based on only 8 data points and thus *not* statistically significant: $r(6) = -0.16$, $p = .71$.

5. Discussion

This section will describe how our findings relate to existing research and how the findings contribute to our knowledge of how interaction design can positively improve digital reading environments for academic reading.

Although we saw a small absolute difference in reading time (WPM) on paper versus computer, with reading on paper being slightly faster, the *difference between reading time on paper and on laptop was statistically non-significant* (**RQ1**). This result replicates other recent comparison studies of reading speed in computer versus paper studies (Delgado et al., 2018; Kara et al., 2019; Kong et al., 2018), and hints that laptop is generally equal to or slightly less time efficient than paper reading, even when the element of digital distractions is removed from the equation (Pálsdóttir, 2019) (in the experiment setup, students read on a laptop with no internet access, and without access to their phones). It is possible that the difference between words per minute in paper and laptop would be even smaller if students had read on their personal laptops using more familiar digital environments/software.

The amount of interactions with the physical versus digital text was interesting, especially that we saw *almost the same average amount of highlights in the paper as the laptop reading* (**RQ2**). The difference between average proportion of document highlighted on laptop was not statistically significant from the average number of highlights on paper, which is interesting when comparing the result to previous findings, e.g. (Johnston & Ferguson, 2020; Kara et al., 2019; Liu, 2005; Mizrachi et al., 2018), which found that students did not interact with digital documents, even if they knew the features to do so were available. In the digital environment provided in this study, highlights seem to be used as frequently. We speculate that this variance could be due to three factors: 1) Students are constantly becoming more familiar with reading digital documents, and newer studies will therefore find that younger students are becoming increasingly used to using interactive features when reading digitally. 2) The interface of the Lix software may offer a particularly good highlighting feature, which is lightweight enough that

it is attractive to use without cognitive effort (Pearson et al., 2013), and 3) There could be a random variance in this particular study, due to factors such as the student population (which were from IT-related programs), or we simply biased the students by introducing them to the highlighting feature before the beginning of the study.

The amount of annotations on laptop was, conversely, significantly lower (more than 80% of participants did not use digital annotation features at all, against 40% not annotating on paper) *than the amount of annotations made on paper* (**RQ2**). This observation corresponds to previous research findings.

This finding attests to the importance of distinguishing between highlighting and other annotation features in digital reading software, in that we observed that the digital annotation features were less popular than their analog counterparts. Typing on a keyboard is generally faster than writing with a pen, and we could therefore expect participants to take more notes during the digital reading than during the paper reading (because typing is simply easier). This does not seem to be the case. We suspect that the low amount of annotations in the digital readings is caused by one of two factors: 1. Either, the interface of the software provided does not provide a good way for students to annotate the text, or 2. The students were not motivated to take notes because the reading was isolated, in the sense that they did not have access to their regular note-taking tools or environments.

Memory test scores were similar from paper to laptop reading, although 8% lower for laptop in absolute numbers (RQ3). This result corresponds to what we would expect based on meta-studies that say the comprehension deficit for digital reading seems to be narrowing in newer studies (Kong et al., 2018).

There was *no relation to observe between reading time and memory on paper, but a weak correlation between reading time and memory on laptop and overall*, where higher reading time correlated with higher score on the memory test (**RQ4**). We interpret from this result that people simply have different academic reading times, and that self-managed reading speed does not appear to be a moderating factor of reading comprehension.[8]

The use of *highlighting appears to correlate with higher reading time on paper, but negatively correlate with reading time on laptop* (**RQ5**). The positive correlation on paper is somewhat contrary to previous findings stating that highlighting slows readers down (Dunlosky et al., 2013a, 2013b), at least for the paper-based reading. It is possible that readers, who read fast, tend to also highlight more.

While we saw no correlational relation between amount of highlighting on a laptop and reading time, we did observe a lower average reading time for those who highlighted - or the relation could be reversed, so that those who read slower tend to highlight more. When looking at the screen captures, we saw that it was a common behavior in the digital environment to use the mouse cursor to "follow" the text as a form of temporary placeholder (as one might do with a finger on paper) (Pearson et al., 2013). When the cursor is already close to the text it is quick to use it to highlight, and this could cause some readers to highlight more. The Observation could also indicate that the software does not support highlighting in a sufficient way to at least not hinder the reading.

