Evaluating Academic Reading Support Tools: Developing the aRSX-Questionnaire

Nanna Inie[1][0000−0002−5375−9542] and Bjørn Hjorth Westh[1]

IT University of Copenhagen, Center for Computing Education Research, Copenhagen, Denmark nans@itu.dk, bjwe@itu.dk
www.itu.dk

Abstract. This paper presents and evaluates a new survey metric, the active Reading Support indeX (aRSX), which was created to help researchers and designers evaluate whether a specific software or hardware tool supports active reading of academic texts. The aRSX is comprised of questions in five categories: *The Text, Cognitive Workload, Physical Workload, Perceived Learning, User Experience and Aesthetics,* and *Flow*, as well as an open-ended question for additional comments. In this paper, we present our initial development of two beta-versions of the questionnaire in two studies of $n = 100$ and $n = 53$ deployments, evaluating paper, laptops, iPads, and reMarkable tablets as reading support tools. These studies led to the current version of the aRSX, and the initial results suggest that the metric is a reliable and valid indicator of a tool's ability to support reading of academic texts.

Keywords: Reading Support Tools · Digital reading · Evaluation metrics · User experience design · User experience methodology

1 Introduction

Reading academic literature with digital tools is becoming more and more of a normalcy for students, yet the diversity of digital reading support tools is surprisingly low [32, 24]. Digital textbooks or research papers are rarely designed to look different from their physical instances, and they often distributed as PDFs, a format which was developed primarily for printing, and which has the purpose of making a document look *stable* on all devices and in all printing conditions – not for accommodating different reading platforms or preferences. The hardware used to consume digital texts is largely confined to personal laptops and, less often, tablets [24, 32, 35]. As Pearson, Buchanan and Thimbleby claimed 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 [32].

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 a declining trajectory in reading abilities and motivation of

students from elementary school through college [15, 40, 30]. In one meta-study of research on digital versus physical reading, Delgado and colleagues [8] 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 complex text, which requires sustained attention [2, 8].

One promising avenue for supporting better digital reading is to study the particularities of the *user experience design* (UX) of digital tools (both hardware and software) – exploring the way these tools afford and support what is necessary for students to engage in *active reading* [1, 32, 29]. The interaction design of digital tools is particularly interesting to explore, because out of the factors 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 [29]), *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 [32].

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, just as 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 presents a novel evaluation metric for reading support tools. We present the beta versions 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 user experience and learning preferences when reading academic texts. In addition, such an evaluation tool can be used to indicate "robust moderating factors" that influence reading performance in different media [8].

In the paper, we evaluate two beta versions of the aRSX and present the current version of the questionnaire. Although the form is in continuous development, we find it useful to share our early experiences with utilizing such a survey metric. This paper explores the question: *How might we evaluate academic reading support tools with a focus on user experience?*, and provides initial findings that the aRSX can give reliable, robust evaluations of the user experience of different digital reading support tools.

2 Background and related work

Surveys of attitudes and preferences continue to conclude that college students slightly prefer physical formats for focused academic reading, generally stating they feel like paper-based reading let them concentrate and remember better [25, 29]. Research in Human-Computer Interaction (HCI) has suggested that especially annotation and note-taking on a computer are activities that compete with the reading itself, due to the lack of direct manipulation [19]. The cognitive workload required to interact with digital devices may be higher than that of interaction with physical media, and annotating introduces another cognitive workload [33, 32]. Indeed, studies have indicated that the simplest textual and interactive environment is associated with the highest comprehension outcomes and better user experience [3, 33, 10].

It is a noteworthy challenge for the development of good reading support software and hardware that few existing studies of digital reading specify the reading environment provided to participants, and even fewer evaluate its interaction design in detail. This means that it is difficult to identify the UX designs that influence the reading experience and performance in a positive way; and thus, difficult to improve the design of novel digital tools for academic reading in a rigorous, systematic manner [18].

2.1 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 reading software – programs for displaying ePub and PDF formats – is often not recognized as a specialized tool, because it mainly displays content, rather than support the reader actively. With the advent and spread of literature in digital formats, however, active reading support tools are in great demand [32]. 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 reading on 'tablets'. 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.

2.2 User experience of reading tools

There are numerous ways of evaluating *usability* of products and tools (such as walkthroughs, think-aloud tests, contextual inquiry, etc.), but a tool's ability to support reading of academic texts is more complex than its usability. The UX goals of reading tools are not efficiency or performance metrics, but rather, that the user feels cognitively enabled to focus on a text content for as little or as

much time as necessary [32], that the user feels enabled to process the text in any way they might need, as well as whether the tool is a good "fit" for the student's preferences and reading context.

User experience as a research agenda is concerned with studying the experience and use of technology in context. The UX of a product is considered 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.)" [14], 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 [16]. 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.

Generally, metric-based research investigating students' opinions and experiences of reading tools has consistently found positive correlation between interaction design and reading performance [11, 21, 10, 22, 43], and between user attitudes and learning outcomes [20, 36, 41]. However, few studies investigate which features in particular foster a good learning experience, although with some exceptions, e.g. [32, 3, 5, 33].

