Impact of English Reading Comprehension Abilities on Processing Magazine Style Narrative Visualizations and Implications for Personalization

Dereck Toker[1], Robert Moro[2], Jakub Simko[2], Maria Bielikova[2], Cristina Conati[1]
1 Department of Computer Science, University of British Columbia, Vancouver, Canada
2 Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Slovakia
{dtoker, conati}@cs.ubc.ca {robert.moro, jakub.simko, maria.bielikova}@stuba.sk

ABSTRACT

In this paper, we present research to uncover how the level of reading comprehension abilities impacts how users process textual documents in English with embedded visualizations (i.e., *Magazine Style Narrative Visualizations* or *MSNVs*). We analyze performance and gaze data of users processing MSNVs from two user studies, one run in Canada and one in a non-English speaking European country. Our findings provide important insights toward developing automatic, real-time support to MSNV processing personalized according to users' English reading comprehension abilities.

KEYWORDS

Narrative visualizations; User-adaptation; Personalization; Eye tracking; English reading comprehension abilities

ACM Reference format:

Dereck Toker, Robert Moro, Jakub Simko, Maria Bielikova and Cristina Conati. 2019. Impact of English Reading Comprehension Abilities on Processing Magazine Style Narrative Visualizations and Implications for Personalization. In *UMAP '19: 27th ACM Conference on User Modelling, Adaptation and Personalization, June 9–12, 2019, Larnaca, Cyprus. ACM New York, NY, USA, 9 pages.* https://doi.org/10.1145/3320435.3320447

1 Introduction

There is increasing evidence that individual differences such as cognitive abilities [6,7,26,43,45], expertise [27], and personality traits [36] impact how users process specific information visualizations. These findings have prompted research on how to *personalize* visualization-based interactions to the specific abilities and traits of each individual user. Most of the research so far has focused on tasks involving just visualizations. However, there has also been initial work [42] on providing personalized support for processing *Magazine Style Narrative Visualizations* (MSNV for short), i.e., visualizations embedded in narrative text [39] (e.g., Figure 1) as they are frequently found in news articles, magazines, reports and instructional material.

Combining text and graphical modalities is a widespread and well-established approach to convey complex information (e.g., [11,28,31,38]), but it can be prone to, e.g., the *split-attention effect*, where the user's attention is split between two information sources with a possible increase in cognitive load and negative impact on comprehension [1]. This can be exacerbated in MSNVs where the text can make multiple *references* to the accompanying visualizations (e.g., two references in Figure 1), each soliciting attention to different aspects of the data being visualized and requiring the user to perform different visual tasks.

To reduce possible negative effects on comprehension that might be generated by repeated transitions between text and graphics, Carenini et al. [4] proposed to provide dynamic attention guidance to visualization processing as users read the corresponding textual reference. This guidance is a form of cuing, which has been mostly investigated to support learning from multi-modal material in instructional settings [14].

Toker et al. [42] followed up on the idea of dynamic guidance for MSNVs by investigating whether it should be personalized to specific user abilities. They conducted a study to test whether lower levels of specific abilities related to visual processing and reading generated difficulties when processing MSNVs. A measure related to *English reading comprehension ability* was one of those identified to have a significant impact on MSNV processing, and thus to be a suitable target for personalization.

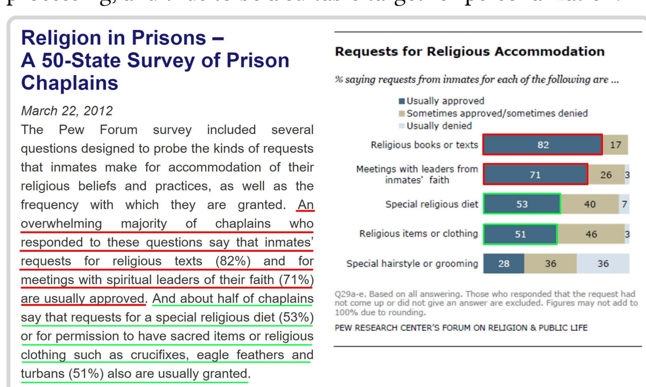

Figure 1: Example of MSNV with two references to the embedded visualization (one underlined in green and one in red, for illustration purposes).

The impact of English comprehension reading ability (simply *reading ability* from now on) found in [42] was notable considering that the study was conducted in an English-speaking country. The majority of participants were educated adults; either native English speakers or using mainly English on a daily basis. These results gave us the idea to further explore the impact of reading ability when considering users in *non-English speaking countries*. Although recent years have seen a rise in non-English content on the Web, a substantial amount of content containing MSNVs are still in English (e.g., textbooks, scientific publications, blogs and other Web content) [17]. Thus, when investigating how to provide personalized support to the consumption of this content, we argue that it is crucial to also study users in non-English speaking countries. These users can consume content in English but do not use English as their primary means of communication on a daily basis. We refer to these users as *NESC* (Non-English speaking country) users.

As a first step in this direction, this paper broadens the work in [42] in two ways:

- We replicate the study in Slovakia and perform an analysis specific to ascertaining the impact of measures related to English reading comprehension ability.
- We analyze gaze data collected with non-intrusive eye-trackers, to ascertain if there are specific attention patterns that negatively affect MSNV processing for users with lower reading ability, and who can be the target for adaptive support.

One challenge in this work was to find a measure of English reading ability suitable to our Slovak users, considering there are many different factors that can influence a user's reading ability, and time constraints can limit the type and number of tests that are feasible to administer. Here, we propose a measure that combines standard tests normally used in English speaking countries along with self-reported data. We show that lower levels of this combined measure negatively impact MSNV processing, specifically by affecting document processing speed. Furthermore, by analyzing users' gaze patterns we then identify several behaviors that explain this negative impact. Notably, these behaviors are related not only to text reading but also to the joint processing of textual and graphical information. Namely, users with lower reading ability need to transition more from text to visualization, and take longer to locate the relevant bars within the visualization. This suggests that that these users would benefit from attention guidance that would help them establish the mapping between textual references and corresponding visualization elements.

