Developing Evaluation Metrics for Active Reading Support

Nanna Inie[1][a] and Louise Barkhuus[1][b]
[1]*Department of Computer Science, IT University of Copenhagen, Rued Langaards Vej 7, Copenhagen, Denmark*
nans@itu.dk, barkhuus@itu.dk

Keywords: Reading Support Tools, digital reading, evaluation metrics, user experience design, user experience methodology

Abstract: Reading academic literature in digital formats is becoming more and more of a normalcy for students, but designers of reading support tools do not share common, metrics for evaluating such tools. This paper introduces our work in developing an evaluation form which we call the aRSX (active Reading Support IndeX). The aRSX-form is a quantitative means for evaluating whether a specific software or hardware tool supports active, academic reading in a way that resonates with personal *user experience* and *learning preferences* - in other words; whether the tool is practical and pleasant to use for the student who consumes academic literature. The paper presents the first and second iterations of the aRSX evaluation survey based on a preliminary exploratory experiment with 50 university students. The paper also describes how the evaluation form can be developed and used by designers of reading support tools.

1 INTRODUCTION

Academic reading is changing. The content and amount of what students read, the way they read, and the platforms they use to consume written content are highly influenced and changed by the availability of novel, digital software and hardware (Hayles, 2012; Pearson et al., 2013; Delgado et al., 2018). While numerous novel platforms and tools are developed to support classroom teaching and various forms of student collaboration and co-working, *reading* has not been the subject of similar innovation and support. Many studies have indicated that performance deficits between physical and digital platforms have narrowed in later years (defined as the post 2013) (Kong et al., 2018), and convenience factors of digital reading (such as cost, accessibility, and environmental impact) make many students choose to read on digital platforms (Vincent, 2016; Mizrachi et al., 2016).

The diversity of digital reading support tools is surprisingly low (Pearson et al., 2013; Mizrachi et al., 2016). Digital textbooks are rarely designed to look different from their physical instances (both are often distributed as PDFs, a format designed for printing to make a document stable and always look the same on all devices), and the hardware used to consume digital texts is largely confined to personal laptops and, less often, tablets (Mizrachi et al., 2016; Pearson et al., 2013; Sage et al., 2019). As Pearson, Buchanan and Thimbleby found in their work on developing lightweight interaction tools, existing digital document formats are "far from ideal", and both the software and hardware used for reading often supports casual reading much better than attentive, close interpretation of the text (Pearson et al., 2013).

From an educational perspective, this is at best an under-utilization of the great potential of digital tools that we could take advantage of. At worst, this may be a causal factor of the decline in reading abilities of students from elementary school through college (Hayles, 2012). In one meta-study of research on digital versus physical reading, Delgado and colleagues (Delgado et al., 2018) found that students seem to have become worse at reading in digital formats, and suggested that one of the causal factors may be *the shallowing hypothesis*, a rationale that states that 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 challenging tasks, such as reading comprehension, which requires sustained attention (Delgado et al., 2018; Annisette and Lafreniere, 2017).

One of the main reasons for student preferences for physical formats in place of digital has been found to be the *interaction* that physical formats af-

[a] https://orcid.org/0000-0002-5375-9542
[b] https://orcid.org/0000-0003-1306-249X

ford (Sage et al., 2019; Rockinson-Szapkiw et al., 2013; Farinosi et al., 2016; Pálsdóttir, 2019). Some evidence also suggests that the process of interacting with a *computer*, specifically, is the main cause of any performance deficits between paper and screen reading (Pearson et al., 2013).

With this paper, we suggest that rather than looking at differences between physical and digital platforms in a broad sense, we should investigate the user experience of different tools in more detail. A tablet is not just a tablet and a computer is not only a computer. A book is not the same as a loose sheet of paper, and a highlight pen is not the same as a blunt pencil. Different digital document readers can be used on the same tablet or computer, rendering broad comparisons between "digital" and "physical" somewhat meaningless. Rather, we should try to evaluate which interface features and interaction formats work well for different users and their learning preferences.

This paper contributes to the emerging field of digital reading support tools by suggesting an evaluation metric for reading support platforms and tools. We present the first version of the active Reading Support indeX, aRSX, a first step towards a standardized evaluation scheme that can be used by researchers and developers to, relatively quickly, identify how a reading tool supports usability and user experience. In addition, such an evaluation tool can be used to indicate "robust moderating factors" which may shed light on many of the seemingly inconsistent findings across studies of reading performance in different media (Delgado et al., 2018).

The paper presents the first and second iterations of the aRSX, which was developed with intent to explore the question: *How might we evaluate reading support tools with regards to user experience and personal learning preferences?* We seek to answer this question by conducting an exploratory experiment with 50 university students, comparing their evaluation of a paper and a computer-based academic reading medium.

2 Background and related work

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, e.g. (Singer and Alexander, 2017; Dillon et al., 1988; Rockinson-Szapkiw et al., 2013; Mangen et al., 2013; Sage et al., 2019). *Interaction* with different reading support tools and milieus has been investigated to a lesser degree, yet with noteworthy exceptions, e.g. (Freund et al., 2016; Kol and Schcolnik, 2000; Brady et al., 2018; Wolfe, 2008; Johnston and Ferguson, 2020). 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 et al., 2018).

Meanwhile, large-scale studies of attitudes and preferences continue to conclude that students slightly prefer physical formats for focused academic reading, generally stating they feel like paper-based reading let them concentrate and remember better (Mizrachi et al., 2016; Pálsdóttir, 2019).