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 [29]. Out of these, the flexibility of the reading support tool, the approachability of the content, the affordances of annotation, and the technological advancement are factors which are (to at least some degree) influenced by the design of the reading support tool.

2.3 Questionnaire-based UX evaluation

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, questionnaire-based evaluation with a foundation in UX research. Quantitative questionnaires are non-costly and time-efficient to execute, and

they have been used for decades as a valuable indicator of tool specifications and requirements [13, 12, 27]. They also have the advantage that most HCI researchers and designers are already familiar with the format and the reporting of its results.

The NASA Task Load indeX (TLX) [13] 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. The TLX consists of six subscales: Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance. All questions are on a 21-point scale of "Very Low" to "Very High." The assumption is that a combination of these dimensions represents the workload experienced by most people performing most tasks. Each subscale is furthermore "weighted" by the individual performing the task, so that the final score is given based on how important each subscale is to the individual. The NASA TLX has been translated into more than a dozen languages, and modified in a variety of ways [12]. It is being used as a benchmark of evaluation, and has proved its value in a wide range of fields from nuclear power plant control rooms to website design.

Other fields have had great success appropriating the TLX to evaluate task-specific tools, for instance *creativity support* [4, 6]. The Creativity Support Index (CSI) is based on the NASA TLX, and is a psychometric survey designed to assess the ability of a digital creativity support tool to support the creative process of its users. Its structure is very similar to the NASA Task Load Index, but its theoretical foundation is based on concepts from creativity and cognition support tools, such as creative exploration, theories of play, Csiksentmihalyi's theory of flow, and design principles for creativity support tools. The CSI includes six subscales or "factors": Collaboration, Enjoyment, Exploration, Expressiveness, Immersion, and Results Worth Effort.

While both the TLX and the CSI are incredibly valuable tools, the surveys do not address the particularity of learning from reading. Inspired by the CSI, we decided to create an evaluation form tailored to uncover the ability of a tool to support textual knowledge acquisition based on theory of learning and UX research.

3 Methodology

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 the questionnaire:

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

2) *Operationlizability*: The evaluation should be operational and useful for researchers and designers developing and evaluating reading support tools. It should be clear and usable for both participants and those who administer the evaluation.

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

4) *Validity*: The survey should accurately measure the factors that it intends to measure.

5) *Reliability*: The survey should produce reliable results, aiming for a Cronbach's alpha above .70.

6) *Empirical grounding*: The framework must be thoroughly tested in practice.

In the study presented in this paper, we focused especially on developing the *theoretical foundation*, *validity*, *reliability*, and *empirical grounding*. The operationalizability and generalizability are somewhat inherent in the original NASA TLX survey, and we borrow some credibility from this thoroughly tested metric.

Further theoretical and empirical grounding must 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 Experimental setups

The first beta-version of the aRSX was created and tested in *Study 1: Laptop and paper reading*. The findings from this study led to the second beta-version, which was much longer than the first. It is common practice in psychometrics to create longer, temporary versions of a survey when developing a new metric. This allowed us to conduct an exploratory factor analyses of the responses, in the interest of identifying the items or questions that performed the best. The second beta-version was tested in *Study 2: iPad and reMarkable reading*. Following this section, we will describe the studies and findings chronologically.

Both studies were designed as controlled within-group studies, where we invited a group of students to read half a text on one medium, asked them to evaluate it, then switched to a different medium for the second half of the text, and asked the participants to evaluate the second medium. That means that 77 students filled out a total of 153 evaluations (one student only completed a reading in one medium), 100 of the first beta-version, and 53 of the second beta-version. The within-group comparison of two reading tools per participant allowed us to explore the aRSX with four different media, as well as to conduct four different reliability analyses of Cronbach's alpha (one per reading tool). Although thorough development of psychometric evaluation forms require hundreds and sometimes thousands of participant numbers for statistically sound analyses to be conducted, we find it valuable to share our initial findings at an early state of development, both to document the development process of the aRSX openly, as well as to share the beta-versions of the questionnaire with the research community for feedback and comments.

The beta-versions of the aRSX were designed as "Raw TLX", eliminating the part of the original TLX which is concerned with pairwise ranking of the subscales to reflect personal importance attributed to each subscale or factor.

The raw-TLX approach is simpler to employ, and does not appear to yield less useful results [12], but more importantly, we wished to thoroughly develop and evaluate subscales to explore which questions and factors carried higher loading. When the question wordings and factors are more resolved, it may make sense to add a pairwise factor rating to the survey, as in the TLX and the CSI [13, 6]. Consequently, no cumulative or final score for each tool was calculated for the beta versions of the aRSX, as the final score is traditionally dependent on each participant's ranking of the different factors.