Our findings are significant because: *(i)* they contribute to the so far limited understanding of the impact of reading ability to the processing of material that combines both text and graphics; and (ii) provide evidence toward the need and design of adaptive interventions that specifically target the multimodal nature of the MSNVs for users with low reading ability. Furthermore, our focus on general purpose MSNVs extends existing research on processing multimodal information that has mostly been related to educational material.

In the rest of the paper, we first present related work. Next, we describe the MSNV user study conducted in the non-English speaking country (Slovakia). We then evaluate the impact of measures related to English reading ability on MSNV performance. After, we provide an analysis of eye-tracking data to identify patterns that negatively affect MSNV processing for users with low English reading abilities. Lastly, we discuss our results, contributions, and implications toward the design of user-adaptive MSNVs.

2 Related Work

There has been extensive research in psychology on investigating how people process combinations of textual and graphical information, mainly related to instructional material. Some of this research focused on the impact of individual differences, but there are very limited findings specific to *reading abilities*.

Kalyuga et al. [22,23] investigated viewer's *expertise* as a factor that influences whether instructional material consisting of two modalities increases comprehension or creates overload. They found for instance, that inexperienced electrical trainees learned better from diagrams of electrical circuits integrated with textual explanations, whereas more experienced trainees performed better with the diagram only [23].

Wiley et al. [47] presented evidence that *working memory capacity* (WMC – a measure of one's ability to use one's working memory system) can predict learning from illustrated text. Specifically, lower WMC reduced a reader's ability to select specific information in each modality and integrate it to develop overall understanding. Based on these findings, the authors discuss various forms of personalized support for learners with low WMC.

Hegarty and Just [19] looked at the impact of both students' *reading proficiency* and *aptitude for reasoning with mechanical principles* when studying material on pulley systems that contained both text and diagrams. There was a marginal effect of mechanical aptitude on learning time (higher for low ability students), explained by significant differences found from the analysis of the learners' gaze patterns: low mechanical ability students re-read more clauses in the text and inspected the diagram more often. No effect, however, was found for reading ability, possibly because participants were students at one of the top U.S. universities and thus likely had high reading abilities.

There has been some work on investigating how to facilitate the processing of educational text with graphics via *cuing*, i.e., adding visual prompts that guide learners' attention to relevant elements in multimodal material (see [14] for an overview). For instance, Folker at al. [13] and Ozcelik et al. [37] showed that color-coding matching parts of the text and the graphics can increase comprehension. However, this approach is limited by possibly not having enough easily distinguishable colors for color matching. Kalyuga [21] addressed this limitation by color matching corresponding parts of text and graphics dynamically. Novices studying electric circuits explained by diagrams and text, received guidance when they clicked on a specific paragraph that consisted of coloring in both the text and the diagram of all the electrical elements mentioned. Novices who received this guidance learned significantly more than those who did not. The results from our work support the idea of using a similar approach to users processing MSNVs, but tailored to a user's reading ability.

Although we are not aware of other work specifically targeting personalized support for processing text with graphics, there is initial research on providing personalized guidance to reading only, and to processing stand-alone visualizations.

For reading, in Loboda et al. [29] eye-tracking was used to infer word relevance and user information needs during reading tasks, useful to provide personalized content. D'Mello et al. [8] looked at supporting reading by detecting instances of users' mind wandering, and then intervening to refocus their attention.

For visualization processing, examples of personalized guidance include suggesting a different visualization based on detected user needs such as suboptimal behaviors [15] and evolving knowledge [18] or changing aspects of the current visualization [35]. Carenini et al. [5] evaluated several forms of dynamic highlighting (e.g., bolding, arrows) to guide attention to relevant data points within grouped bar charts, and showed a significant improvement in task performance compared to using no interventions.

3 User Study

The study that we conducted to collect data on how users in a non-English speaking country (*NESC* users) process MSNVs in English was held at the Slovak University of Technology in Bratislava, Slovakia[1]. The study included 52 participants (15 females), ranging in age from 20 to 35 (avg. = 23.1). Participants were recruited among university students and using the university Facebook page). 80% of the participants were students from computer science or engineering, 10% were students from other fields (e.g., chemistry or finances), and 10% had a variety of other occupations (e.g., veterinary surgeon, civil engineer, marketing assistant). In comparison, the original study in [42] which was carried out in an English speaking country (*ESC* users), included 56 participants (32 female) ranging in age from 19 to 69 (avg. = 28.02). 60% of participants in the original study were university students, and the others were from a variety of backgrounds (e.g., retail manager, restaurant server, retired).

The study was conducted in a room with 20 computer stations [2] arranged as shown in Figure 2. Each station was equipped with a Tobii Pro X2-60, a non-intrusive camera-based eye-tracker mounted below a 1920x1200 pixel screen. During the study, at most 12 stations were used at once; the participants were seated so that they would not disturb each other.

3.1 Study Procedure

The study procedure was directly taken from [42], with the difference that in [42], the participants were run one at a time

Figure 2: Eye tracker lab setup used for conducting the study in the non-English speaking country.

(due to eye-tracker availability). The group set-up in the new study resulted in minor changes to the procedure, mainly related to synchronizing the various stages of the study and avoiding disturbance in the group setting, as described later.

The experiment was a within-subjects repeated measures design, lasting at most 120 minutes. Four to five experimenters were in the lab for the duration of each session. One of them would start the session by providing a scripted summary of the study objectives and structure, which included the administration of a variety of well-established standardized tests for cognitive abilities/traits that were tested in [42]. Although here in this work we focus only on the impact of measures relating to reading ability, we chose to test for all the user characteristics measured in [42] for consistency and possible further analysis. Four of these tests (not the focus of this paper) were administered at the start of the session with an average duration of 14.5 minutes[2].

Next, participants underwent eye tracker calibration, lasting a few seconds. They then answered a short pre-study questionnaire to obtain demographic information and to self-report on their preference and ability with English. After, participants were tasked to read 15 different MSNVs displayed one at a time on their screen in randomized order. For each MSNV, participants signaled that they were done reading by clicking a 'next' button at the bottom left of the screen. After this, they received a screen with questions eliciting their opinion and testing their comprehension of the recent document. Participants were not given a time limit to read the MSNVs. However, to increase their motivation to put effort in the tasks, they were told that their performance would be evaluated in terms of speed and accuracy and the top three participants would receive a 50€ reward.

