CrowdHiLite: A Peer Review Service to Support Serious Reading on the Screen

Nan Jiang
Bournemouth University
Fern Barrow, Poole, BH12 5BB, UK
njiang@bournemouth.ac.uk

Huseyin Dogan
Bournemouth University
Fern Barrow, Poole, BH12 5BB, UK
hdogan@bournemouth.ac.uk

The advent of smart devices and consumerisation of IT has produced a significant and permanent shift away from print-based reading to digital reading. This, in turn, has changed people's reading behaviours and suggests that adapted mechanisms should be considered to support digital reading. It is particularly important for novice readers who need to read in-depth scientific literature in their chosen field. In this paper, we propose CrowdHiLite, a peer review service that allows expert readers to provide suggestion on individual readers' highlights to support their reading through the use of crowdsourcing technique. A demonstration was also provided to show how it would work in real world. A preliminary experiment comparing novice readers' reading performance with expert-rated highlights and normal highlights on the same document found improved reading efficiency and comprehension with the former.

Crowdsourcing. Communitysourcing. Reading Comprehension. Reading Speed. Collaborative Reading.

1. INTRODUCTION

The wide use of digital devices and Internet over the past two decades has produced a significant and permanent shift away from print-based reading to digital reading. This, in turn, has changed peoples' reading patterns. In general, reader spend more time on browsing and scanning, keyword spotting, one-time reading and selective reading (Liu, 2006; Rowlands et al., 2008; Hillesund, 2010). Despite new technologies such as e-ink, touch screen and high definition screen being introduced to provide better digital reading experience, a few studies comparing reading pixels and prints found that current devices still lack the functionality required for supporting some serious reading tasks (Aaltonen et al., 2011; Siegenthaler et al., 2010 & 2011). This confirmed Noyes and Garland's finding (2008) that the total equivalence in computer- and paper-based reading tasks is not possible to achieve despite the development in computer technology today. To this end, Tashman and Edwards (2011a) suggest that it needs an adapted mechanism to support digital reading rather than purely mimicking the affordances of paper in a computer-based system. One approach is to develop innovative systems that support the digital reading process through changing the way a reader interacts with digital documents (Tashman and Edwards, 2011b; Chen et al., 2012). Another approach is to direct readers to the most relevant regions on the page by using highlight enabled interfaces when they are reading (Boguraev et al., 1998; Graham, 1999; Chi et al., 2007).

In this paper, we propose CrowdHiLite, a crowdsourcing based peer review service that provides novice readers insights when reading scientific literature through a collaborative highlighting practice with expert readers. The service can be used in conjunction with any digital reading program to convert user highlights in a text stored in the program into highlight rating tasks and send them to expert readers for feedback and gather the feedback results to suggest the users whether they should pay attention to the highlights when they are reading. We conducted a preliminary experiment based on the service implementation and found improved reading efficiency and comprehension among novice readers when reading a document with expert rated highlights. The paper proceeds by discussing the design approach in Section 2, followed by a service demonstration in Section 3 and experiment design in Section 4. Results and discussions are provided in Section 5 and 6, followed by conclusion in Section 7.

2. DESIGN APPROACH

Our main design goal was to create a service that can be used to get expert feedback (ratings) on individual user highlights and use the feedback to support more readers when practising serious

*© The Authors. Published by
BCS Learning and Development Ltd.
Proceedings of HCI 2014, Southport, UK*

reading tasks on the screen. In order to achieve this, it needs to (1) define expert readers; (2) acquire personal highlights from individual users of a digital reading program and (3) reuse these highlights for obtaining expert feedback. Here we define an expert reader as an individual working in the relevant field who is not only experienced in reading scientific literature but also has some *specific* domain knowledge. For example, a lecturer or researcher in the field of computer science who has some knowledge in the HCI domain should be able to understand an HCI article even though they are not specialised in this area. Moreover, collecting highlights from individual readers and syncing these highlights over a number of devices to benefit more users has been already a popular feature in many modern reading programs (e.g., "LiveMinutes" with Evernotes, "Popular Highlights" in Amazon Kindle). Therefore, the core task was to (1) identify a cost effective mechanism to support this rating practice and (2) propose a universal solution that can be used with any digital reading program.