2.1 Active (academic) reading

Active reading was first described by Mortimer Adler in 1940 in the piece *How to Read a Book* (Adler and Van Doren, 2014). Active reading means reading while actively thinking and learning, and is often accompanied by interaction activities such as note-taking, highlighting and underlining the text (O'Hara, 1996). Pearson et al. list the main interaction features of active reading as *adding placeholders* in the text (temporary or permanent), *creating annotations* in the text, *taking notes* both in the text and on separate media, and *navigating* the text with the help of indexing (Pearson et al., 2013). Several studies indicate that it is beneficial for "secondary tasks" - such as annotation or navigation - to be as *minimally cognitively demanding as possible* in active reading contexts (Pearson et al., 2013; DeStefano and LeFevre, 2007).

Active reading is an example of how academic students read texts, though active reading is not limited to academic reading. We focus, in this paper, on academic reading as the type of reading that academic students perform of texts that they wish to memorize and learn from, according to the Remember-Know paradigm, which states that Knowledge which is *Remembered* is typically recalled 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 (Tulving, 1985; Conway et al., 1997).

2.2 Reading Support Tools

A reading support tool can be defined as any tool that can be used by people to read or support the reading of documents that primarily consist of written text. A reading support tool can can be analog or digital, and

it can be software or hardware. Hardware platforms, of course, need software to display a text.

Digital readers – such as ePub and PDF readers – are often not recognized as distinct tools, because they mainly display content, rather than support the reader actively, such as by facilitating active knowledge construction in the interaction with the tool (Freund et al., 2016; Sage et al., 2019). With the advent and spread of literature in digital formats, however, active reading support tools are in great demand (Pearson et al., 2013). Studies in this area are often not specific about the reading tool they are evaluating. Reading on an iPad with GoodReader may yield different reading performance and user experience than using Adobe Acrobat Reader on a Samsung Galaxy Tab, even though these could both be categorized as "tablet reading". There is a difference in evaluating the iPad versus the Samsung Galaxy, or evaluating GoodReader versus Adobe Acrobat. Providing clarity and distinctions between these tools is necessary for research to be comparable and findings to be widely applicable.

2.3 User experience of reading tools

User experience (UX) as a research agenda is concerned with studying the experience and use of technology in context. The UX of a product is a consequence of

> "a user's **internal state** (predispositions, expectations, needs, motivation, mood, etc.), the **characteristics of the designed system** (e.g. complexity, purpose, usability, functionality, etc.) and the **context** (or the environment) within which the interaction occurs (e.g. organisational/social setting, meaningfulness of the activity, voluntariness of use, etc.)" (Hassenzahl and Tractinsky, 2006), our emphases.

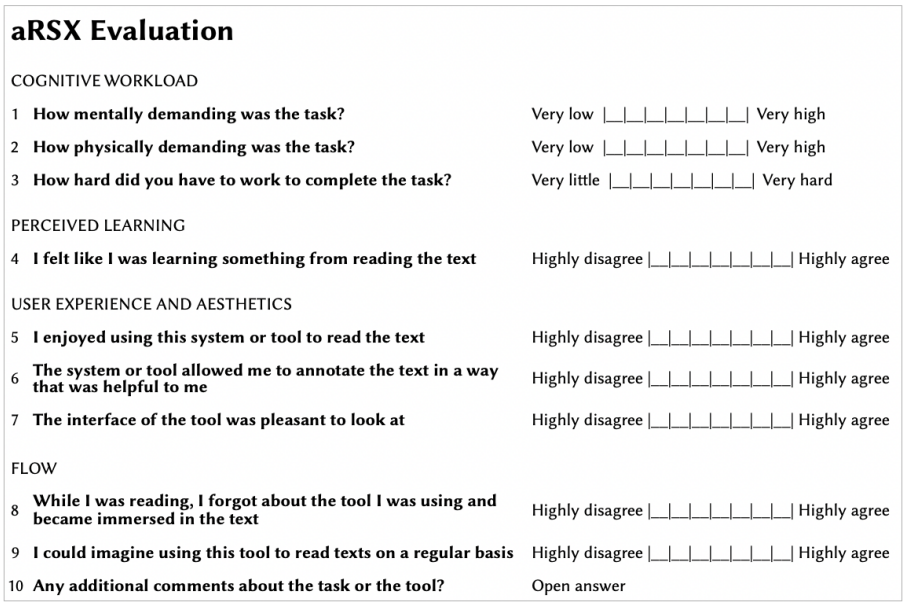
Models of UX usually separate a product's pragmatic from its hedonic qualities, where pragmatic attributes advance the user toward a specific goal and depend on whether the user sees a product as simple, predictable, and practical. Hedonic attributes, on the other hand, are related to whether users identify with a product or find it appealing or exhilarating (Hornbæk and Hertzum, 2017). Pragmatic attributes are often found to exert a stronger influence on the evaluation of a product than hedonic attributes.

Although text is presented linearly, learning by reading is not a linear process. Reading, and particularly academic reading, is open-ended. An academic reader depends on constant self-evaluation of whether the material is understood and internalized or not, rather than defined and well-known external objectives. According to Csikszentmihalyi's concept of *flow* in a learning context, student engagement is a consequence of simultaneous occurence of *concentration*, *enjoyment*, and *interest*. These states are readily related to the UX qualities of internal state, system characteristics, and context, and should be evaluated in relation to any reading support tool.

There are numerous ways of evaluating *usability* of products and tools, but none of the methods address the specific requirements of reading, such as whether the tool helps the student read, process and understand – in other words, whether the tool helps the student *know* the material. The UX goals of such tools are not efficiency or performance metrics, but rather that the user feels cognitively enabled to focus on the text content for as little or as much time as necessary (Pearson et al., 2013).