For the nine participants who created annotations on paper, *we saw a decrease in reading time* (**RQ5**) of 138 WPM (versus 142 WPM for those who did not annotate at all). The group of participants creating any digital annotations at all was so low (eight students, out of whom two only created one note/comment each) that it is, at best, questionable to derive conclusions about their performance in relation to the other participants. The three participants who created a substantial amount of annotations (around 20 or more) spent, on average, longer than the overall average on the reading, which is unsurprising, in that writing notes takes time.

We observed *no relationship between highlighting and memory test score*, neither on paper, nor on laptop (**RQ6**). The average memory test score was 17% higher in absolute numbers for students who highlighted on laptop than for students who did not, and the average score was the same for highlighters and non-highlighters on paper. This finding mirrors the conclusions of Dunlosky and colleagues, stating that highlighting does not work for obtaining memorizing goals (Dunlosky et al., 2013a, 2013b).

We observed no correlation between annotations and memory test score, neither on paper, nor on laptop (**RQ6**). On paper, there was almost no difference in memory test scores between those who annotated and those who did not, and no correlation between creating more annotations and memory test score. As expected based on previous research, the amount of students who used annotation features on laptop was small, however, the students who did annotate scored, on average, 29% higher on the memory test in absolute numbers. The difference was not significant, and we found a weak, but statistically non-significant negative correlation between creating more annotations and memory test score. This indicates that the relation could be reversed, so that those who found the text more challenging tried to compensate for this by taking more notes.

The observations based on research question 2: H*ow does academic reading on a laptop compare to physical paper in terms of interaction?* and 6: *What is the relationship between interaction and memory on paper vs. laptop?* support recent research findings in that there are few statistically significant differences between the two media to be found. Our experiment recruited students to read a text that was part of their normal curriculum in a natural setting, and the main difference we saw was that students created significantly more annotations on paper than on laptop, but that creating annotations did not correlate with a higher score in a subsequent memory test.

We can therefore conclude that highlighting and annotation creation do *not* seem to be moderating factors in comprehension performance between academic students reading on paper versus laptop - at least not in the simplest form that these interaction features are deployed in this software. There were weak differences in performance in absolute numbers between some of the metrics, and we are looking forward to investigating these further, such as the slightly higher memory test score for the few students who annotated on laptop and the positive correlation between reading time and highlighting on paper, while a negative correlation on laptop.

6. Threats to validity

6.1. Internal threats

Are the metrics and measurements valid? A potential threat to validity is of course whether the metrics of the study are the correct ones, and whether they are accurate. Particularly the measurement of reading speed/time, which is derived from self-evaluation. It was clear from the screen captures, that different students approached the reading task in different ways – some skim read the text before or after the thorough reading, and some did not – this behavior is to be expected (Johnston & Ferguson, 2020). The participants were under Observation for the duration of the study, and their laptop reading was captured on video which mitigates the risk that the time measurements are skewed. The memory tests were designed with either *correct* or *incorrect* answers, so no subjective evaluations took place in the evaluation of these. The 'interactions' (highlights and annotations) were counted in the read documents, and we have clarified our approach to these measurements in accordance with previous research (Marshall, 1997; Pearson et al., 2013), but some subjective assessment has taken place in categorizing 'self-sufficient' segments of particularly highlights. We do expect that any discrepancies in these counts even out with increased numbers.

[8] Time constraints, on the other hand, have shown to be a moderating factor which enhances the advantage in comprehension when reading on paper (Delgado et al., 2018).

Is interaction with paper texts comparable to interaction with digital texts? To achieve as reliable comparability as possible between the interaction modalities of the paper and the digital reading, we tested with a reading software with lightweight interaction features that corresponded to the tools provided in the paper reading; highlighting, annotation, and notes. Although interaction with a laptop will not correspond directly to the interaction with paper and pens—which would defeat the whole purpose of this study—research suggests that it is possible to *imitate* physical interaction features in a way that fosters equally lightweight interaction (Pearson et al., 2013).

Is the sample size too limited? For pragmatic reasons, we conducted the study with N = 50 participants. This amount of participants was a non-trivial setup due to availability of equipment, as well as correct management of texts and memory tests. The amount of participants should be enough to uncover stronger effects, even when we subdivide the data into smaller groups (for instance, highlighters and non-highlighters). The amount may, however, be insufficient for uncovering more subtle effects. It would obviously be interesting to run larger experiment which would perceive more effects and yield stronger conclusions.

Was there a subject selection bias? The participants signed up by responding to in-class presentations about the study as well as flyers distributed around campus. They were told that the study was about investigating digital reading, so there could have been some self selection bias in that mostly students who are interested in digital reading would sign up. However, we made it clear in the presentations that the study was for everyone who were interested in helping us learn more about reading course material (even if their personal preferences were reading in physical media).