After reading the 15 MSNVs, a participant would leave the room (to avoid disturbing those still working) and would take a short break. Next, the participant proceeded to take three remaining tests from the set administered in [42]. First, they moved to a new room where they took a short personality test (not used in this paper) followed by the *X-Lex* vocabulary test. This part lasted about 6 minutes. Then they proceeded to another room where they took the *NAART* test, lasting 5 minutes on average. Because NAART requires reading aloud, it was administered one participant at a time while others waited their

[1] http://uxi.sk, User Experience and Interaction Research Centre

[2] These tests measured: *perceptual speed, spatial memory, disembedding,* and *visual working memory.* For test sources and definitions see [42].

turn outside. Since the *X-Lex* and *NAART* tests are the focus of this paper, they are described in more detail in the next section.

3.2 Materials

All study materials were the same as those used in [42].

3.2.1 Study Tasks. The 15 MSNVs used for the study were derived from an existing dataset of magazine-style documents where the textual references to the accompanying visualizations in each document had been manually identified via crowdsourcing to indicate which data points in each visualization correspond to each reference [24]. All 15 MSNVs were self-contained excerpts from longer articles extracted from real-world sources including *Pew Research*, *The Guardian*, and *The Economist*. They were selected to each include one visualization (a simple, stacked, or grouped bar chart [34]), and a body of narrative text ranging between 42 and 228 words (avg. = 91) containing 1 to 7 references (avg. = 2.6). The number of words and references were varied to account for their potential influence on MSNV processing.

3.2.2 Dependent Measures. For each MSNV, we collected:

- Two measures of performance: *Time on Task*, as the time to read the MSNV; *Task Accuracy*, as the percentage of correct questions answered for the MSNV.
- Two subjective measures related to perceived *Ease-of-Understanding* and document *Interest*.

Accuracy and subjective measures were assessed via a set of questions shown to the users on one screen after they read each MSNV document (see Figure 3). The first two questions in Figure 3 measured on a 5-point Likert Scale perceived ease-of-understanding and interest based on the work of [46]. The remaining questions measured document comprehension, based on the work [10] consisting of:

- One or two (depending on document length) *recognition questions* asking to recall specific information from the MSNV: either a named entity discussed in the text (e.g., question 3 in Figure 3); or the magnitude/directionality of two named entities (e.g., question 4 in Figure 3).
- One *title question* which asks to select a suitable alternative title for the MSNV (see question 5, bottom of Figure 3) and provides a simple way to ensure that the user had a grasp of the general document narrative.

The comprehension questions were designed so they could not be answered relying solely on general knowledge (i.e., measured comprehension actually reflects the content of the MSNVs)[3].

3.2.3 Measures to Assess Reading Ability. Many standard tests for reading ability tend to be quite long. For instance, the *ESOL*, *IELTS*, and *TOEFL iBT* all require more than an hour to administer, which was not feasible for this user study. Instead, following [42], we selected two well established tests to measure

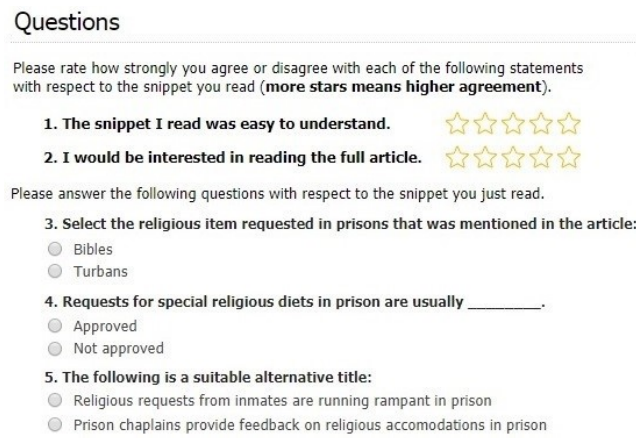

Figure 3: Questions presented to users after reading each MSNV document.

constructs relating to English reading comprehension ability that are designed for time-constrained settings:

- *X_Lex Vocabulary Test* [32]. This paper-based test consists of reading through a list of 180 words from which some are non-existent (i.e., fake). The users are instructed to indicate all the words for which they know the meaning and are scored based on the number of selected words that are real as well as the number of selected words that are non-existent. This test provides a quick method for profiling the English vocabulary of users, and scores from this test correlate well with English reading comprehension [33].
- *North American Adult Reading Test (NAART)* [40]. This is a spoken test, for which users are recorded as they read aloud a series of 61 increasingly difficult English words and they are scored based on the number of correctly pronounced words. The recordings from our study were scored by a native English speaker following the *NAART* scoring manual. *NAART* is a good predictor of several constructs related to reading ability, i.e., verbal reasoning, verbal comprehension, or semantic knowledge [3].

In addition, users self-reported on 3 different questions relating to their English reading ability:

- *"What is your native language/first language learned?"* Scored as 1 for English, 0 otherwise.
- *"Currently in your everyday life, is English your preferred language to speak/read?"* Scored 1 for English, 0 otherwise.
- *"Please self-report your English language proficiency."* Measured on a 4-point scale; from beginner to expert. Scored from 1 to 4.

4 Impact of Reading Ability on Task Performance

As we discussed in the introduction, our goal with this work is two-fold: *(i)* ascertain if we can reproduce with users in a non-English speaking country (NESC) the impact of measures related to English reading comprehension abilities on MSNV performance found in [42]; *(ii)* see if we can explain this effect in

[3] This was tested in a pilot study were users were given these questions without having seen the MSNVs to ensure that the overall mean accuracy achieved on each question was at most 50%.

terms of suboptimal attention patterns. This section describes the analysis we conducted toward (i).

For our analysis, we selected Linear Mixed-Effects Models since they can handle multiple random effects at once. Because our study was a repeated measures design where all users were exposed to the same set of 15 documents, there are two random effects to handle. The first random effect *user_id* accounts for within-subject correlations due to the fact that multiple measurements were collected from the same user. The second random effect *MSNV_id* accounts for within-document correlation due to repeated measurements being collected from the same MSNV document. We used the *lmerTest* software package in R [25] to construct a separate mixed model for each dependent measure to be analysed, along with *NAART* and *X-Lex* as covariates, and *user_id* and *MSNV_id* as random effects.