While qualitative research such as detailed interviews and observations are traditional methods for conducting UX evaluations, these methods are time-consuming and not easy to implement on a large scale. The goal of the aRSX is develop a quantitative, survey-based evaluation with a foundation in qualitative UX research. Quantitative surveys are non-costly and time-efficient to execute, and they have been used for decades as a valuable indicator of tool specifications and requirements (Hart and Staveland, 1988; Hart, 2006; Nielsen, 1995).

Generally, metric-based research investigating students' opinions and experiences of reading tools has consistently found correlation between interaction design and reading performance (Haddock et al., 2019; Léger et al., 2019; Freund et al., 2016; Lim and Hew, 2014; Zeng et al., 2016), and a positive correlation between user attitudes and learning outcomes (Kettanurak et al., 2001; Sage et al., 2016; Teo et al., 2003). However, few studies investigate specifically which features foster a good learning experience, although with some exceptions, e.g. (Pearson et al., 2013; Buchanan and Pearson, 2008; Chen et al., 2012; Pearson et al., 2012). One survey identified some of the most important themes for academic students when choosing between digital and paper as the following: Flexibility, ability to concentrate, ability to remember what was read, organizing, approachability and volume of the material, expenses, making notes, scribbling and highlighting, and technological advancement (Pálsdóttir, 2019). It is our research agenda to develop an evaluation survey which is founded in these and similar findings, and which evaluates the user experience design of any given reading support tool.



aRSX Evaluation

COGNITIVE WORKLOAD

1	**How mentally demanding was the task?**	Very low |__|__|__|__|__|__|__| Very high
2	**How physically demanding was the task?**	Very low |__|__|__|__|__|__|__| Very high
3	**How hard did you have to work to complete the task?**	Very little |__|__|__|__|__|__|__| Very hard

PERCEIVED LEARNING

4	**I felt like I was learning something from reading the text**	Highly disagree |__|__|__|__|__|__|__| Highly agree

USER EXPERIENCE AND AESTHETICS

5	**I enjoyed using this system or tool to read the text**	Highly disagree |__|__|__|__|__|__|__| Highly agree
6	**The system or tool allowed me to annotate the text in a way that was helpful to me**	Highly disagree |__|__|__|__|__|__|__| Highly agree
7	**The interface of the tool was pleasant to look at**	Highly disagree |__|__|__|__|__|__|__| Highly agree

FLOW

8	**While I was reading, I forgot about the tool I was using and became immersed in the text**	Highly disagree |__|__|__|__|__|__|__| Highly agree
9	**I could imagine using this tool to read texts on a regular basis**	Highly disagree |__|__|__|__|__|__|__| Highly agree
10	**Any additional comments about the task or the tool?**	Open answer

Figure 1: The first iteration of the aRSX evaluation. The questions are divided into categories addressing cognitive workload, perceived learning, user experience and aesthetics, and flow of studying, which are categories amalgamated from previous research findings. The form also has an open-ended question, which has shown to lead to very interesting thematic responses (Pálsdóttir, 2019), and serves as a way for us to become aware of salient themes for the students.

3 Methodology: Developing the aRSX

The first iteration of the aRSX (Figure 1) consists of nine Likert-scale questions and an open-ended prompt for additional comments. We included four categories of questions based on UX and active reading theory.

Cognitive workload. One of the main goals of a reading support tool is to minimize the cognitive effort required of the reader to interact with the tool itself, so they can focus completely on the content (Pearson et al., 2013; DeStefano and LeFevre, 2007). We therefore used the NASA Task Load Index (TLX) (Hart and Staveland, 1988) as inspiration for the survey format. The NASA TLX has been used for over 30 years as an evaluation method to obtain workload estimates from 'one or more operators' either while they perform a task or immediately afterwards. Other fields have had great success appropriating the TLX to evaluate task-specific tools, for instance, creativity support tools (Cherry and Latulipe, 2014). The first three questions of our evaluation form are copied from the TLX-questions concerning mental and physical demand – the latter potentially distracting from the content itself. The first iteration of the aRSX is designes as a "Raw TLX", eliminating the part of the original TLX which is concerned with weighting the individual questions to reflect personal importance attributed to each question. The Raw-TLX approach is simpler to employ, and does not yield less useful results (Hart, 2006). Future versions of the aRSX may include a weighted scoring part, in any instance, to yield insights about preferences across different populations.

Perceived learning. An academic reader depends on constant self-evaluation of whether the material is understood and internalized or not (Tulving, 1985; Conway et al., 1997). The survey should evaluate whether the tool generally lives up to the reader's expectation of pragmatic qualities, i.e. whether the tool helps them learn from the text. Self-evaluation is often used in learning research, and has been proven reliable (Sage et al., 2019; Paas et al., 2003), and the fourth question of the survey simply asks the reader whether they believed they learned from the reading.

User experience and aesthetics. As described in section 2.3, good user experience and interaction design have a positive correlation with learning outcomes. Although we expected readers to have higher *pragmatic* expectations of reading support tools than *hedonic* expectations, user enjoyment and aesthetics of the reading tool are important to the overall user experience and technology adaptation (Hornbæk and Hertzum, 2017). The fifth, sixth and seventh ques-

tion ask whether the reader enjoyed using the tool, whether it allowed them to annotate the text in a helpful way, and whether the tool was pleasant to look at.