2.1 Crowdsourcing as a channel

Crowdsourcing is a flexible and cost-effective technique to get tedious work done by subdividing it into small Human Intelligent Tasks (HITs) and soliciting contributions from a large group of people (Estellés-Arolas and González-Ladrón-de-Guevara, 2013). Although most crowdsourcing applications focus on generic work that need contributions from the general public (e.g., Kickstarter and Crowdfynd), recent research is exploiting how it can be practised to support more advanced tasks that require expert knowledge through approaching a certain community of workers. For example, Manohar and Roy (2013) proposed a novel assessment method to enhance teachers' productivity using crowdsourcing technique. Zhang et al. (2013) proposed a solution that draws on the crowds in the CHI community to help schedule large-scale conferences. These studies suggest that crowdsourcing could be considered as an ideal channel for reading-related tasks. In fact, a similar study has been done and received some good initial results (Chircop et al., 2013).

2.2 Highlight rating as HIT

Since the main role of expert readers in our proposal was to provide feedback on individual users' highlights in a text and suggest whether novice readers should pay attention to these highlights, the HIT should be highlight rating tasks. Note that in order to rate the importance of a highlight, the related text paragraph where the highlight belongs to must be provided at the same time. In terms of the feasibility, this, in nature, is similar to common data annotation tasks where workers are asked to annotate and tagging a video, text or image based on their comprehension (Nowak and Rüger, 2010; Welinder and Perona; 2010). Considering the complexity level, it is also moderate as more complex text reading tasks have already been used in other problem domains and received positive results (Kittur et al., 2008).

2.3 CrowdHiLite as the solution

The service architecture of CrowdHiLite is shown in Fig. 1 where the numbers are used to indicate the work flow. First, collecting user highlight (Step 1 and 2) is often natively supported by a reading program. Once this is done, the program can call the CrowdHiLite service to convert its user highlights along with the related text paragraphs into HITs (Step 3 – 4). After that, the service will distribute the HITs to expert readers for completion and collect results from them (Step 5 – 7). Then the reading program can access the service again to use these results to filter its user highlights (Step 8) so as to provide reading suggestions to its users (Step 9 – 10). In this crowdsourcing based architecture, the *crowdsourcer* is the reading program (on behalf of its users), the *workers* are expert readers and the *HITs* are highlight rating tasks.

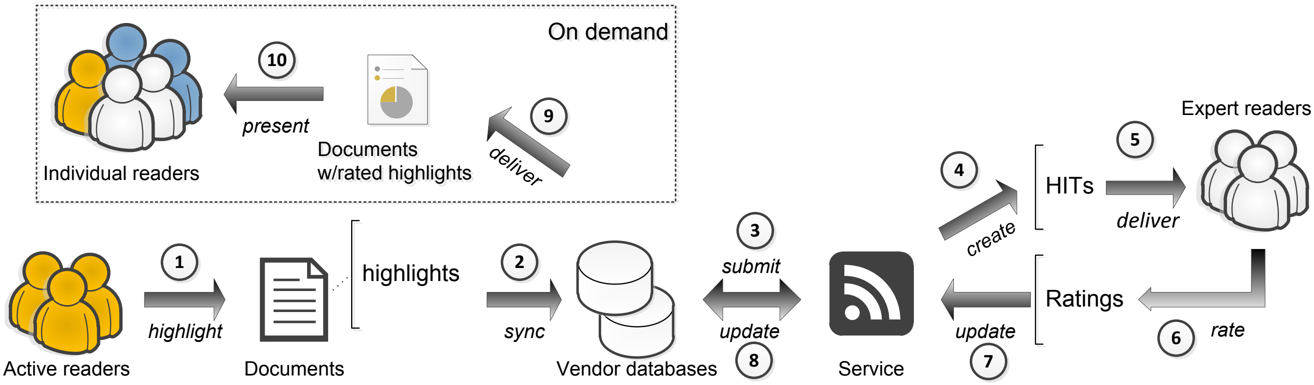
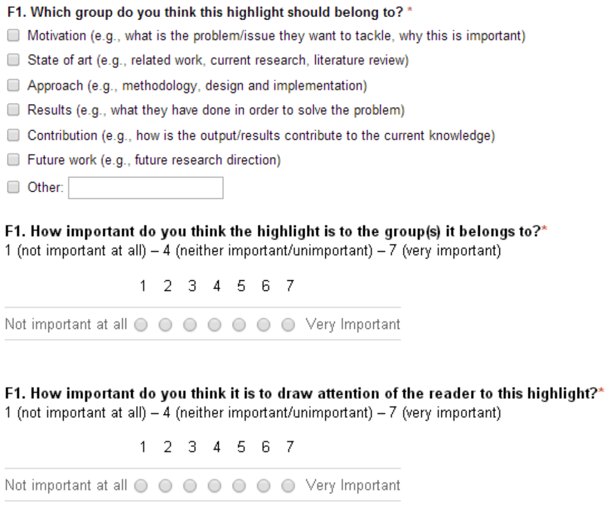