We started by testing whether we could replicate with our new pool of *NESC* users, the effects of measures related to reading abilities (i.e., *NAART* and *X-Lex*) on task performance that were found in [42], but we found no effects. One possible explanation for this outcome is that the *NESC* users were all comparable in terms of task performance. Given these users reside in a non-English speaking country, it's also possible they were equally slower compared to users from an English speaking country. A comparison of *time on task* indeed shows that *NESC* users were slower reading the MSNV documents (Mean=69.2 sec, SD=32.7) vs. the *ESC* users in [42] (Mean=57.2 sec, SD=33.1). Therefore, we chose to combine our pool of *NESC* users with the pool of *ESC* study users to increase the range of performance in our dataset and thus improve our ability to detect potential differences due to English reading ability. The pool of users we added from the *ESC* consisted of 56 subjects (see Section 3 for their demographics); yielding a combined dataset of 108 users.

4.1 Combining Users' Reading Ability Scores

There are many factors in various combinations that can explain the reading abilities of users [16]. In addition to the *X_Lex* and *NAART* tests that were administered, we also collected three *self-assessed* measures related to English reading ability (described in the previous section). Among these self-assessed measures, noteworthy differences were reported between the two populations. For instance, in the *NESC* only 3.9% of users' native language was English vs. 50.0% in ESC; and for preferred language only 5.5% of users reported English in the *NESC* vs. 76.8% in the *ESC*). Given the potential impact that these distinguishing measures could have on English reading ability (in addition to the *NAART* and *X_LEX* tests), we wanted to ensure that we leveraged as much information from these different sources as possible to characterize users' reading ability across the pooled dataset of *ESC* and *NESC* users. To this end, we performed a dimensional reduction, so that users' English reading ability could be modeled in as few variables as necessary. We opted to use Principal Component Analysis (PCA), which facilitates the identification and combination of groups of inter-related variables into components more suitable for data analysis [12]. A PCA on the five different measures resulted in one output component. Bartlett's test of sphericity x^2 (10) = 3103.37, p < .001, indicated that the PCA was appropriate. Kaiser's sampling adequacy was good at 0.79 [20], and all variables showed a communality > 0.65 which is above the acceptable limit of 0.51 [12]. The component we generated had an eigenvalue over Kaiser's criterion of 1 and explained 59.4% of the variance, providing us with one suitable measure that alleviates the need to include multiple measures related to reading ability in our subsequent analyses. In the rest of the paper we refer to the component we generated as the *Combined English Reading Proficiency* score (or *CERP* for short).

Figure 4 shows the distribution of *CERP* scores in our pooled dataset based on users' study location (*ESC* vs. *NESC*). In general, users from the *NESC* population have lower *CERP* scores (not a surprise), and the range of scores, when compared to the *ESC* population, is much narrower. In contrast, the scores in the *ESC* population are overall much higher (to be expected), but the spread of scores is much wider. This larger spread in the *ESC* population is likely attributed to the fact that 50% of them are non-native English speakers. However, unlike users in the *NESC* population, non-native English speakers in the *ESC* receive more regular exposure to English and are likely more comfortable working in English (as evidenced 76.8% of *ESC* users who prefer to speak/write in English in their everyday lives). In other words, there appears to be an important distinction between non-native English speakers living in a non-English speaking country vs. non-native English speakers living in an English speaking country, and what we observe in Figure 4 appears to be consistent with this idea.

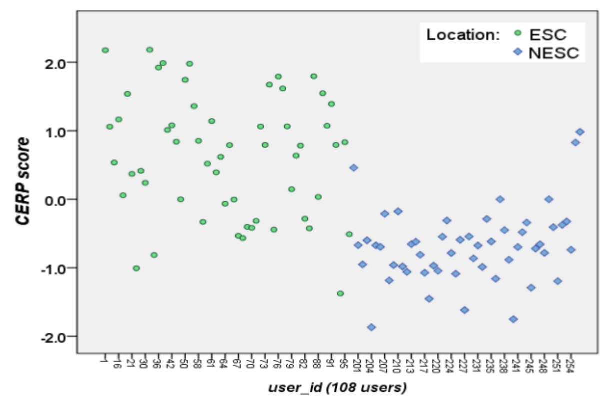

Figure 4: Distribution of users from the English Speaking Country and Non-English Speaking Country (NESC) populations according to their Combined English Reading Proficiency (*CERP*) scores.

4.2 Analysis of *CERP* on MSNV Performance

We constructed one Mixed Model for each measure of task performance as the dependent measure, along with *CERP* (continuous variable) as a covariate, and *user_id* and *MSNV_id* as random effects. Results are as follows:

For the two objective measures, a significant effect, $b = -5.93$, $t(-3.73)$, p < .001, was found for *CERP* on *Time on Task*. The negative slope of b indicates that users with lower *CERP* spent more time on task. Using a median split on users' *CERP* scores, we found that it took the *low CERP* group 70.7 seconds on average to read the excerpts, while only 55.7 seconds were needed for *high CERP* users. No effect however, was found for *CERP* on *Task Accuracy*, $b = -0.003$, $t(-0.46)$, $p = .64$. In conjunction though, these first two results provide evidence that all users were similarly accurate regardless of *CERP* score such that low *CERP* users likely needed extra time to achieve comparable accuracy as their counterparts. The most straightforward explanation as to why these users are struggling is precisely because of their low English reading comprehension abilities as captured by *CERP*.

For the two subjective measures, no significant effect was found for *CERP* on document *Ease-of-Understanding*, $b = 0.10$, $t(1.77)$, $p = .08$, indicating that users rated the documents similarly regardless of their *CERP* score. Lastly, a main effect was found for *CERP* on document *Interest*, $b = 0.29$, $t(3.91)$, $p < .001$. The positive slope of b indicates that users with low *CERP* rated the documents as less interesting.

Given our results, *Time on Task* offered the strongest indication that users with low *CERP* were objectively struggling on task, and thus could likely benefit the most from some form of adaptive support. Therefore, we select *Time on Task* as the primary measure of performance for further investigation in the next section.