Flow. While theory of learning is often concerned with the *content* of a given text, the aRSX focuses on the capacity of the *tool* to allow the student to process and engage with the text. The experience of flow while reading academic texts can occur as the result of a well-written, interesting or challenging text, but it can be enhanced or disrupted by *contextual factors* such as the tool used to consume the text (Shernoff et al., 2014; Pearson et al., 2013). By definitions by the IFLA Study Group (on the functional requirements for bibliographic records, 1998) we can say that the aRSX pertains to evaluating the *item* – the physical representation of the book – rather than the *work* or the *manifestation* i.e. the author's creation or a particular translation of that work. The last two questions of the aRSX address whether the tool was invisible while reading, and whether the tool seems appropriate to student's normal practices.

Finally, the aRSX includes an open-ended question which we used to uncover additional themes of the students' experience of the tool.

3.1 Criteria for a usable evaluation form

In order to evaluate the usefulness of the aRSX, we specified the following criteria as ideals for a survey evaluation:

1) *Theoretical foundation*: The evaluation should be grounded in prior research on reading support tools and user experience design.

2) *Operationlizability*: The evaluation should be operational and useful for researchers developing and evaluating reading support tools.

3) *Generalizability*: The evaluation must enable researchers to analyse different kinds of reading support tools with different types of populations in different types of settings.

4) *Comparability*: The framework must enable researchers to compare evaluations of different tools, also between studies.

5) *Reliability*: The survey should produce reliable results, aiming for a Cronbach's alpha above 0.7.

6) *Empirical grounding*: The framework must be tested in practice, ensuring that it measures the intended aspects and that it is clear and usable for both study participants and researchers.

In the study presented in this paper, we focused especially on developing empirical grounding through an early pilot study – testing the survey in an exploratory experiment. Theoretical and empirical grounding must also be developed through applying and evolving the evaluation form in different studies and communities. Through this paper we share the aRSX with other researchers and invite them to use, evaluate, and modify the evaluation form.

3.2 Testing the aRSX: Experimental Setup

3.2.1 Subjects and treatments

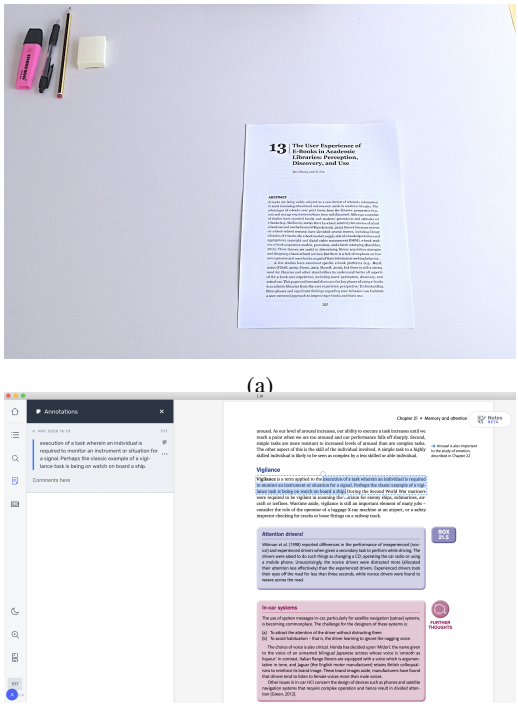
The first version of the aRSX was tested with 50 students at the IT University of Copenhagen, Denmark, during fall 2019. The students were from four different study programs, and primarily bachelor level students. The students demographics were: 27 male and 23 female, 22 on their 1st year, 18 on their 2nd year and 10 on their 3rd year or more of studies. The median age was 21 years old with the majority of subjects in the age group of 21-23 years old.

As compensation for their participation, all participants were gifted a semester's worth of free textbooks (digital or physical) of their own choosing.

The test setup was a controlled, *within-group* setup, where each student were subjected to both treatments; a paper reading treatment, and a digital reading treatment.

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 12pt Times New Roman with headlines in 14pt Arial. The text had 2cm 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 laptop. The text was a PDF file and formatted exactly as the paper reading for comparability. The text was provided in the software Lix, a reading support software for PDF readings. 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. A picture of the two treatments is shown in Figure 2.



(a)

(b)

Figure 2: The paper reading treatment (top) and digital reading treatment (bottom).

3.2.2 Experimental design and execution

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 ten pages, which you believe corresponds to the course level 2-3 months in the future". We performed tests of LIX scores (Björnsson, 1968) and Flesch–Kincaid readability tests (Kincaid et al., 1975) on each of the texts, and selected texts of similar reading difficulty. 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 avoid variation in scores as a result of varying reading content.

After reading the first half of the text, each student filled out the aRSX survey on paper for that treatment. They then read the other half of the text in the opposite medium, and filled out the aRSX on paper for that medium. To avoid adverse effects from information overload or fatigue, 26 of the students read the first half of the text on paper and the second half of the text on computer (condition A) and 24 students read the first half of the text on computer and the second half of the text on paper (condition B)[1]. The students in condition A and B were not in the same room. They were not informed about the focus on the evaluation form. The students 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".

3.2.3 Data gathering and analysis

Experimental observations. The experiment was run as an explorative experiment, where we were interested in discovering if the evaluation was generally meaningful for participants, and whether it yielded significant and useful results. The first type of data we gathered was *experimental observations*, primarily questions from participants about the wording of the survey or how to complete it. Those data did not need thorough analysis, as the questions we received were quite straightforward.