There could have been some selection bias rising in selecting the different texts for the experiment. We attempted to mitigate such a bias by reaching out to the course responsible lecturers for selection of texts.

Was the hardware and software familiar to the participants? The laptops provided were of a brand and size familiar to university students, which we confirmed with study participants at the beginning of the experiment. As we received some questions about the Lix software, we learned that we could have carried out a more comprehensive demo task, allowing the students to familiarize themselves even more with the program. The interface of the program is very simple, and all students managed to complete the reading task with no major obstacles, but there could potentially be a small bias in the reading time and interaction metrics

6.2. External threats

Beyond students? Our experiment studied university students of a European university. The students came from four different study programs, albeit all related to information technology. Also, they all have easy access to technology as part of their daily practice. A drop in time, comprehension and ease of interaction might be observed in different populations, such as older students, non-students, non-Western students, or perhaps students in different fields, such as classical music or medicine, where learning strategies and use of literature may vary.

Same text or different texts? As described in the *Texts* section (3.4.1), we prioritized that the text fit the students' *internal motivation* and *skill/specialization*. Our belief is that humans are inherently different in their approach to a text, and that we would achieve higher external and ecological validity of the results by selecting a text appropriate for the individual subject's current situation and educational level. For internal validity, it would have been more valid to have all subjects read the *same* text. However, we have essentially traded internal for external validity in that it is a more realistic setting to read literature that corresponds to one's own curriculum.

Beyond lab setting? One of the main caveats of temporally constrained experiments is that the results only measure short-term reading and learning, and we can not know if the findings generalize to students' own, long term habits and practices. We have devoted some effort into designing the experiment to be as true to natural study settings as possible, but further studies should be conducted to understand more about students' day-to-day practices with paper and digital reading. One of the main problems often mentioned by students when studying digitally is the constant exposure to digital environment distractions (Pálsdóttir, 2019), and these are usually ignored in a lab experiment setup. Another caveat with lab settings is that most research tends to make broader comparisons between hardware, rather than studying the unique interaction features of different digital platforms in detail (Kara et al., 2019), which is something we have attempted to pursue in this study.

7. Conclusions

As digital reading becomes increasingly prevalent in university education, it becomes more important to know how digital reading shapes reading comprehension, and which factors of digital software influence reading experiences, preferences, and learning. The exploratory study presented in this paper contributes to this knowledge by exploring 50 university students' active reading styles, closely mimicking their normal study circumstances. We looked at reading time, highlighting, and annotation of the text, and subsequent reading comprehension in the form of memory tests.

We did not find statistically significant differences in the memory score for the students who used highlights and annotations, and those who did not – neither on paper, nor on laptop. This study suggests that highlighting and annotation features in a simple form are *not* moderating factors of digital reading performance for academic students. In terms of designing digital reading tools, these findings demonstrate the importance of investigating interaction features further. Our broader contribution to education research is to expand understanding of how different tools for reading in a university environment are used.

CRediT authorship contribution statement

Nanna Inie: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Louise Barkhuus:** Writing – original draft. **Claus Brabrand:** Formal analysis, Writing – original draft, Writing – review & editing, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the anonymous reviewers for their incredibly constructive and valuable reviews, and the students who participated in this study. This research was funded by the Innovation Fund Denmark, grant number 9066-00006B.

References

Abuloum, A., Farah, A., Kaskaloglu, E., & Yaakub, A. (2019). College students' usage of and preferences for print and electronic textbooks. *International Journal of Emerging Technologies in Learning, 14*, 7, 2019.

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied, 17*(1), 18, 2011.

Ackerman, R., & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior, 28*(5), 1816–1828, 2012.

Adler, A., Gujar, A., Harrison, B. L., O'hara, K., & Sellen, A. (1998). A diary study of work-related reading: Design implications for digital reading devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 241–248).

Adler, M. J., & Van Doren, C. (2014). *How to read a book: The classic guide to intelligent reading*. Simon and Schuster.

Annisette, L. E., & Lafreniere, K. D. (2017). Social media, texting, and personality: A test of the shallowing hypothesis. *Personality and Individual Differences, 115*, 154–158, 2017.

Baron, N. S. (2015). *Words onscreen: The fate of reading in a digital world*. USA: Oxford University Press.

Ben-Yehudah, G., & Eshet-Alkalai, Y. (2018). The contribution of text-highlighting to comprehension: A comparison of print and digital reading. *Journal of Educational Multimedia and Hypermedia, 27*(2), 153–178, 2018.