5 Gaze Analysis of MSNV Processing

In this section, we leverage eye tracking data to address our second research goal, namely, to see if we can explain the effect of low English reading ability (as captured by users' *CERP* score) in terms of suboptimal attention patterns during MSNV processing. First, Section 5.1 explains how the eye tracking data collected during the user study was utilized to generate numerous gaze metrics that capture various MSNV processing behaviors. Next, in Section 5.2 we analyze the set of generated gaze metrics, to identify which among them are relevant to performance with the MSNVs (i.e., time on task). Lastly, in section 5.3 we conduct an analysis to see if any of the gaze metrics relevant to MSNV performance are significantly influenced by users' *CERP* score.

5.1 Computing Gaze Metrics

Raw gaze data comprises of fixations (points of gaze on the screen) and saccades (quick movements between fixations). In order to capture a more detailed understanding of users' MSNV processing, we compute from the raw gaze data a set of summary statistics describing numerous aspects of their gaze behaviors following a standard approach adopted in many other works [9,30,41,44]. Users' raw gaze data was processed using EMDAT, an open source library *(github.com/ATUAV/EMDAT)* that produces a comprehensive set of gaze-metrics specified over the entire display, and over specific *areas of interests* (AOIs). For our analysis, we selected only AOI-based gaze metrics because those specified over the entire display do not capture any information relating to the content of the MSNVs and thus are not useful for our research goal. The complete set of gaze metrics we selected are listed in Table 1. These metrics are defined over four AOIs chosen to gain a general sense of MSNV document processing with respect to the two sources of information, text and visualization. The four AOIs (see Figure 5) are defined as:

- ***Refs AOI:*** The combined areas of all the reference phrases contained in the MSNV document (purple-shaded boxes).
- ***Text AOI:*** The rest of the MSNV document text (orange box minus purple boxes).
- ***Referenced Bars (R-Bars) AOI:*** The combined area of all the bars in the visualization that are mentioned by any of the references (green boxes).
- **Viz AOI:** The rest of the visualization region (pink box minus green boxes).

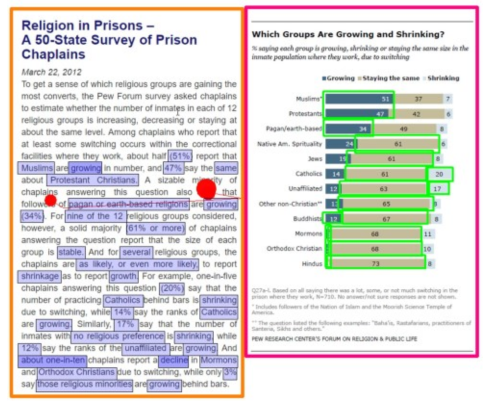

Figure 5: Example of the four AOIs we defined to capture MSNV processing.

Table 1: Set of 17 gaze metrics generated for each of the 4 AOIs shown in Figure 5.

No.	Gaze Metric	Description
1	• *fixation_rate*	Fixation rate in AOI (number of fixations ÷ total time spent in AOI)
2	• *number_of_fixations*	Total number of fixations in AOI
3	• *longest_fixation*	Longest fixation in AOI
4-6	• *sum_fix_durations* • *mean_fix_durations* • *stddev_fix_durations*	Sum, Mean, and Std. Deviation of fixation durations in AOI
7-8	• *time_to_first_fix* • *time_to_last_fix*	Time to first and last fixation in AOI
9-12	• *transitions_to_Text* • *transitions_to_Viz* • *transitions_to_Refs* • *transitions_to_R-Bars*	Number of gaze transitions from this AOI to every AOI
13-16	• *prop_ trans_to_Text* • *prop _trans_to_Viz* • *prop _trans_to_Refs* • *prop _trans_to_R-Bars*	Proportion of gaze transitions from this AOI to every AOI (according to total gaze transitions in all AOIs)
17	• *prop_num_fixations*	Proportion of fixations in AOI (according to total fix. in all AOIs)

5.2 Identifying Relevant Gaze Metrics

Here, our goal is to identify which gaze metrics have a significant relationship with MSNV *time on task.* For the purposes of our research, gaze metrics found with no significant relationship to time on task are non-relevant and do not warrant further consideration. Non-relevant gaze metrics offer no concrete indication on how the captured processing behavior translates to MSNV performance, and thus provide little guidance towards designing meaningful adaptive support.

We construct one Mixed Model for each of the 56 gaze metrics (14 gaze metrics x 4 AOIs)[4], using gaze metric as the independent measure, with time on task as the dependent measure, and *user_id* and *MSNV_id* as random effects. Given the relatively high number of models to check, we account for multiple comparison error by adjusting the obtained *p*-values using a Bonferroni correction [12] equal to the total number of gaze metrics within each family of AOIs (in this case 14). All significant results we found are reported in Table 2.

Table 2: Gaze metrics in each AOI found to be significant with *time on task*. The coefficient *b* indicates the directionality of the relationship (+/–). Grey cells indicate metrics excluded due to high correlations (i.e., transitions within the same AOI are highly correlated to *sum_fix_durations*).

Gaze Metric	Text AOI	Refs AOI	Viz AOI	R-Bars AOI
sum_fix_durations	$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$
longest_fixation	$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$
time_to_first_fix	not sig.	$p < .001$ $b = +$	$p < .01$ $b = +$	$p < .001$ $b = +$
time_to_last_fix	$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$
mean_fix_durations	$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$
fixation_rate	$p < .001$, $b = –$	not sig.	not sig.	not sig.
transitions_to_Text		$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$
transitions_to_Refs	$p < .001$ $b = +$		$p < .001$ $b = +$	$p < .001$ $b = +$
transitions_to_Viz	$p < .001$ $b = +$	$p < .001$ $b = +$		$p < .001$ $b = +$
transitions_to_R-Bars	$p < .001$ $b = +$	$p < .001$ $b = +$	$p < .001$ $b = +$	

In all cases but one, the relationship between gaze metric and time on task has positive directionality indicated by the estimated coefficient *b* (i.e., higher values of the corresponding gaze metric relate to longer times on task). The exception is *fixation_rate* on the Text AOI, which has negative directionality (i.e., higher fixation rates in the Text related to lower time on task). We found no significant results for any of the 20 proportion-based gaze metrics (cf. metrics 13-17 in Table 1), and are thus not shown in Table 2. Overall, we identified a total of 32 relevant gaze metrics that have a significant effect on MSNV time on task. Even though it is not surprising that many of these gaze metrics are highly correlated to time on task (i.e., they get bigger as more time is spent on task), we report them here for completeness. The interesting part of these identified gaze metrics will surface in the next part of our analysis, when they are examined to see to what extent any of these relationships are qualified by users' reading ability.