Figure 1: *Service architecture of CrowdHiLite*

3. IMPLEMENTATION

We created a desktop PDF document reader, an online repository and a highlight rating template to demonstrate the use of CrowdHiLite in real world. The reader, which was used to simulate a real document reader, supports highlighting, annotating, highlight sync and update to/from the online repository (Step 1 to 2 and 9 to 10 in Fig. 1). The online repository, which was used to simulate the central database of the reader (vendor databases in Fig. 1), supports highlight and related text paragraph storage (Step 2, 8) and retrieval (Step 3, 9). The highlight rating template (in Excel format) which was used to generate highlight rating HITs, consists of three predefined questions as shown in Fig. 2.



F1. Which group do you think this highlight should belong to? *

□ Motivation (e.g., what is the problem/issue they want to tackle, why this is important)
□ State of art (e.g., related work, current research, literature review)
□ Approach (e.g., methodology, design and implementation)
□ Results (e.g., what they have done in order to solve the problem)
□ Contribution (e.g., how is the output/results contribute to the current knowledge)
□ Future work (e.g., future research direction)
□ Other:

F1. How important do you think the highlight is to the group(s) it belongs to?*
1 (not important at all) – 4 (neither important/unimportant) – 7 (very important)

	1	2	3	4	5	6	7	
Not important at all	○	○	○	○	○	○	○	Very Important

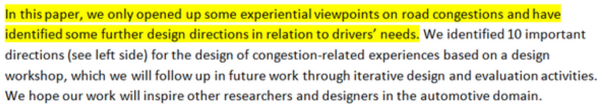
F1. How important do you think it is to draw attention of the reader to this highlight?*
1 (not important at all) – 4 (neither important/unimportant) – 7 (very important)

	1	2	3	4	5	6	7	
Not important at all	○	○	○	○	○	○	○	Very Important

Figure 2: *Questions in the template*

Question 1 was about classifying a user highlight based on the main topic areas of an academic paper. This was used to help expert readers to better understand the purpose of the highlight rating from a novice reader's point of view. Question 2 and 3 focused on the relevancy and importance of the highlight to its related text paragraph where an expert reader will be asked to rate the highlight based on a 7-point Likert scale ranging from "1: not important at all" to "7: very important".

We then created an instance of CrowdHiLite to (1) retrieve user highlights and related text paragraphs from the online repository using Amazon SimpleDB service; (2) convert them into HITs based on the template using Google Drive API and (3) release the HIT in the form of an online survey using Google Form Service API. Since each HIT is based on a highlight and its related text paragraph, the highlight and the text paragraph will be converted into an image as shown in Fig. 3 and embedded into the survey form presenting the HIT.



In this paper, we only opened up some experiential viewpoints on road congestions and have identified some further design directions in relation to drivers' needs. We identified 10 important directions (see left side) for the design of congestion-related experiences based on a design workshop, which we will follow up in future work through iterative design and evaluation activities. We hope our work will inspire other researchers and designers in the automotive domain.

Figure 3: *Image generated from a highlight and its related text paragraph*

4. EXPERIMENT SETUP

As mentioned in Section 1, several studies have already noted that participants read faster using highlight-enabled interfaces than non-highlight versions with improved comprehension. It is relatively more important to understand whether expert-rated highlights are more effective than the 'normal' collection of personal highlights. Therefore, we decided to run a comparative study on the same document with two versions of highlights: one with expert rated highlights (Version 1) and the other with normal aggregated but unrated highlights (Version 2).

4.1 Reading subject

We randomly selected a CHI Work-In-Progress (WIP) paper (Obrist et al., 2013) as the reading subject. The paper contains 3,144 words in 6 pages. The main body (excluding acknowledgement and reference) is 5 pages long with 79 sentences in 14 paragraphs.

4.2 Collecting highlights from individuals

Six MSc IT students (non-dyslexia) were asked to read the paper as if they were reading it for understanding all aspects of the work presented in the paper (e.g., motivation, state of art, approach, results, contribution and future work). They were asked to read the document naturally by using the PDF reader we developed so that they could highlight text and make annotations as usual. Note highlights were set to sentence level to minimise overlapping and duplicate issues. A total of 55 non-repeated highlighted sentences were collected in the end.