4 Study 1: Laptop and paper reading

4.1 First beta-version subscales: Cognitive Workload, Physical Workload, Perceived Learning, User Experience and Aesthetics, and Flow

The first beta-version of the aRSX is shown in figure 1. It had 10 basic questions.

COGNITIVE & PHYSICAL WORKLOAD		
1	**How mentally demanding was the task?**	Very low I__I__I__I__I__I__I__I Very high
2	**How physically demanding was the task?**	Very low I__I__I__I__I__I__I__I Very high
3	**How hard did you have to work to complete the task?**	Very little I__I__I__I__I__I__I__I Very hard
PERCEIVED LEARNING		
4	**I felt like I was learning something from reading the text**	Highly disagree I__I__I__I__I__I__I__I Highly agree
USER EXPERIENCE AND AESTHETICS		
5	**I enjoyed using this system or tool to read the text**	Highly disagree I__I__I__I__I__I__I__I Highly agree
6	**The system or tool allowed me to annotate the text in a way that was helpful to me**	Highly disagree I__I__I__I__I__I__I__I Highly agree
7	**The interface of the tool was pleasant to look at**	Highly disagree I__I__I__I__I__I__I__I Highly agree
FLOW		
8	**While I was reading, I forgot about the tool I was using and became immersed in the text**	Highly disagree I__I__I__I__I__I__I__I Highly agree
9	**I could imagine using this tool to read texts on a regular basis**	Highly disagree I__I__I__I__I__I__I__I Highly agree
10	**Any additional comments about the task or the tool?**	Open answer

Fig. 1. The first beta-version of the aRSX, from [17].

Cognitive and Physical Workload. The first three questions were copied from the TLX-questions concerning mental and physical demand. In the first beta-version, the cognitive and physical workload were collected under one subscale, as we did not anticipate the physical workload to play a significant role to the evaluation, other than as a potentially distracting factor to the cognitive workload.

One of the most basic issues in the study of cognitive workload is the problem of how to actually measure it, but since we are concerned with the user *experience* of cognitive workload, it is appropriate to let the user self-evaluate this aspect. The cognitive attention required to read and annotate on paper is minimal for most people, which often means that people can do it without thinking [34]. 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 [32, 9].

Perceived learning. The questionnaire should evaluate whether the tool allows for learning from reading the text. This aspect is an evaluation of the tool's *pragmatic* qualities, i.e. whether it advances a user toward their specific goal [16]. An academic reader depends on constant self-evaluation of whether the material is understood and internalized or not [42, 7]. Self-evaluation is often used in learning research, and has been proven reliable [35, 28]. The fourth question of the survey simply asked the reader whether they believed they learned from the reading.

User experience and aesthetics. As described in section 2.2, 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 extremely important to the overall user experience and technology adaptation [16, 29]. The fifth and seventh question asked whether the reader enjoyed using the tool and whether the tool was pleasant to look at. Since the questionnaire was designed for evaluating *active* reading [1], it should evaluate how the tool supports annotation of the text. When engrossed in academic reading, many users will accompany reading with annotating; for example, highlighting, commenting and underlining. Annotations can be considered a by-product of the active reading process, and they should be supported by digital reading tools [31]. Therefore, the sixth question asked whether the tool allowed annotation in a form that was helpful to the user.

Flow. According to Csikszentmihalyi's concept of *flow* in a learning context, student engagement is a consequence of simultaneous occurence of *concentration*, *enjoyment*, and *interest* - of high challenge in combination with high skill [38]. 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 reader to process and engage with the text. The experience of flow can happen to individuals who are deeply engrossed in activity which is intrinsically enjoyable, and the activity is perceived as worth doing for its one sake, even if no further goal is achieved [26]. 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 [39, 32]. Question eight and nine of the aRSX addressed whether the tool fosters immersion while reading, and whether the tool would fit into the reader's regular practices.

Finally, the questionnaire finished with an open-ended question, so we could learn about other factors that may have been relevant to the reader during the first evaluation of the questionnaire.

4.2 Participants and execution

Subjects and treatments 50 students from the IT University of Copenhagen, Denmark, were recruited during fall 2019. The test setup was a controlled, *within-group* setup, where each student was subjected to both treatments; a paper reading treatment, and a digital reading treatment. The study is described in further detail in Inie and Barkhuus [17].

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.

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.

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. 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".

4.3 Findings from Study 1

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.

The second type of data from the study were the **quantitative results** of the evaluations. We performed Cronbach's alpha tests on the responses for reliability.

The third type of data was the **qualitative responses** to the final open-ended question. The question was optional, and we received 25 comments regarding the paper reading and 37 comments regarding the digital reading. The

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

length of the comments varied from one to eight sentences. We clustered the comments into themes according to their relevance to the questionnaire, rather than their distinct evaluation of the tool.

Experimental observations

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 [23, 32]. Construction of knowledge and meaning during reading happens through activities such as 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.

Quantitative results

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 .732 for paper, and .851 for laptop, which indicates satisfying reliability. 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) were very low for both paper and laptop. The question is copied directly from the NASA TLX, and was deemed relevant because eye strain from digital reading has often been mentioned as a negative factor of screen reading in previous research (e.g. [37]). '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.