Quantitative data analysis. The second type of data we collected were the results of the aSRX, the purely quantitative data. Running studies with a number of participants large enough to yield statistical significance is not trivial, and the evaluation method should not depend on it (i.e. the aRSX should produce reliable small-scale results). We used the Cronbach's alpha as an indicator of internal reliability. To investigate statistical significance, we conducted Kolmogorov-Smirnov and Wilcoxon Signed-Rank tests, aiming for p-values below 0.05. The goal of these analyses was to investigate whether we could observe statistically significant within-group differences in the evaluation scores of two media, and whether a potential difference would be reliable according to classical test theory (DeVellis, 2006).

Qualitative data analysis. We designed the aRSX with a final open-ended question to discover salient themes that may not have been addressed by the questions of the survey. The question was optional, and we received 25 comments regarding the paper reading and 37 comments regarding the digital reading. The length of the comments varied from one to eight sentences. We clustered the comments into themes according to their contribution or evaluation focus, rather than necessarily the topic they addressed. For instance; *"I was missing some words to accompany the icons on the screen. It wasn't intuitive for me what*

[1]This slight unevenness in distribution of condition A and B was due to student availability at the time of the experiment.

icons meant what" (Participant 170001, digital reading) was coded as *#interface/ux design* rather than, for instance, 'icons' or 'UI design', because the focus of the study was to evaluate whether the aRSX addressed important themes in reading support, and not to evaluate the individual tool interface.

4 Findings

4.1 Experimental observations

The first field the evaluation page had an empty field asking the participant to fill in the 'tool' they were using, in this study referring to either the paper or the digital reading. The first observation we made during the experiment was that it was not clear to all participants what to put in this field – they, of course, did not know we were comparing "paper" and "digital" reading. This attests to the importance of clarifying which hardware or software is being evaluated. For future studies we would recommend pre-filling in that field for the participants.

Finding 1: "Annotating" is not an obvious concept. Several participants asked what was meant by 'annotating' in question 6: "The system or tool allowed me to annotate the text in a way that was helpful to me". This could be exacerbated by the fact that only few of the students were native English speakers, and the survey was conducted in English. In addition, we observed this theme in the open-ended survey responses (9 out of 37 participants commented on highlighting features), e.g.: *"It is very easy to highlight, but a little more confusing to make comments to the text"* (Participant 190802, digital reading). According to reading research, annotating a text consists of, for instance, highlighting, underlining, doodling, scribbling, and creating marginalia and notes (Marshall, 1997; Pearson et al., 2013). Construction of knowledge and meaning during reading happens through activities such these, making the possibility of annotation extremely important when supporting active reading. The question of annotation should be clarified.

Finding 2: "Interface" is a concept that works best for evaluation of digital tools. The word 'interface' in question 7: "The interface of the tool was pleasant to look at" prompted some questions in the paper treatment. An interface seems to be interpreted as a feature of a digital product, and this was not a useful term when evaluating an analog medium. In the interest of allowing the aRSX to be used in the evaluation of both digital and analog tools, this question should designate a more general description of the aesthetics of the tool.

4.2 Quantitative data

The quantitative results of the first iteration of the aRSX scale-based questions of the survey are shown in Table 1. We performed a Kolmogorow-Smirnov test of normal distribution, which showed general non-normal distributions, and low probability of results arising from chance alone. Five p-values pointed to a normal distribution.

Because of the non-normal distribution of results, we conducted a Wilcoxon Signed-Rank test. This showed that differences between the sum scores of the two tools (paper and laptop) were *not* statistically significant, with p-values ranging between 0.08 and 0.7. In line with our expectations based on previous research (e.g. (Mizrachi et al., 2016; Abuloum et al., 2019; Pálsdóttir, 2019) the paper-based reading condition was rated slightly more favorable on all parameters except for the physical strain of reading (Question 2, mean 2.18 for computer vs. 2.20 for paper) and the appearance of the interface (Question 7, mean 4.86 for paper vs. 5.34 for computer). The fact that the differences in scores between the two treatments were not statistically significant does not attest to the validity of the evaluation, rather, it is likely an accurate picture of students not having strong preferences for either paper or digital formats when reading shorter texts (Mizrachi et al., 2016; Pálsdóttir, 2019).

Finding 3: The survey appears to be internally reliable. We performed an ANOVA two-factor analysis without replication, and calculated a Cronbach's alpha of 0.7182, which indicates that the survey is likely to be internally reliable. Question one, two and three (pertaining to cognitive workload) ask the user to rate their mental and physical strain or challenge from 1 (Very low) to 7 (Very high). In these questions a high score corresponds to a negative experience, and the scores therefore had to be reversed to calculate sum score and Cronbach's alpha. Further tests are needed to investigate whether positively/negatively worded statements produce different results.

Finding 4: "Physical demand" should be specified. The average scores for question two (physical demand) are very low for both paper and Lix. The question is copied directly from the NASA TLX, and was deemed relevant because eye strain from digital