4.3 Crowdsourcing highlight rating tasks

55 HITs were created from the 55 non-repeated highlights collected from above readers. These HITs were distributed to all academic staff at the Department of Computing, Bournemouth University. Nine lecturers and researchers in the department responded to the HITs and completed them within 11.35 minutes on average.

4.4 Constructing two versions for comparison

Version 1 contained all expert rated highlights with an average of 5 and above ratings (7-point Likert

Scale ranging from 1: not important at all to 7: very important) in both group importance and reading attention question. Version 2 contained individual highlights from MSc students with three and more occurrences. In other words, a highlight will be displayed in Version 2 only if it has been made by over a half of students. The highlight inclusion criteria for both Version 1 and 2 were used to make sure these visible highlights reflect the common understanding of students and experts to minimise appropriateness and relevancy risks (Silvers and Kreiner, 1997). As a result, Version 1 contained 18 highlights and Version 2 contained 30 highlights where there were 14 highlights in common.

4.5 Reading comprehension test

We presented the two versions of paper in Adobe Acrobat reader to 35 non-dyslexia, Level C (1^{st} Year) computing student volunteers where 16 and 19 volunteers were assigned to Version 1 and 2 respectively. They were asked to read the paper in order to answer five reading comprehension questions. They were also told not to look at the paper again when answering questions to ensure they fully understood the paper before proceeding any further. Their reading speed was recorded by using an on-screen timer to count the elapsed time between the time when they opened the PDF and the time when they closed our PDF reader program. There was an observer in the same room to ensure participants did not break the rule. It should also be noted that these participants did not know the difference between the two versions before coming to the test.

5. RESULTS

Participants' reading performance on the two versions was compared based on three aspects: reading speed, reading comprehension and general user feedback. The first aspect was captured by recoding completion time; the second aspect was examined through using a comprehension test and the last was studied based on three semi open questions.

5.1 Reading speed

Participants who read Version 1 (expert rated version) performed marginally faster (M = 14.75 min, SD: 5.0398) than those who read Version 2 (aggregated version) (M = 14.947 min, SD: 4.0889). There was no significant difference between the two groups (Mann-Whitney U test, $p > 0.05$).

5.2 Reading comprehension

There were 5 multiple choice type comprehension questions covering the key aspects presented in the paper including (1) the research problem; (2) the state of art; (3) the approach, (4) the conclusion and (5) the future work. These questions were set by two domain experts. The results were measured by taking the proportion of participants who provided correct answers. As shown in Fig. 4, the proportion of participants who answered correctly after reading Version 1 was higher than those reading Version 2 especially for Q2 (69% vs. 30%) and Q5 (38% vs. 10%).

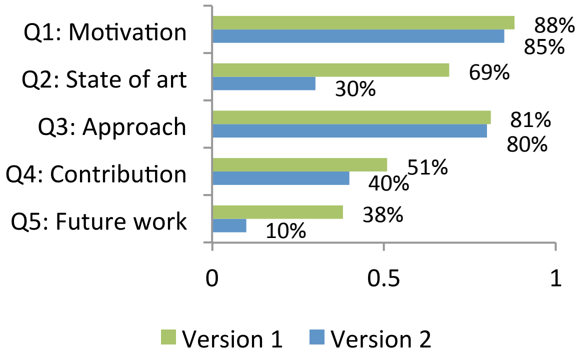


Figure 4: *Reading comprehension test results*

5.3 General user feedback

We were interested in finding out whether (1) highlights were helpful for improving reading comprehension; (2) readers felt confident with highlights displayed in the text and (3) readers felt distracted with highlights displayed in the text. A 7-point Likert scale was used in these three questions where 1 indicates "strongly disagree" and 7 indicates "strongly agree". Fig. 5 shows positive (5 and above), neutral (4) and negative responses (3 and below). In general, most participants acknowledged the presence of highlights in the text when reading but participants who read Version 2 (aggregated) felt more positive and less negative than those who read Version 1 (expert rated) in all questions. Note that for distraction, "disagree to distraction" indicates positive reading experience while "agree" to distraction indicates negative reading experience.