Qualitative responses 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.

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).

5 Study 2: iPad and reMarkable reading

5.1 Second beta-version subscales: The Text, Cognitive Workload, and Physical Workload

The second beta-version of the aRSX is shown in Figure 2. Based on the findings of Study 1, we rephrased some of the questions of the aRSX, as well as added several questions. For this beta-verison, we wanted to conduct a factor analysis to determine optimal statement wording, and we therefore split many of the questions into several options.

We 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

THE TEXT		
0	**How would you rate the difficulty level of the text?**	Very easy I__I__I__I__I__I__I__I Very difficult
COGNITIVE WORKLOAD		
1a	**It was mentally effortless for me to complete the task**	Highly disagree I__I__I__I__I__I__I__I Highly agree
1b	**Reading the text took a lot of mental exertion**	Highly disagree I__I__I__I__I__I__I__I Highly agree
1c	**My brain did a lot of work during this task**	Highly disagree I__I__I__I__I__I__I__I Highly agree
1d	**I experienced the work load for this task as low**	Highly disagree I__I__I__I__I__I__I__I Highly agree
PHYSICAL WORKLOAD		
2	**It was physically effortless for me to complete the task**	Highly disagree I__I__I__I__I__I__I__I Highly agree
2a	**If you experienced physical strain, describe which kind**	Open answer
2b	**I felt physically tired/uncomfortable during reading**	Highly disagree I__I__I__I__I__I__I__I Highly agree
PERCEIVED LEARNING		
3a	**The content of the text was easy for me to understand**	Highly disagree I__I__I__I__I__I__I__I Highly agree
3b	**I felt like I was learning something from reading the text**	Highly disagree I__I__I__I__I__I__I__I Highly agree
3c	**I did not acquire a lot of information from this text**	Highly disagree I__I__I__I__I__I__I__I Highly agree
USER EXPERIENCE AND AESTHETICS		
4a	**I enjoyed using this system or tool to read the text**	Highly disagree I__I__I__I__I__I__I__I Highly agree
4b	**I found it fun to use this tool or system to read the text**	Highly disagree I__I__I__I__I__I__I__I Highly agree
4c	**This tool was annoying to use for reading**	Highly disagree I__I__I__I__I__I__I__I Highly agree
5a	**It was easy to interact with the tool or system**	Highly disagree I__I__I__I__I__I__I__I Highly agree
5b	**The functions of this tool or system were difficult to use**	Highly disagree I__I__I__I__I__I__I__I Highly agree
5c	**It was straightforward to use the tool or system in the way I wanted to**	Highly disagree I__I__I__I__I__I__I__I Highly agree
6a	**I liked the way the tool or system looked**	Highly disagree I__I__I__I__I__I__I__I Highly agree
6b	**The tool or system was attractive or beautiful**	Highly disagree I__I__I__I__I__I__I__I Highly agree
6c	**The tool or system was ugly or boring**	Highly disagree I__I__I__I__I__I__I__I Highly agree
Int.a	**The system or tool allowed me to highlight the text in a way that was helpful to me**	Highly disagree I__I__I__I__I__I__I__I Highly agree
Int.b	**The system or tool allowed me to take notes in a way that was helpful to me**	Highly disagree I__I__I__I__I__I__I__I Highly agree
FLOW		
Flow	**While I was reading, I forgot about the tool I was using and became immersed in the text**	Highly disagree I__I__I__I__I__I__I__I Highly agree
7a	**I could imagine using this tool to read texts on a regular basis**	Highly disagree I__I__I__I__I__I__I__I Highly agree
7b	**I would like to own this tool to use for reading more often**	Highly disagree I__I__I__I__I__I__I__I Highly agree
7c	**I think this tool fits my reading requirements**	Highly disagree I__I__I__I__I__I__I__I Highly agree
8	**Write your additional comments about the task or tool**	Open answer

Fig. 2. The second beta-version of the aRSX.

their perception of the tool. While the aRSX is not designed to evaluate the quality of the text, this question acknowledges that there is an difference between evaluating a text as difficult, and experiencing difficulty reading it.

Cognitive and Physical Workload was split into two subscales, and we added an open question to understand the Physical Workload further: "If you experienced physical strain, describe which kind". This was added in response to Finding 4: '"Physical demand" should be specified'.

Furthermore, we split the question about annotations into two questions, one about the tool's ability to support highlights and one about the support of creating 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 anticipated the scores would be neutral.

Finally, the word "interface" was removed from the questionnaire 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'.

In the second beta-version, we primarily focused on analyzing quantitative data from the responses. The qualitative responses to the open-ended questions fell in two categories; they elaborated either a participant's individual state, i.e. "I felt very tired due to lack of sleep", or the answers to the evaluation, i.e. "I normally don't highlight, so that was not relevant to me". The answers indicated that the task of assessing the tool's reading support ability was straightforward to understand to the participants.