Björnsson, C.-H. (1968). *Läsbarhet*. Liber.

Brady, K., Cho, S. J., Narasimham, G., Fisher, D., & Goodwin, A. (2018). *Is scrolling disrupting while reading?* International Society of the Learning Sciences, Inc.[ISLS].

Buchanan, G., & Pearson, J. (2008). Improving placeholders in digital documents. In *International conference on theory and practice of digital libraries* (pp. 1–12). Springer.

Chen, N., Guimbretiere, F., & Sellen, A. (2012). Designing a multi-slate reading environment to support active reading activities. *ACM Transactions on Computer-Human Interaction, 19*(3), 1–35, 2012.

Conway, M. A., Gardiner, J. M., Perfect, T. J., Anderson, S. J., & Cohen, G. M. (1997). Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General, 126*(4), 393, 1997.

Delgado, P., & Salmerón, L. (2021). The inattentive on-screen reading: Reading medium affects attention and reading comprehension under time pressure. *Learning and Instruction, 71*(2021), 101396.

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review, 25*, 23–38, 2018.

Dillon, A., McKnight, C., & Richardson, J. (1988). Reading from paper versus reading from screen. *The computer journal, 31*(5), 457–464, 1988.

Du Prel, J.-B., Hommel, G., Röhrig, B., & Blettner, M. (2009). Confidence interval or p-value?: Part 4 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International, 106*(19), 335, 2009.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013a). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58, 2013.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013b). What works, what doesn't. *Scientific American Mind, 24*(4), 46–53, 2013.

Frank, S., Price, M., Marshall, C. C., & Golovchinsky, G. (2003). Identifying useful passages in documents based on annotation patterns. In *International conference on theory and practice of digital libraries* (pp. 101–112). Springer.

Freund, L., Kopak, R., & O'Brien, H. (2016). The effects of textual environment on reading comprehension: Implications for searching as learning. *Journal of Information Science, 42*(1), 79–93, 2016.

Hillesund, T. (2010). *Digital reading spaces: How expert readers handle books, the Web and electronic paper*, 2010.

Inie, N., & Barkhuus, L. (2021). Developing evaluation metrics for active reading support. *CSEDU*, (1), 177–188.

Jennifer Sarah Pearson. (2012). *Investigating lightweight interaction for active reading in digital documents*. United Kingdom): Swansea University.

Johnston, N., & Ferguson, N. (2020). University students' engagement with textbooks in print and E-book formats. *Technical Services Quarterly, 37*(1), 24–43, 2020.

Kara, S., Augustine, H., Shand, H., Bakner, K., & Rayne, S. (2019). Reading from print, computer, and tablet: Equivalent learning in the digital age. *Education and Information Technologies, 24*(4), 2477–2502, 2019.

Kawase, R., Herder, E., & Nejdl, W. (2009). A comparison of paper-based and online annotations in the workplace. In *European conference on technology enhanced learning* (pp. 240–253). Springer.

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*, 1975.

Kobayashi, K. (2005). What limits the encoding effect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology, 30*(2), 242–262, 2005.

Kol, S., & Schcolnik, M. (2000). Enhancing screen reading strategies. *Calico journal*, 67–80, 2000.

Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education, 123*, 138–149, 2018.

Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior, 35*, 455–463, 2014.

Liu, Z. (2005). Reading behavior in the digital environment. *Journal of Documentation, 61*(6), 700–712, 2005 https://doi.org/10.1108/00220410510632040.

Logan, S., Medford, E., & Hughes, N. (2011). The importance of intrinsic motivation for high and low ability readers' reading comprehension performance. *Learning and Individual Differences, 21*(1), 124–128, 2011.

Loh, K. K., & Kanai, R. (2016). How has the Internet reshaped human cognition? *The Neuroscientist, 22*(5), 506–520, 2016.

Mangen, A., & Don, K. (2014). Lost in an iPad: Narrative engagement on paper and tablet. *Scientific Study of Literature, 4*(2), 150–177, 2014.

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research, 58*(2013), 61–68.

Margolin, S. J., Driscoll, C., Toland, M. J., & Kegler, J. L. (2013). E-readers, computer screens, or paper: Does reading comprehension change across media platforms? *Applied Cognitive Psychology, 27*(4), 512–519, 2013.

Marshall, C. C. (1997). Annotation: From paper books to the digital library. In *Proceedings of the second ACM international conference on Digital libraries* (pp. 131–140).