		PAPER			COMPUTER			COMPARISON	
	Question	**Mean**	σ	p	**Mean**	σ	p	z	p
	COGNITIVE WORKLOAD	3.36	1.02		3.49	1.14			
1	**How mentally demanding was the task?**	4.20	1.31	0.003	4.28	1.51	0.122	-0.377	0.704
2	**How physically demanding was the task?**	2.22	1.23	0.019	2.18	1.23	0.003	-0.222	0.826
3	**How hard did you have to work to complete the task?**	3.66	1.32	0.041	4.02	1.49	0.082	-1.680	0.093
	PERCEIVED LEARNING	5.3	—		5.08	—			
4	**I felt like I was learning something from reading the text**	5.3	1.25	0.026	5.08	1.50	0.014	-0.852	0.395
	USER EXPERIENCE AND AESTHETICS	4.98	0.11		4.87	0.41			
5	**I enjoyed using this system or tool to read the text**	5.02	1.46	0.032	4.58	1.76	0.016	-1.262	0.208
6	**The system or tool allowed me to annotate the text in a way that was helpful to me**	5.06	1.49	0.006	4.70	1.59	0.038	-0.892	0.373
7	**The interface of the tool was pleasant to look at**	4.86	1.63	0.061	5.34	1.69	0.005	-1.755	0.080
	FLOW	4.96	0.54		4.34	0.62			
8	**While I was reading, I forgot about the tool I was using and became immersed in the text**	4.58	2.04	0.225	3.90	1.78	0.167	-1.598	0.110
9	**I could imagine using this tool to read texts on a regular basis**	5.34	1.64	0.014	4.78	1.92	0.010	-1.385	0.168
	SUM SCORE (AVERAGE OF CATEGORIES)	**4.72**			**4.26**				

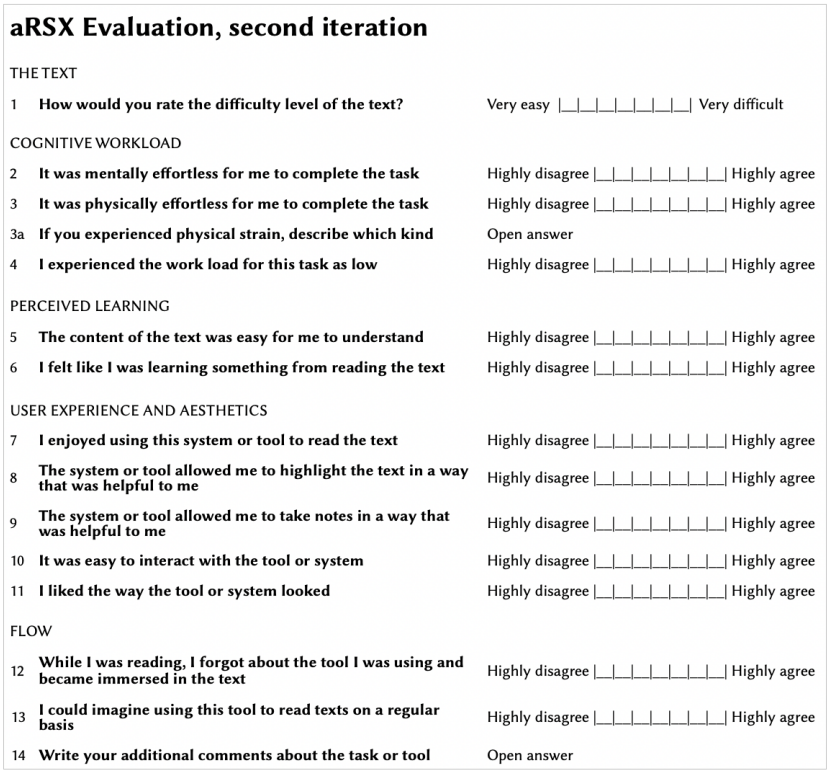
Table 1: aRSX results of Kolmogorov-Smirnov and Wilcoxon Signed-Rank calculations. Normal distributions within question scores are marked with green, and insignificant p-values of comparisons between the two tools are marked with red (all).

reading has often been mentioned as a negative factor of screen reading in previous research (e.g. (Sheppard and Wolffsohn, 2018)). 'Physical demand', however, may be associated with hard, physical labor, and should be specified further to gain useful knowledge from the score. This was exacerbated by some of the comments from the open-ended question: *"I think it would be better to do the test on my own computer. The computer was noisy and the screen was small"* (Participant 190211, digital reading), and *"Reading on a pc is not pleasant when the paper is white. Use some sort of solarized"* (Participant 170303, digital reading). These comments demonstrate that types of experienced physical strain can very a lot, and the question of physical demand does not, in itself, yield useful insights. Based on the low scores, the high p-value, and the comments on specific, physical difficulties, we speculate that participants may interpret the term 'physical demand' differently, and that the scores will be skewed because of this.

4.3 Qualitative data

The open-ended question responses generally showed that participants were aware of the evaluation setup, and that they were focused on evaluating the usability and experience of the software tool. The responses also showed that many of the students were willing to reflect on and compare the different tools in a meaningful way during the same setup or session. The responses were extremely valuable in elaborating the measured experience reflected in the quantitative measures, and we would recommend to keep this question in future iterations of the survey. In the interest of keeping the survey very light-weight and quick to use, we have not made the question very elaborate, or split it into more questions, although that could be considered in future uses of the survey.