5.2 Participants and execution

Subjects and treatments 27 students at the IT University of Copenhagen, Denmark, were recruited to participate in the second study. It took place in fall 2021. Like Study 1, the setup for Study 2 was a controlled, within-group setup, where each student were subjected to both treatments; an iPad reading treatment, and a reMarkable[2] reading treatment. In the **iPad reading** treatment, students were provided with the text in PDF Expert, as well as an Apple pencil for annotations. In the **reMarkable** reading treatment, students were provided with the text in reMarkable's own PDF reading software and the reMarkable pen. The text was a PDF file and separated into two halves. 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. 16 students read on the iPad first and the reMarkable second, and 11 students read on the reMarkable first, and the iPad second. The students were not informed about the focus on the evaluation form, but were told to evaluate the tool they were using. 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".

[2] https://remarkable.com/

5.3 Findings from Study 2

The focus on Study 2 was to investigate optimal wording for the individual questions. We opted to identify two to three items per subscale, so that each subscale or factor can be tested for reliability in future deployments. This beta-version of the aRSX contained 26 individual items, clustered around very similar questions to those in the first beta-version (although the questions were shuffled around randomly in the actual deployment, rather than organized in their respective subscales), as well as an open-ended question at the end. The open-ended question was rephrased in this version to an imperative "Write your additional comments ...", and this wording yielded much more detailed qualitative responses to the evaluations. 20 out of 26 (one student only read in on the reMarkable, and not the iPad) participants wrote comments for this question for the iPad, and 23 out of 27 participants wrote comments for the reMarkable, all comments of at least three sentences or more.

The overall averages of the evaluations are shown in Figure 3. We conducted an exploratory factor analysis for the responses, and calculated Cronbach's alpha for the responses for iPad versus reMarkable, respectively. We also calculated Cronbach's alpha for each subscale.

Reliability The overall Cronbach's alpha for the questionnaire was .919 for the iPad evaluation, and .904 for the reMarkable evaluation, indicating an extremely high internal reliability. This is likely a consequence of the many questions, and also indicates that we can safely remove some questions and likely still achieve a high reliability.

Factor analysis As we were interested in the factor loadings for each question, we present the rotated component matrix in Figure 4 (KMO .790, $p < .001$). We can see that the questions cluster in 5 different components, which actually *almost* match the number of subscales (6), and that they are somewhat matched to their subscales: questions 7a, 7b, and 7c are all in cluster 1, question 1a, 1b, 1c, and 1d are all in cluster 2, and so forth.

With this analysis, we are interested in *eliminating* questions, so that we identify the two items that are most likely to represent the underlying factor. We see that question 7c and 2b load heavily (above .40) in more than one cluster, so we would like to eliminate those. Both 5a, 5b, and 5c also load in two clusters, but 5c loads lowest in both clusters, so we would also like to eliminate that. Using this process – identifying the questions that load highest in *one* category alone, we were left with questions 0, 1a, 1b, 2a, 3b, 3c, 5a, 5b, 6a, 6b, 7a, 7b, and the two questions about interaction. The questions about the text (0) and about flow are kept as they are, as they have not been divided into subquestions.

Post item-removal reliability After eliminating some questions, we re-calculated the reliability for the whole questionnaire, as well as for each individual factor. The results are shown in Table 1. We see that the survey is still highly internally reliable, both in total, and on each subscale level.

	iPad	**reMarkable**
The Text	N/A	N/A
Cognitive Workload	.837	.872
Physical Workload	N/A	N/A
Perceived Learning	.752	.670
UX & Aesthetics	.905	.790
Flow	.874	.874
Total	**.883**	**.861**

Table 1. Cronbach's alpha for the adjusted survey. The scales "The Text" and "Physical Workload" are not calculated, as there is only one question in these categories.

6 Discussion and current version of the aRSX

The current version of the aRSX is shown in Figure 5. A main change is that the subscale User Experience and Aesthetics has been split into two: Interaction Design and Aesthetics, in an attempt to distinguish between pragmatic and hedonic qualities of the reading support tool. The 16 questions are based on findings from Study 1 and Study 2, and should provide a good basic evaluation of a given reading support tool.

Overall, it was simple to use the aRSX as an evaluation method. We identified some possibilities for improvement, which have been integrated into the current iteration. In our studies, 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 current iteration of the aRSX is 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.

The aRSX is simple to deploy and analyze for any designer, developer or researcher, and the current iteration should present a question wording which is general enough to be applicable in a multitude of contexts. Building upon findings from well-established frameworks, such as the NASA TLX and the CSI, the current state of the survey is *generalizable*.

In this paper we presented an initial analysis of the survey's *validity*. We recognize that further studies and higher participant numbers are necessary to make rigorous claims about the survey's validity.

The survey responses had high reliability scores, which satisfy the criterion *Reliability*. All deployments of the survey had a general Cronbach's alpha above .70. This criterion will require more test-retest and split-half studies to confirm.