[4] 17 gaze metrics were generated per AOI, however 3 gaze metrics were removed in each AOI from further analysis due to very strong positive correlations (r > 0.9). In each AOI: *number_of_fixations* and *transitions_to_self* were removed due to high correlation with *sum_fixation_durations*; and *stddev_fix_durations* was removed due to high correlation with *longest_fixation*.

5.3 Impact of *CERP* on Relevant Gaze Metrics

As identified in Section 4.2, we found that the users' *CERP (combined English reading proficiency)* impacts task performance with MSNVs in a manner that may call for personalized support: namely users with low *CERP* spend significantly more time on task to achieve comparable accuracy compared to users with higher *CERP* scores. Here, our goal is to see if *CERP* impacts any of the 32 relevant gaze metrics identified in the previous subsection, so as to identify possible gaze processing behaviors exuded by users with low *CERP* scores which are causing lower MSNV task performance. We construct one Mixed Model for each of the 32 relevant gaze metrics as the dependent measure, with *CERP* as a continuous covariate, and *user_id* and *MSNV_id* as random effects. As before, we apply a Bonferroni correction equal to the total number of gaze metrics within each family of AOIs (in this case 8).

Results revealed a significant effect of *CERP* on seven of the tested gaze metrics (reported in Table 3). For all seven cases, the slope *b* is negative, indicating that lower *CERP* users produced higher values of the corresponding gaze metric.

Table 3: Results showing in which AOIs a significant effect of *CERP* was found on the corresponding gaze metric. Metrics in grey cells were not relevant to time on task.

Gaze Metric	Text AOI	Refs AOI	Viz AOI	R-Bars AOI
sum_fix_durations	***CERP*** $p < .05$ $b = –$	not sig.	not sig.	not sig.
time_to_first_fix		not sig.	not sig.	***CERP*** $p < .05$ $b = –$
time_to_last_fix	***CERP*** $p < .001$ $b = –$	***CERP*** $p < .05$ $b = –$	***CERP*** $p < .01$ $b = –$	***CERP*** $p < .05$ $b = –$
transitions_to_Viz	***CERP*** $p < .05$ $b = –$	not sig.		not sig.

First, we found that low *CERP* users have significantly higher times to last fixation in all of the AOIs (see third row of Table 3). Since we observe it in all AOIs, little guidance is offered by this result and may just be a direct consequence of users with low *CERP* spending overall more time processing MSNVs. Similarly, we would expect by nature of *CERP* the low *CERP* users to spend more time processing the narrative parts of MSNVs, i.e., reading the text, which is confirmed by a significant effect of *CERP* on *sum_fixation_durations* in the Text AOI (first row of Table 3). In addition, the lack of effect on *sum_fixation_durations* for the visual information (Viz and R-Bars AOIs) suggests that users spent the same amount of time processing the visualization regardless of their *CERP* score.

Next, we found that users with low *CERP* are taking significantly longer to fixate (i.e., *time_to_first_fixation*) on the relevant bars in the visualization (see second row of Table 3). On average, we found that users with low *CERP* fixated on the R-Bars 32.1 seconds into the task compared to 23.0 seconds for users with high *CERP*. Interestingly, no significant effect of *CERP* was detected on *time_to_first_fixation* in the Viz AOI, meaning that users are looking at the visualization for the first time at comparable times regardless of their *CERP*. Therefore, it is likely not the case that users with low *CERP* are failing to look at the visualization soon enough, but rather they require more time to find the relevant information within the visualization; a behavior which is negatively impacting task performance and could be helped with an adaptive intervention.

We also found a significant effect of *CERP* on one transition-based gaze metric (last row of Table 3) indicating that users with low *CERP* transitioned more often from the Text AOI to the Viz AOI. This result provides further evidence that users with low *CERP* are struggling with the mappings between textual and visual information (i.e., references), and makes the case for adaptive support even stronger.

6 Discussion and Conclusion

In this paper, we conducted an exploratory user study with Magazine Style Narrative Visualizations (MSNVs) in a non-English speaking country (*NESC*), thus broadening the work in [42]. Our aim was to ascertain (*i*) if the impact of measures related to English reading comprehension abilities on MSNV processing found in [42] can be reproduced in a *NESC*; and (*ii*) if we can explain this effect in terms of suboptimal attention patterns.

Regarding (*i*), we found no effect of measures related to reading abilities using solely *NESC* users (likely due to low variance in their task performance which was overall significantly lower than those of *ESC* users from [42]). We then proceeded to pool the two datasets together from both countries, thus producing a sample of users with a substantially wider range of task performance and consequently English reading abilities. Furthermore, in order to leverage all of the information we collected relating to users' English reading abilities, we introduced a combined measure of English reading proficiency (*CERP* score) generated using PCA.

Our results from the pooled data confirmed the original findings in [42]. Namely, users' English reading comprehension ability (as expressed by their *CERP* score) significantly impacted task performance with MSNVs. We found that users with lower reading abilities (*CERP* score) required significantly *more* time to reach the same degree of MSNV comprehension as higher ability users (about 12 seconds on average per MSNV). Even though 12 seconds may seem a rather short amount of time to warrant adaptive interventions, bear in mind that the MSNV documents we administered consisted of very short excerpts of much longer documents. Therefore, in a real-world setting where MSNVs are typically much longer (i.e., many paragraphs and pages) and contain many visualizations, it is very likely that the effect we found of reading ability on task time would be greatly exacerbated for lower reading ability users.

Concerning (*ii*), our analysis of eye tracking data identified several MSNV processing behaviors of users with low English reading abilities (*CERP*) from the *ESC* and *NESC* that significantly contributed to lower task performance. Users with low reading ability were primarily struggling (i.e., behaviors that negatively impact performance) in two ways. First, they spent more time processing the narrative parts of MSNVs (i.e., reading the text); a result to be expected given the text is where the bulk of the reading occurs in the MSNV. Second, they struggled to locate the relevant bars in the visualization. Namely, it took them significantly longer to fixate on them for the first time, and they transitioned significantly more often from the narrative text to the visualization.