Finding 5: The content may influence the evaluation of the tool. A theme in the comments which was not addressed by the questions in the aRSX was the content of the specific text which was read. Nine participants commented on the text e.g.: *"Really interesting text"*. (Participant 170001, digital reading) and *"The text was more of a refresher than new learning"* (Participant 150302, paper reading). Although it seemed from the comments like the students were able to distinguish the text from the tool, and some of these effect would be mitigated by the fact that the students read from the same text in both treatments, we believe it to be a relevant observation that the text which is being read may influence the experience of using a tool. Furthermore, the type of text may also require different tools for annotating, cf. *"When I tried to highlight mathematical formulas it would sometimes try to highlight additional text that I couldn't remove from the little highlight box"*. (Participant 180403, digital reading). For the next iteration of the aRSX, we suggest adding a question about the experienced difficulty of the text to allow transparency of potential effects of this.



aRSX Evaluation, second iteration

	THE TEXT	
1	**How would you rate the difficulty level of the text?**	Very easy |__|__|__|__|__|__|__| Very difficult
	COGNITIVE WORKLOAD	
2	**It was mentally effortless for me to complete the task**	Highly disagree |__|__|__|__|__|__|__| Highly agree
3	**It was physically effortless for me to complete the task**	Highly disagree |__|__|__|__|__|__|__| Highly agree
3a	**If you experienced physical strain, describe which kind**	Open answer
4	**I experienced the work load for this task as low**	Highly disagree |__|__|__|__|__|__|__| Highly agree
	PERCEIVED LEARNING	
5	**The content of the text was easy for me to understand**	Highly disagree |__|__|__|__|__|__|__| Highly agree
6	**I felt like I was learning something from reading the text**	Highly disagree |__|__|__|__|__|__|__| Highly agree
	USER EXPERIENCE AND AESTHETICS	
7	**I enjoyed using this system or tool to read the text**	Highly disagree |__|__|__|__|__|__|__| Highly agree
8	**The system or tool allowed me to highlight the text in a way that was helpful to me**	Highly disagree |__|__|__|__|__|__|__| Highly agree
9	**The system or tool allowed me to take notes in a way that was helpful to me**	Highly disagree |__|__|__|__|__|__|__| Highly agree
10	**It was easy to interact with the tool or system**	Highly disagree |__|__|__|__|__|__|__| Highly agree
11	**I liked the way the tool or system looked**	Highly disagree |__|__|__|__|__|__|__| Highly agree
	FLOW	
12	**While I was reading, I forgot about the tool I was using and became immersed in the text**	Highly disagree |__|__|__|__|__|__|__| Highly agree
13	**I could imagine using this tool to read texts on a regular basis**	Highly disagree |__|__|__|__|__|__|__| Highly agree
14	**Write your additional comments about the task or tool**	Open answer

Figure 3: The second iteration of the aRSX

5 Discussion: Evaluating the aRSX

Based on the findings of our experiments, we have rephrased some of the questions of the aRSX, as well as added several questions. The second iteration is shown in figure 3.

We have added an initial question pertaining the general difficulty level the text response to Finding 5: 'The difficulty of the text may influence the evaluation of the tool'. If participants experience the text as very difficult, this may impact their perception of the tool. Additionally, we added a question under Perceived Learning: "The content of the text was easy for me to understand". While the aRSX is not attempting to evaluate the quality of the text, this questing, together with the new question 1, acknowledges that there is an difference between evaluating a text as difficult, and experiencing difficulty reading it.

The questions regarding cognitive workload, previously 1-3 and now 2-4 have been rephrased as positive statements, cf. the observation in Finding 3, making it easier to calculate a sum score. Hopefully, this also avoids any potential confusion in decoding the scales for participants, who may inadvertently confuse a positive with a negative answer, if the scales vary.

Question 3: "It was physically effortless for me to complete the task" now has an added open question of "If you experienced physical strain, describe which kind". This has been added as a consequence of Finding 4: '"Physical demand" should be specified'.

We split the question about annotations into one question about the tool's ability to support highlights and to create notes cf. Finding 1: '"Annotating" is not an obvious concept'. We recognize that not all students may need to highlight or annotate the texts they read, in which case we anticipate the score would be neutral. Based on a few (less than three) of the responses to the open-ended question, we anticipate that participants may mention this themselves if they find it relevant: *"I normally don't take notes for texts like this. Usually I just read the text and take notes for the lectures. Therefore it is quite different for me"*. (Participant 190105, paper reading).

The word "interface" is removed from the survey in the interest of making it usable for evaluation of analog tools as well as digital, cf. Finding 2: '"Interface" is a concept that works best for evaluation of digital tools', and this question has been rephrased.

Overall, it was simple to use the aRSX as an evaluation method. We identified some possibilities for improvement, which have been integrated into the sec-

ond iteration. In this study, the aRSX was distributed on paper, which was simple in terms of execution, and a little more cumbersome in terms of digitizing the data - especially transcribing the open-ended question responses might be problematic with large participant numbers. The survey may also be distributed digitally, which we hypothesize could have a positive effect on the open-ended question responses, both due to the possibility of making the question mandatory, and because of the ease of writing comments on the computer versus in hand. In its simple form, the survey should be straightforward to moderate for other studies. The evaluation does thus fulfill the criterion of *Operationalizability*, as per section 3.1.

We believe the first iteration of the aRSX was well founded in theory, described in the criterion *theoretical foundation*. Neither reading support nor user experience design are new fields, and there is a solid foundation of knowledge on which to build the selection of good evaluation questions. The novelty consists primarily in developing a consistent, reflective practice around such evaluation, so that both developers and researchers can best benefit from the work of colleagues and peers.

In terms of *Generalizability* and *Comparability*, we need to conduct further studies of the aRSX in use, to assess the internal reliability of the questions and validity of results. Developing psychometric evaluation tools is not a trivial task, and we are looking forward to explore this avenue in depth.

The survey responses had a Cronbach's alpha of 0.718, which is a satisfying result. We will explore different terminology and phrasing in future iterations of the aRSX and aim for an *a* above 8.5.

Finally, we have achieved better *empirical grounding* of the aRSX in conducting the first study and evaluating its outcomes. The second iteration of the aRSX is more detailed, and will hopefully yield more valuable insights to researchers and developers of reading support tools. We of course invite peers and colleagues to use, moderate, and evolve the survey in their research, further extending the level of empirical grounding.