Finally, we have achieved better *empirical grounding* of the aRSX in conducting the first two studies and evaluating their outcomes.

7 Conclusion

In this paper, we presented two beta-versions of the aRSX, a novel survey metric designed to evaluate the ability of a tool to support active reading. We tested the two beta versions in two different studies, one focused on defining the questions to ask in such an evaluation, and one focused on defining the most salient factors to evaluate. Future work will include further deployment of the current version of the aRSX in different contexts and with different users. 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 user experience, and we invite the research community to apply and appropriate the survey.

8 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

1. Adler, M.J., Van Doren, C.: How to read a book: The classic guide to intelligent reading. Simon and Schuster (2014)
2. Annisette, L.E., Lafreniere, K.D.: Social media, texting, and personality: A test of the shallowing hypothesis. Personality and Individual Differences **115**, 154–158 (2017)
3. Buchanan, G., Pearson, J.: Improving placeholders in digital documents. In: International Conference on Theory and Practice of Digital Libraries. pp. 1–12. Springer (2008)
4. Carroll, E.A., Latulipe, C., Fung, R., Terry, M.: Creativity factor evaluation: towards a standardized survey metric for creativity support. In: Proceedings of the seventh ACM conference on Creativity and cognition. pp. 127–136 (2009)
5. Chen, N., Guimbretiere, F., Sellen, A.: Designing a multi-slate reading environment to support active reading activities. ACM Transactions on Computer-Human Interaction (TOCHI) **19**(3), 1–35 (2012)

6. Cherry, E., Latulipe, C.: Quantifying the creativity support of digital tools through the creativity support index. ACM Transactions on Computer-Human Interaction (TOCHI) **21**(4), 21 (2014)
7. Conway, M.A., Gardiner, J.M., Perfect, T.J., Anderson, S.J., Cohen, G.M.: Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. Journal of Experimental Psychology: General **126**(4), 393 (1997)
8. Delgado, P., Vargas, C., Ackerman, R., Salmerón, L.: 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)
9. DeStefano, D., LeFevre, J.A.: Cognitive load in hypertext reading: A review. Computers in human behavior **23**(3), 1616–1641 (2007)
10. Freund, L., Kopak, R., O'Brien, H.: The effects of textual environment on reading comprehension: Implications for searching as learning. Journal of Information Science **42**(1), 79–93 (2016)
11. Haddock, G., Foad, C., Saul, V., Brown, W., Thompson, R.: The medium can influence the message: Print-based versus digital reading influences how people process different types of written information. British Journal of Psychology (2019)
12. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 50, pp. 904–908. Sage Publications Sage CA: Los Angeles, CA (2006)
13. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. In: Advances in psychology, vol. 52, pp. 139–183. Elsevier (1988)
14. Hassenzahl, M., Tractinsky, N.: User experience-a research agenda. Behaviour & information technology **25**(2), 91–97 (2006)
15. Hayles, N.K.: How we think: Digital media and contemporary technogenesis. University of Chicago Press (2012)
16. Hornbæk, K., Hertzum, M.: Technology acceptance and user experience: A review of the experiential component in hci. ACM Transactions on Computer-Human Interaction (TOCHI) **24**(5), 1–30 (2017)
17. Inie, N., Barkhuus, L.: Developing evaluation metrics for active reading support. In: CSEDU (1). pp. 177–188 (2021)
18. Inie, N., Barkhuus, L., Brabrand, C.: Interacting with academic readings—a comparison of paper and laptop. Social Sciences & Humanities Open **4**(1), 100226 (2021)
19. Kawase, R., Herder, E., Nejdl, W.: A comparison of paper-based and online annotations in the workplace. In: European Conference on Technology Enhanced Learning. pp. 240–253. Springer (2009)
20. Kettanurak, V.N., Ramamurthy, K., Haseman, W.D.: 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 (2001)
21. Léger, P.M., An Nguyen, T., Charland, P., Sénécal, S., Lapierre, H.G., Fredette, M.: How learner experience and types of mobile applications influence performance: The case of digital annotation. Computers in the Schools **36**(2), 83–104 (2019)
22. Lim, E.L., Hew, K.F.: 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 (2014)
23. Marshall, C.C.: Annotation: from paper books to the digital library. In: Proceedings of the second ACM international conference on Digital libraries. pp. 131–140 (1997)