Our findings contribute to the future design of MSNVs with user-adaptive support in two major ways. First, we identified a significant user trait that such a system can leverage to determine which users to provide adaptations to; namely, to help users with low English reading abilities who are slower on task. Second, we provided significant evidence regarding how these users are struggling while processing MSNVs, which directly supports the implementation of highlighting interventions (e.g., bolding, arrows) to help guide users' attention to relevant information in the visualization, similar to what was shown in [6] for tasks involving a single visualization only. In that work however, highlighting was provided only once per task, and only for a single subset of bars in the visualization. Since MSNVs typically contain many different references, each eliciting a different subset of data in the visualization, we propose using eye tracking to detect which part of the narrative text users with low reading ability are currently processing, so that highlighting can be provided in real-time to the corresponding set of bars in the visualization.

Our results are of interest because they contribute to existing work on how to enhance the value of multimodal presentation of information based on written text and graphics. Whereas most of the existing works focus on learning with instructional texts, here we investigated a more open-ended task of processing MSNVs. Furthermore, our work adds English reading comprehension abilities to the list of user characteristics identified to be important in multimodal processing, which thus far has consisted mainly of domain expertise.

As future work, we are planning to conduct a study with MSNVs to test the effectiveness of the adaptive strategy that our findings advocated. Namely, we will use eye tracking to detect which part of the narrative text a user is currently reading and highlight the corresponding set of bars in the visualization to help users with low English reading abilities find them sooner.

ACKNOWLEDGMENTS

This work was partially supported by: the Natural Sciences and Engineering Research Council of Canada (NSERC) project No. 22R8054; and the education and research development project No. 002STU-2-1/2018, projects No. APVV-15-0508, VG 1/0725/19, and VG 1/0667/18, in the Slovak Republic.

REFERENCES

[1] P Ayres and G Cierniak. 2012. Split-Attention Effect. In *Encyclopedia of the Sciences of Learning, ed. Norbet M. Seel.* Springer, 3172–3175.

[2] Maria Bielikova, Martin Konopka, Jakub Simko, Jozef Tvarozek, Patrik Hlavac, and Eduard Kuric. 2018. Eye-tracking en masse: Group user studies, lab infrastructure, and practices. *Journal of Eye Movement Research* 11, (3): 6. https://doi.org/10.16910/jemr.11.3.6

[3] Jennifer R. Blair and Otfried Spreen. 1989. Predicting premorbid IQ: A revision of the national adult reading test. *Clinical Neuropsychologist* 3, 2: 129–136. https://doi.org/10.1080/13854048908403285

[4] Giuseppe Carenini, Cristina Conati, Enamul Hoque, and Ben Steichen. 2013. User Task Adaptation in Multimedia Presentations. In *Proceedings of the 1st International Workshop on User-Adaptive Information Visualization (WUAV 2013), in conjunction with UMAP 2013.*

[5] Giuseppe Carenini, Cristina Conati, Enamul Hoque, Ben Steichen, Dereck Toker, and James T. Enns. 2014. Highlighting Interventions and User Differences: Informing Adaptive Information Visualization Support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI'14), 1835–1844. https://doi.org/10.1145/2556288.2557141

[6] Cristina Conati, Giuseppe Carenini, Enamul Hoque, Ben Steichen, and Dereck Toker. 2014. Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. In *Computer Graphics Forum*, 371–380. https://doi.org/10.1111/cgf.12393

[7] Cristina Conati and Heather Maclaren. 2008. Exploring the role of individual differences in information visualization. In *Proc. of the working conference on Advanced visual interfaces*, 199–206. https://doi.org/10.1145/1385569.1385602

[8] Sidney K. D'Mello, Caitlin Mills, Robert Bixler, and Nigel Bosch. 2017. Zone out no more: Mitigating mind wandering during computerized reading. In *Proc. of the 10th International Conference on Educational Data Mining (EDM 2017)*, 8–15.

[9] Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *Journal of Human-Computer Studies* 70, 5: 377–398. https://doi.org/10.1016/j.ijhcs.2012.01.004

[10] Mary C. Dyson and Mark Haselgrove. 2001. The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies* 54, 4: 585–612. https://doi.org/10.1006/ijhc.2001.0458

[11] ETSI. 2002. *Human Factors (HF); Guidelines on the multimodality of icons, symbols and pictograms.* European Telecommunications Standards Institute.

[12] Andy P Field. 2009. *Discovering statistics using SPSS.* SAGE.

[13] Sonja Folker, Helge Ritter, and Lorenz Sichelschmidt. 2005. Processing and Integrating Multimodal Material — The Influence of Color-Coding. *Proceedings of the Annual Meeting of the Cognitive Science Society* 27, 27: 690–695.

[14] Tamara van Gog. 2014. The Signaling (or Cueing) Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning* (2nd ed.), Cambridge, 263–278. https://doi.org/10.1017/CBO9781139547369.014

[15] David Gotz and Zhen Wen. 2009. Behavior-driven visualization recommendation. In *Proceedings of the 14th international conference on Intelligent user interfaces*, 315–324. https://doi.org/10.1145/1502650.1502695

[16] William Grabe and Xiangying Jiang. 2013. Assessing Reading. In *The Companion to Language Assessment.* John Wiley & Sons, Inc., Hoboken, NJ, USA, 185–200. https://doi.org/10.1002/9781118411360.wbcla060

[17] David Graddol. 2006. *English next: why global English may mean the end of "English as a foreign language."* British Council, London.