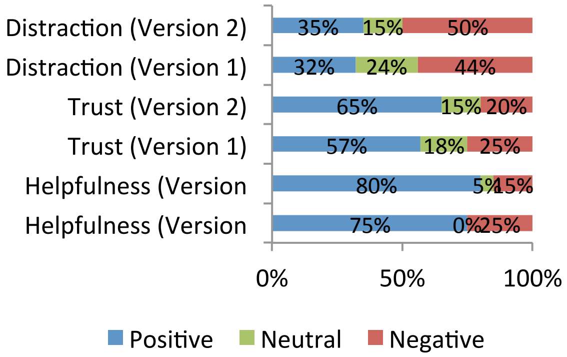


***Figure 5:** General user feedback results*

6. DISCUSSION

The results show that participants reading Version 1 performed generally better than those reading Version 2 in terms of both reading speed and comprehension. In particular, the latter has been improved noticeably with Version 1. However, it should be noted that there was no significant difference found between Version 1 and Version 2 in term of reading speed. This may suggest that the reading speed will not be affected by the number and filtering of the highlights if they come from the same source (i.e., same collection of individual reader highlights). Eye-tracking study could be followed to further investigate this. Moreover, since between-group design was used for the experiment, some variables such as participants' reading skills, group size and thresholds for highlight selection might have impact on the results, too. In that case, within-subject design could be used in the future to minimise the effect of such variables with more documents. Nevertheless, the results can be seen positive when considering the fact that participants' comprehension has been improved noticeably with Version 1 without affecting reading speed.

Although reading comprehension improvement on Version 1 has been reported when compared to Version 2, participants who read Version 2 felt more confident, helped and less distracted than those who read Version 1. Since the number of highlights in Version 2 was almost doubled compared to Version 1, a correlation between the number of highlights and the novice reader preference may exist. Evidence could be seen when participants were asked to provide any thought about their reading experience in an open question. 6 participants reading Version 1 said they would need to read highlights carefully in order to justify whether they were truly relevant or important as they seemed to be "random" or "irrelevant". In comparison, only 3 participants reading Version 2 had similar view. The difference between Version 1 and Version 2 was not revealed to the participants and there was no annotation provided for the text highlights during the experiment. Therefore, novice readers' confidence level might be affected by the number of highlights presented in a text due to their limited reading skills and domain knowledge. In other words, their confidence might be improved if the program could explicitly state how the highlights were derived and why the experts would like them to draw attention to the highlights. Again, this would need to be validated in future developments and experiments.

In addition, for this experiment, we sent the rating tasks to our colleagues in the same department directly and received good responses. However, in real world, sourcing experts and motivating them to complete these tasks can be challenging (Doan et al., 2011). This might be solved by specifying the purpose of such tasks and application domain (e.g., supporting students who study online or supporting students' self-learning).

7. CONCLUSION AND FUTURE WORK

In this paper, we presented CrowdHiLite, a crowdsourcing based peer review service designed for supporting novice readers when performing serious reading tasks on the screen by allowing expert readers to provide suggestion on their highlights. Using crowdsourcing technique to streamline this peer review process in a cost effective way also offers a prospective opportunity for automating large-scale highlight rating practice combined with supervising learning methods (Brew et al., 2010). A demonstration was provided to validate our idea and show how it can be realised in a real world. Positive results about improved reading efficiency and reading comprehension with expert rated highlights were also reported in the preliminary experiment. However, this approach still needs to be further investigated and consolidated with more experiments featuring large sample sizes and focusing on different types of academic papers across heterogeneous domains.

8. ACKNOWLEDGEMENT

We thank all staff and students for taking part in this research. This research is partially supported by Bournemouth University Fusion Grant: "Gamifying Enjoyable Learning Experiences on Social Networks".

9. REFERENCES

Aaltonen, M., Mannonen, P., Nieminen, S. & Nieminen M. (2011) Usability and compatibility of e-book readers in an academic environment: A collaborative study. IFLA Journal, 37(1), 16-27.

Boguraev, B., Kennedy, C., Bellamy, R., Brawer, S., Wong, Y. Y. & Swartz J. (1998) Dynamic presentation of document content for rapid on-line skimming. In Proc. AAAI Symposium on Intelligent Text Summarization, 118-128.

Brew, A., Greene, D. & Cunningham P. (2010) The Interaction Between Supervised Learning and Crowdsourcing', Computational Social Science and the Wisdom of Crowds. In 24th Annual Conference on Neural Information Processing Systems (NIPS 2010), Vancouver, Canada.

Chen, N., Guimbretiere, F., & Sellen, A. (2012) Designing a multi-slate reading environment to support active reading activities. ACM Transactions on Computer-Human Interaction (TOCHI), 19(3), 18.

Chi, E. H., Gumbrecht, M. & Hong L. (2007) Visual Foraging of Highlighted Text: An Eye-Tracking Study. In Proc. HCI'07, Beijing, China, 22 – 27 July 2007, 589-598. Springer-Verlag, Berlin, Heidelberg.