24. Mizrachi, D., Boustany, J., Kurbanoğlu, S., Doğan, G., Todorova, T., Vilar, P.: The academic reading format international study (arfis): Investigating students around the world. In: European Conference on Information Literacy. pp. 215–227. Springer (2016)
25. Mizrachi, D., Salaz, A.M., Kurbanoglu, S., Boustany, J., Group, A.R.: Academic reading format preferences and behaviors among university students worldwide: A comparative survey analysis. PloS one **13**(5), e0197444 (2018)
26. Nakamura, J., Csikszentmihalyi, M.: Flow theory and research. Handbook of positive psychology pp. 195–206 (2009)
27. Nielsen, J.: How to conduct a heuristic evaluation. Nielsen Norman Group **1**, 1–8 (1995)
28. Paas, F., Tuovinen, J.E., Tabbers, H., Van Gerven, P.W.: Cognitive load measurement as a means to advance cognitive load theory. Educational psychologist **38**(1), 63–71 (2003)
29. Pálsdóttir, Á.: 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 (2019)
30. Parsons, A.W., Parsons, S.A., Malloy, J.A., Gambrell, L.B., Marinak, B.A., Reutzel, D.R., Applegate, M.D., Applegate, A.J., Fawson, P.C.: Upper elementary students' motivation to read fiction and nonfiction. The Elementary School Journal **118**(3), 505–523 (2018)
31. Pearson, J., Buchanan, G., Thimbleby, H.: Hci design principles for ereaders. In: Proceedings of the third workshop on Research advances in large digital book repositories and complementary media. pp. 15–24 (2010)
32. Pearson, J., Buchanan, G., Thimbleby, H.: Designing for digital reading. Synthesis lectures on information concepts, retrieval, and Services **5**(4), 1–135 (2013)
33. Pearson, J., Buchanan, G., Thimbleby, H., Jones, M.: The digital reading desk: A lightweight approach to digital note-taking. Interacting with Computers **24**(5), 327–338 (2012)
34. Pearson, J.S.: Investigating lightweight interaction for active reading in digital documents. Swansea University (United Kingdom) (2012)
35. Sage, K., Augustine, H., Shand, H., Bakner, K., Rayne, S.: Reading from print, computer, and tablet: Equivalent learning in the digital age. Education and Information Technologies **24**(4), 2477–2502 (2019)
36. Sage, K., Rausch, J., Quirk, A., Halladay, L.: Pacing, pixels, and paper: Flexibility in learning words from flashcards. Journal Of Information Technology Education **15** (2016)
37. Sheppard, A.L., Wolffsohn, J.S.: Digital eye strain: prevalence, measurement and amelioration. BMJ open ophthalmology **3**(1), e000146 (2018)
38. Shernoff, D.J., Csikszentmihalyi, M.: Cultivating engaged learners and optimal learning environments. Handbook of positive psychology in schools **131**, 145 (2009)
39. Shernoff, D.J., Csikszentmihalyi, M., Schneider, B., Shernoff, E.S.: Student engagement in high school classrooms from the perspective of flow theory. In: Applications of flow in human development and education, pp. 475–494. Springer (2014)
40. Spichtig, A.N., Hiebert, E.H., Vorstius, C., Pascoe, J.P., David Pearson, P., Radach, R.: The decline of comprehension-based silent reading efficiency in the united states: A comparison of current data with performance in 1960. Reading Research Quarterly **51**(2), 239–259 (2016)
41. Teo, H.H., Oh, L.B., Liu, C., Wei, K.K.: An empirical study of the effects of interactivity on web user attitude. International journal of human-computer studies **58**(3), 281–305 (2003)

42. Tulving, E.: Memory and consciousness. Canadian Psychology/Psychologie canadienne **26**(1), 1 (1985)
43. Zeng, Y., Bai, X., Xu, J., He, C.G.H.: 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)