[18] Beate Grawemeyer. 2006. Evaluation of ERST: an external representation selection tutor. In *Proc. of the 4th International Conference on Diagrammatic Representation and Inference*, 154–167. https://doi.org/10.1007/11783183_21

[19] M. Hegarty and M.A. Just. 1993. Constructing Mental Models of Machines from Text and Diagrams. *Journal of Memory and Language* 32, 6: 717–742. https://doi.org/10.1006/jmla.1993.1036

[20] Graeme Hutcheson. 1999. *The Multivariate Social Scientist.* SAGE Publications, Ltd., 6 Bonhill Street, London. https://doi.org/10.4135/9780857028075

[21] Slava Kalyuga. 2009. *Managing Cognitive Load in Adaptive Multimedia Learning:* IGI Global. https://doi.org/10.4018/978-1-60566-048-6

[22] Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. 2003. The Expertise Reversal Effect. *Educational Psychologist* 38, 1: 23–31. https://doi.org/10.1207/S15326985EP3801_4

[23] Slava Kalyuga, Paul Chandler, and John Sweller. 1998. Levels of Expertise and Instructional Design. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 40, 1: 1–17. https://doi.org/10.1518/001872098779480587

[24] Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. 2014. Extracting references between text and charts via crowdsourcing. In *Proc. of the SIGCHI conference on Human Factors in Computing Systems* (CHI'14), 31–40. https://doi.org/10.1145/2556288.2557241

[25] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 13. https://doi.org/10.18637/jss.v082.i13

[26] Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2017. Impact of Individual Differences on User Experience with a Visualization Interface for Public Engagement. In *Proceedings of the 2nd International Workshop on Human Aspects in Adaptive and Personalized Interactive Environments (in conjunction with UMAP'17)*, 247–252. https://doi.org/10.1145/3099023.3099055

[27] Sébastien Lallé, Dereck Toker, Cristina Conati, and Giuseppe Carenini. 2015. Prediction of Users' Learning Curves for Adaptation while Using an Information Visualization. In *Proc. of the 20th International Conference on Intelligent User Interfaces* (IUI'15), 357–368. https://doi.org/10.1145/2678025.2701376

[28] Jason Lankow, Josh Ritchie, and Ross Crooks. 2012. *Infographics: the power of visual storytelling.* John Wiley & Sons, Inc, Hoboken, N.J.

[29] Tomasz D. Loboda, Peter Brusilovsky, and Jöerg Brunstein. 2011. Inferring word relevance from eye-movements of readers. 175. https://doi.org/10.1145/1943403.1943431

[30] Pascual Martínez-Gómez and Akiko Aizawa. 2014. Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (IUI'14), 95–104. https://doi.org/10.1145/2557500.2557546

[31] Richard E. Mayer. 2009. *Multimedia learning.* Cambridge University Press, Cambridge ; New York.

[32] Paul Meara. 2010. *EFL Vocabulary Tests.* Lognostics, Swansea: Wales.

[33] Paul Meara and Glyn Jones. 1990. *Eurocentres Vocabulary Size Test 10KA.* Eurocentres Learning Service, Zurich.

[34] Tamara Munzner. 2015. *Visualization analysis and design.* CRC Press, Taylor & Francis Group.

[35] Kawa Nazemi, Reimond Retz, Jürgen Bernard, Jörn Kohlhammer, and Dieter Fellner. 2013. Adaptive Semantic Visualization for Bibliographic Entries. In *Advances in Visual Computing* (Lecture Notes in Computer Science), 13–24. https://doi.org/10.1007/978-3-642-41939-3_2

[36] Alvitta Ottley, Huahai Yang, and Remco Chang. 2015. Personality as a Predictor of User Strategy: How Locus of Control Affects Search Strategies on Tree Visualizations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI'15), 3251–3254.

[37] Erol Ozcelik, Ismahan Arslan-Ari, and Kursat Cagiltay. 2010. Why does signaling enhance multimedia learning? Evidence from eye movements. *Computers in Human Behavior* 26, 1: 110–117. https://doi.org/10.1016/j.chb.2009.09.001

[38] Katharina Scheiter, Eric Wiebe, and Jana Holsanova. 2011. *Theoretical and Instructional Aspects of Learning with Visualizations. Instructional Design: Concepts, Methodologies, Tools and Applications.* IGI Global. https://doi.org/10.4018/978-1-60960-503-2

[39] E Segel and J Heer. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6: 1139–1148. https://doi.org/10.1109/TVCG.2010.179

[40] Esther Strauss, Elisabeth M. S. Sherman, Otfried Spreen, and Otfried Spreen. 2006. *A compendium of neuropsychological tests: administration, norms, and commentary.* Oxford University Press, Oxford ; New York.

[41] Dereck Toker and Cristina Conati. 2014. Eye tracking to understand user differences in visualization processing with highlighting interventions. In *Proceedings of the 22nd international conference on User Modeling, Adaptation, and Personalization* (UMAP'14). https://doi.org/10.1007/978-3-319-08786-3_19

[42] Dereck Toker, Cristina Conati, and Giuseppe Carenini. 2018. User-adaptive Support for Processing Magazine Style Narrative Visualizations: Identifying User Characteristics that Matter. In *Proc. of the 23rd International Conference on Intelligent User Interfaces* (IUI'18), 199–204. https://doi.org/10.1145/3172944.3173009

[43] Dereck Toker, Cristina Conati, Giuseppe Carenini, and Mona Haraty. 2012. Towards adaptive information visualization: on the influence of user characteristics. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization* (UMAP'12), 274–285. https://doi.org/10.1007/978-3-642-31454-4_23

[44] Dereck Toker, Cristina Conati, Ben Steichen, and Giuseppe Carenini. 2013. Individual user characteristics and information visualization: connecting the dots through eye tracking. *In Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (CHI'13), 295–304. https://doi.org/10.1145/2470654.2470696

[45] Maria C. Velez, Deborah Silver, and Marilyn Tremaine. 2005. Understanding visualization through spatial ability differences. In *Proc. of the IEEE Conference on Visualization*, 511–518. https://doi.org/10.1109/VISUAL.2005.1532836

[46] T. Franklin Waddell, Joshua R. Auriemma, and S. Shyam Sundar. 2016. Make it Simple, or Force Users to Read?: Paraphrased Design Improves Comprehension of End User License Agreements. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (CHI'16), 5252–5256. https://doi.org/10.1145/2858036.2858149

[47] Jennifer Wiley, Christopher A. Sanchez, and Allison J. Jaeger. 2014. The Individual Differences in Working Memory Capacity Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning* (2nd ed.), Cambridge, 598–620. https://doi.org/10.1017/CBO9781139547369.029