Mizrachi, D., Salaz, A. M., Kurbanoglu, S., Boustany, J., & ARFIS Research Group. (2018). Academic reading format preferences and behaviors among university students worldwide: A comparative survey analysis. *PLoS One, 13*, 5, 2018, e0197444.

Morehead, K., Dunlosky, J., Rawson, K. A., Blasiman, R., & Hollis, R. B. (2019). Note-taking habits of 21st century college students: Implications for student learning, memory, and achievement. *Memory, 27*(6), 807–819, 2019.

Norman, E., & Furnes, B. (2016). The relationship between metacognitive experiences and learning: Is there a difference between digital and non-digital study media? *Computers in Human Behavior, 54*, 301–309, 2016.

Noyes, J. M., & J Garland, K. (2008). Computer-vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*(9), 1352–1375, 2008.

OECD. PISA programme for international student assessment. n.d. https://www.oecd.org/pisa/

O'Hara, K. (1996). *Towards a typology of reading goals*, 1996.

Pálsdóttir, Á. (2019). Advantages and disadvantages of printed and electronic study material: Perspectives of university students. *Information Research, 24*, 2 (2019), Retrieved from http://InformationR.net/ir/24–2/paper828.html.

Pearson, J., Buchanan, G., & Thimbleby, H. (2013). Designing for digital reading. *Synthesis lectures on information concepts, retrieval, and Services, 5*(4), 1–135, 2013.

Pearson, J., Buchanan, G., Thimbleby, H., & Jones, M. (2012). The digital reading desk: A lightweight approach to digital note-taking. *Interacting with Computers, 24*(5), 327–338, 2012.

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology, 80*(1), 16, 1988.

Rockinson-Szapkiw, A. J., Courduff, J., Carter, K., & Bennett, D. (2013). Electronic versus traditional print textbooks: A comparison study on the influence of university students' learning. *Computers & Education, 63*, 259–266, 2013.

Schilit, B. N., Golovchinsky, G., & Price, M. N. (1998). Beyond paper: Supporting active reading with free form digital ink annotations. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 249–256).

Schutte, N. S., & Malouff, J. M. (2007). Dimensions of reading motivation: Development of an adult reading motivation scale. *Reading Psychology, 28*(5), 469–489, 2007.

Singer Trakhman, L. M., Alexander, P. A., & Berkowitz, L. E. (2019). Effects of processing time on comprehension and calibration in print and digital mediums. *The Journal of Experimental Education, 87*(1), 101–115, 2019.

Singer Trakhman, L. M., Alexander, P. A., & Silverman, A. B. (2018). Profiling reading in print and digital mediums. *Learning and Instruction, 57*, 5–17, 2018.

Singer, L. M., & Alexander, P. A. (2017a). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education, 85*(1), 155–172, 2017.

Singer, L. M., & Alexander, P. A. (2017b). Reading on paper and digitally: What the past decades of empirical research reveal. *Review of Educational Research, 87*(6), 1007–1041, 2017.

Subrahmanyam, K., Michikyan, M., Clemmons, C., Carrillo, R., Uhls, Y. T., & Greenfield, P. M. (2013). Learning from paper, learning from screens: Impact of screen reading and multitasking conditions on reading and writing among college students. *International Journal of Cyber Behavior, Psychology and Learning, 3*(4), 1–27, 2013.

Tashman, C. S., & Edwards, W. K. (2011). LiquidText: A flexible, multitouch environment to support active reading. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3285–3294).

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie canadienne, 26*(1), 1, 1985.

Vincent, J. (2016). Students' use of paper and pen versus digital media in university environments for writing and reading–a cross-cultural exploration. *Journal of Print Media and Media Technology Research, 5*(2), 97–106, 2016.

Witherby, A. E., & Tauber, S. K. (2019). The current status of students' note-taking: Why and how do students take notes? *Journal of Applied Research in Memory and Cognition, 8*(2), 139–153, 2019.

Wolfe, J. (2008). Annotations and the collaborative digital library: Effects of an aligned annotation interface on student argumentation and reading strategies. *International Journal of Computer-Supported Collaborative Learning, 3*(2), 141, 2008.

Zeng, Y., Xue, B., Xu, J., Gong, C., & He, H. (2016). The influence of e-book format and reading device on users' reading experience: A case study of graduate students. *Publishing Research Quarterly, 32*(4), 319–330, 2016.

Zhang, T., & Niu, X. (2016). *The user experience of e-books in academic libraries: Perception, discovery, and use*. Academic e-books: Publishers, librarians, and users, 2016.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 920, 1994.