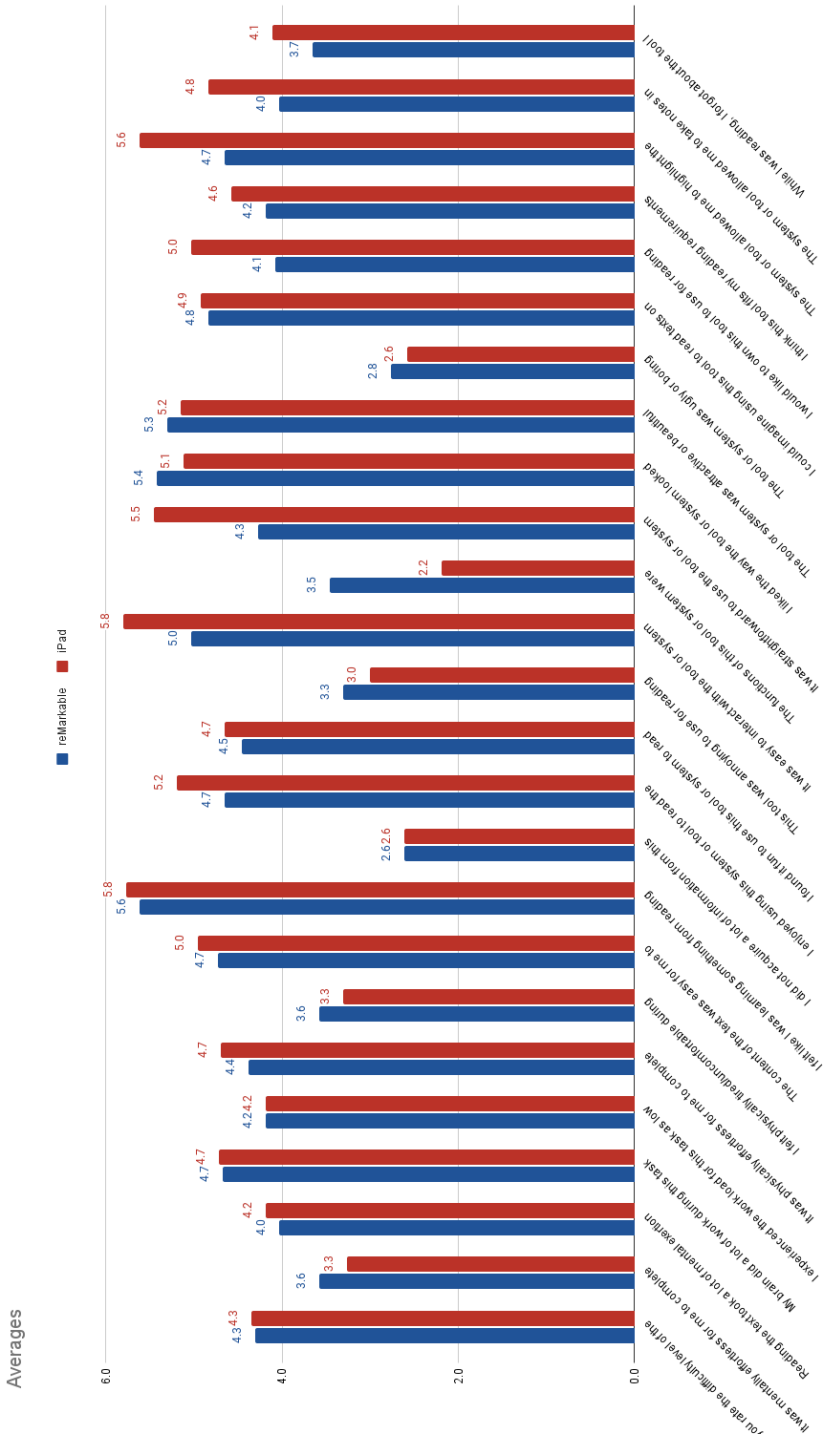

Fig. 3. Average scores for the iPad and the reMarkable.

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
4b	.908				
7b	.887				
7a	.839				
4a	.837				
7c	.798				.422
6a	.752				
4c	.751				
6b	.705				
Flow	.682			.528	
6c	.665				
1a		.828			
@0		-.826			
1b		.826			
1d		.714			
3a		.707			
1c		.697			
Interaction			.801		
5a	.497		.673		
5b	.454		.563		
5c	.417		.533		
Interaction			.511		
3b				.805	
3c				.784	
2b		.456			.627
2a					.621

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
[a]

a. Rotation converged in 9 iterations.

Fig. 4. Rotated component metrix showing factor loadings for all questions.

THE TEXT		
1	**How would you rate the difficulty level of the text?**	Very easy I__I__I__I__I__I__I Very difficult
COGNITIVE WORKLOAD		
2	**It was mentally effortless for me to complete the task**	Highly disagree I__I__I__I__I__I__I Highly agree
3	**Reading the text took a lot of mental exertion**	Highly disagree I__I__I__I__I__I__I Highly agree
PHYSICAL WORKLOAD		
4	**It was physically effortless for me to complete the task**	Highly disagree I__I__I__I__I__I__I Highly agree
4a	**If you experienced physical strain, describe which kind**	Open answer
PERCEIVED LEARNING		
5	**I felt like I was learning something from reading the text**	Highly disagree I__I__I__I__I__I__I Highly agree
6	**I did not acquire a lot of information from this text**	Highly disagree I__I__I__I__I__I__I Highly agree
INTERACTION DESIGN		
7	**It was easy to interact with the tool or system**	Highly disagree I__I__I__I__I__I__I Highly agree
8	**The functions of this tool or system were difficult to use**	Highly disagree I__I__I__I__I__I__I Highly agree
9	**The system or tool allowed me to highlight the text in a way that was helpful to me**	Highly disagree I__I__I__I__I__I__I Highly agree
10	**The system or tool allowed me to take notes in a way that was helpful to me**	Highly disagree I__I__I__I__I__I__I Highly agree
AESTHETICS		
11	**I liked the way the tool or system looked**	Highly disagree I__I__I__I__I__I__I Highly agree
12	**The tool or system was attractive or beautiful**	Highly disagree I__I__I__I__I__I__I Highly agree
FLOW		
13	**While I was reading, I forgot about the tool I was using and became immersed in the text**	Highly disagree I__I__I__I__I__I__I Highly agree
14	**I could imagine using this tool to read texts on a regular basis**	Highly disagree I__I__I__I__I__I__I Highly agree
15	**I would like to own this tool to use for reading more often**	Highly disagree I__I__I__I__I__I__I Highly agree
16	**Write your additional comments about the task or tool**	Open answer

Fig. 5. The current version of the aRSX.