Chircop, L., Radhakrishnan, J., Selener, L. & J. Chiu (2013) Markitup: Crowdsourced Collaborative Reading. In CHI 2013 Extended Abstracts, Paris, France, 27 April – 2 May 2013, 2567 - 2572. ACM.

Doan, A., Ramakrishnan, R. & Halevy, A. Y. (2011) Crowdsourcing systems on the World-Wide Web. Communications of the ACM, 54(4), 86 - 96.

Estelles-Arolas, E., Gonzalez-Ladron-de-Guevara, F. (2012) Towards an integrated crowdsourcing definition. Journal of Information Science, 38(2), 189 - 200.

Graham, J. (1999) The reader's helper: a personalized document reading environment. In Proceedings of the SIGCHI conference on human factors in computing systems, Pittsburgh, Pennsylvania, USA, May 15-20, 1999, 481-488. ACM.

Hillesund, T. (2010) Digital reading spaces: How expert readers handle books, the Web and electronic paper. First Monday, 15(4).

Kittur, A., Chi, E. H., & Suh, B. (2008) Crowdsourcing user studies with Mechanical Turk. In Proceedings of the SIGCHI conference on human factors in computing systems (CHI '08), Florence, Italy, 5 – 10 April, 2008, 453-456. ACM.

Liu, M. (2006) Print vs. electronic resources: A study of user perceptions, preferences, and use. Information Processing & Management, 42(2), 583–592.

Nowak, S., Rüger, S. (2010) How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In Proceedings of the International Conference on Multimedia Information Retrieval (MIR 2010), Philadelphia, USA, 29 – 31 March, 2010, 557-566. ACM.

Obrist, M., Wurhofer, D., Krischkowsky, A., Karapanos, E., Wilfinger, D. H., Perterer, N. & Tscheligi, M. (2013) Experiential perspectives on road congestions. In CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13), Paris, France, 27 April – 2 May 2013, 535 – 540. ACM.

Manohar, P., Roy, S. (2013) Crowd, the Teaching Assistant: Educational Assessment Crowdsourcing. In First AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2013), Palm Springs, California, USA, 7 – 9 November 2013, 50 – 51.

Morris, M. R., Brush, A. B., & Meyers, B. R. (2007) Reading revisited: Evaluating the usability of digital display surfaces for active reading tasks. In Second IEEE International Workshop on Horizontal Interactive Human-Computer Systems Tabletop 2007 (TABLETOP '07), Newport, Rhode Island, USA, 10 – 12 October 2007, 79 – 86.

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

Rowlands, I., Nicholas, D., Williams, P., Huntington, P., Fieldhouse, M., Gunter, B., Withey, R. Jamali, H. R., Dobrowolski, T. & Tenopir C. (2008) The Google generation: The information behaviour of the researcher of the future. Aslib Proceedings, 60(4), 290–310.

Siegenthaler, E., Wurtz, P., Bergamin, P. & Groner, R. (2011) Comparing reading processes on e-ink displays and print. Displays, 32(5), 268 – 273.

Siegenthaler, E., Wurtz, P., and Groner R. (2010) Improving the Usability of E-Book Readers. Journal of Usability Studies, 6(1), 25 – 38.

Silvers, V. L., Kreiner, D. S. (1997) The effects of pre-existing inappropriate highlighting on reading comprehension. Reading Research and Instruction, 36(3), 217-223.

Tashman, C. S., Edwards, W. K. (2011a) Active reading and its discontents: the situations, problems and ideas of readers. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11), Vancouver, Canada, 7 – 12 May, 2011, 2927-2936. ACM.

Tashman, C. S., Edwards, W. K. (2011b) Active reading and its discontents: the situations, problems and ideas of readers. In Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11), Vancouver, Canada, 7 – 12 May, 2011, 3285-3294. ACM.

Welinder, P. and Perona P. (2010) Online crowdsourcing: rating annotators and obtaining cost effective labels. IEEE Conference on Computer Vision and Pattern, 1526-1534.

Zhang, H., André, P., Chilton, L. B., Kim, J., Dow, S. P., Miller, R. C., Mackay, W. E. & Beaudouin-Lafon, M. (2013) Cobi: communitysourcing large-scale conference scheduling. In CHI '13 Extended Abstracts (CHI EA '13), Paris, France, 27 April – 2 May 2013, 3011 – 3014. ACM.