6 Conclusion and future work

In this paper, we presented the active Reading Support indeX (aRSX), an evaluation form designed to assess the reading support of a tool in active reading based on user experience and learning preferences. Currently, such evaluation happens ad hoc in research, and the evaluation methods vary from study to study, making it difficult to compare evaluations of different reading support tools across studies.

An initial test of the first iteration of the aRSX yielded valuable insights about the framing of the evaluation, as well as revealed themes of questions that should be included in the next iteration of the form.

We continue to use the aRSX to evaluate reading support tools and prototypes of different kinds. The second iteration of the aRSX will be tested with a focus on reliability and validity assessment metrics. Future work will include further experimentation with the same and different user groups, providing additional longitudinal comparative data with the same cohort group and several tools as well as different populations. So far, the aRSX has shown to provide meaningful data with a relatively small sample, and we believe the iterations suggested in this paper will make it a stronger evaluation tool.

We believe that the aRSX is a very promising avenue for evaluating reading support tools based on personal user experience, and we invite the research community to apply and appropriate the survey.

7 Acknowledgments

We thank the students who participated in this experiment. This research was funded by the Innovation Fund Denmark, grant 9066-00006B: Supporting Academic Reading with Digital Tools.

REFERENCES

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

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

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

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

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

Buchanan, G. and Pearson, J. (2008). Improving placeholders in digital documents. In *International Conference on Theory and Practice of Digital Libraries*, pages 1–12. Springer.

Chen, N., Guimbretiere, F., and Sellen, A. (2012). Designing a multi-slate reading environment to support active

reading activities. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3):1–35.

Cherry, E. and Latulipe, C. (2014). Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(4):21.

Conway, M. A., Gardiner, J. M., Perfect, T. J., Anderson, S. J., and 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.

Delgado, P., Vargas, C., Ackerman, R., and 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.

DeStefano, D. and LeFevre, J.-A. (2007). Cognitive load in hypertext reading: A review. *Computers in human behavior*, 23(3):1616–1641.

DeVellis, R. F. (2006). Classical test theory. *Medical care*, pages S50–S59.

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

Farinosi, M., Lim, C., and Roll, J. (2016). Book or screen, pen or keyboard? a cross-cultural sociological analysis of writing and reading habits basing on germany, italy and the uk. *Telematics and Informatics*, 33(2):410–421.

Freund, L., Kopak, R., and 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.

Haddock, G., Foad, C., Saul, V., Brown, W., and Thompson, R. (2019). The medium can influence the message: Print-based versus digital reading influences how people process different types of written information. *British Journal of Psychology*.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage Publications Sage CA: Los Angeles, CA.

Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier.

Hassenzahl, M. and Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology*, 25(2):91–97.

Hayles, N. K. (2012). *How we think: Digital media and contemporary technogenesis*. University of Chicago Press.

Hornbæk, K. and Hertzum, M. (2017). Technology acceptance and user experience: A review of the experiential component in hci. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):1–30.

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

Kettanurak, V. N., Ramamurthy, K., and Haseman, W. D. (2001). User attitude as a mediator of learning performance improvement in an interactive multimedia environment: An empirical investigation of the degree of interactivity and learning styles. *International Journal of Human-Computer Studies*, 54(4):541–583.

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

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

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

Léger, P.-M., An Nguyen, T., Charland, P., Sénécal, S., Lapierre, H. G., and Fredette, M. (2019). How learner experience and types of mobile applications influence performance: The case of digital annotation. *Computers in the Schools*, 36(2):83–104.

Lim, E.-L. and Hew, K. F. (2014). Students' perceptions of the usefulness of an e-book with annotative and sharing capabilities as a tool for learning: a case study. *Innovations in Education and Teaching International*, 51(1):34–45.

Mangen, A., Walgermo, B. R., and Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International journal of educational research*, 58:61–68.

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

Mizrachi, D., Boustany, J., Kurbanoğlu, S., Doğan, G., Todorova, T., and Vilar, P. (2016). The academic reading format international study (arfis): Investigating students around the world. In *European Conference on Information Literacy*, pages 215–227. Springer.

Nielsen, J. (1995). How to conduct a heuristic evaluation. *Nielsen Norman Group*, 1:1–8.

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

on the functional requirements for bibliographic records, I. S. G. (1998). Functional requirements for bibliographic records: final report.

Paas, F., Tuovinen, J. E., Tabbers, H., and Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1):63–71.

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

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

Pearson, J., Buchanan, G., Thimbleby, H., and Jones, M. (2012). The digital reading desk: A lightweight ap-

proach to digital note-taking. *Interacting with Computers*, 24(5):327–338.

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

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

Sage, K., Rausch, J., Quirk, A., and Halladay, L. (2016). Pacing, pixels, and paper: Flexibility in learning words from flashcards. *Journal Of Information Technology Education*, 15.

Sheppard, A. L. and Wolffsohn, J. S. (2018). Digital eye strain: prevalence, measurement and amelioration. *BMJ open ophthalmology*, 3(1):e000146.

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., and Shernoff, E. S. (2014). Student engagement in high school classrooms from the perspective of flow theory. In *Applications of flow in human development and education*, pages 475–494. Springer.

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

Teo, H.-H., Oh, L.-B., Liu, C., and Wei, K.-K. (2003). An empirical study of the effects of interactivity on web user attitude. *International journal of human-computer studies*, 58(3):281–305.

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

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.

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.

Zeng, Y., Bai, X., Xu, J., and He, C. G. 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.