User Modeling and User-Adapted Interaction (2019) 29:977–1011
https://doi.org/10.1007/s11257-019-09244-5

Gaze analysis of user characteristics in magazine style narrative visualizations

Dereck Toker[1] · Cristina Conati[1] · Giuseppe Carenini[1]

Received: 29 June 2018 / Accepted in revised form: 19 August 2019 / Published online: 30 August 2019

Abstract

Previous research has shown that various user characteristics (e.g., cognitive abilities, personality traits, and learning abilities) can influence user experience during information visualization tasks. These findings have prompted researchers to investigate user-adaptive information visualizations that can help users by providing personalized support based on their specific needs. Whereas existing work has been mostly limited to tasks involving just visualizations, the aim of our research is to broaden this work to include scenarios where users process textual documents with embedded visualizations, i.e., Magazine Style Narrative Visualizations, or MSNVs for short. In this paper, we analyze eye tracking data collected from a user study with MSNVs to uncover processing behaviors that are negatively impacting user experience (i.e., time on task) for users with low abilities in these user characteristics. Our analysis leverages Linear Mixed-Effects Models to evaluate the relationships among user characteristics, gaze processing behaviors, and task performance. Our results identify several MSNV processing behaviors within the visualization that contribute to poor task performance for users with low reading proficiency. For instance, we identify that users with low reading proficiency transition significantly more often compared to their counterparts between relevant and non-relevant bars, and transition more often from bars to the labels. We present our findings as a step toward designing user-adaptive support mechanisms to alleviate these difficulties with MSNVs, and provide suggestions on how our results can be leveraged for creating a set of meaningful interventions for future evaluation (e.g., dynamically highlighting relevant bars and labels in the visualization to help users with low reading proficiency locate them more effectively).

✉ Dereck Toker
dtoker@cs.ubc.ca

Cristina Conati
conati@cs.ubc.ca

Giuseppe Carenini
carenini@cs.ubc.ca

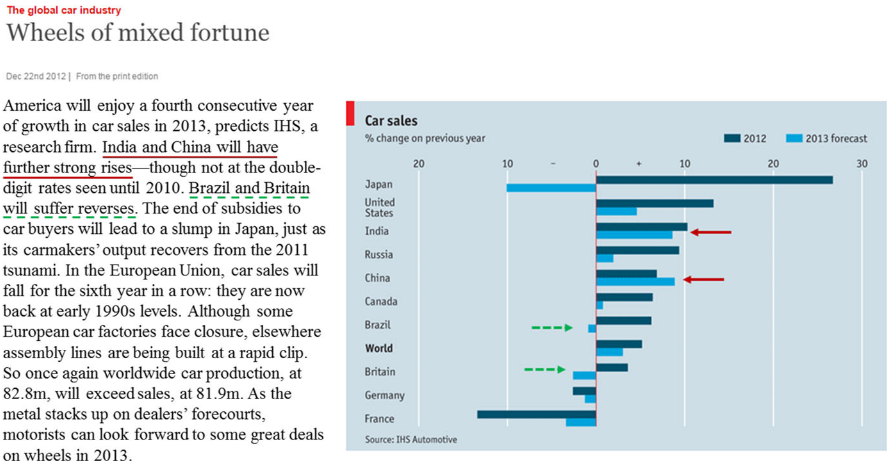
[1] University of British Columbia, Vancouver, Canada

Keywords Narrative visualization · User modeling · Eye tracking · Mixed models · Adaptive visualizations · Individual differences · Users characteristics

1 Introduction

As digital information continues to accumulate in our lives, information visualizations have become increasingly relevant for discovering trends and shaping stories from this overabundance of data (Huang et al. 2015). However, visualizations are typically designed and evaluated following a one-size-fits-all approach, meaning they do not take into account the specific needs of individual users. This is problematic because there is mounting evidence that user characteristics, such as cognitive abilities, personality traits, and learning abilities, can significantly influence user experience (e.g., performance and satisfaction) during information visualization tasks (Lallé et al. 2015; Ottley et al. 2015; Toker et al. 2012; Velez et al. 2005). These findings have prompted researchers to investigate user-adaptive information visualizations, i.e., visualizations that aim to recognize and adapt to each user's specific needs. Whereas existing work has been mostly limited to tasks involving just visualizations, the aim of our research is to broaden this work to include scenarios where users interact with visualizations embedded in narrative text, known as *Magazine Style Narrative Visualization* (Segel and Heer 2010), or MSNV for short (e.g., Fig. 1).

Combining text and graphical modalities is a widespread and well-established approach to convey complex information (e.g., Mayer 2009; Human Factors 2008; Lankow et al. 2012; Scheiter et al. 2011). In a narrative visualization, graphics and text play complementary roles. While graphics can convey large amounts of data compactly and support discovery of trends and relationships, text is much more effective at pointing out and explaining key points about the data, in particular, by focusing on specific temporal, causal, and evaluative aspects (Tufte 1997). As a result,


The global car industry

Wheels of mixed fortune

Dec 22nd 2012 | From the print edition

America will enjoy a fourth consecutive year of growth in car sales in 2013, predicts IHS, a research firm. India and China will have further strong rises—though not at the double-digit rates seen until 2010. Brazil and Britain will suffer reverses. The end of subsidies to car buyers will lead to a slump in Japan, just as its carmakers' output recovers from the 2011 tsunami. In the European Union, car sales will fall for the sixth year in a row: they are now back at early 1990s levels. Although some European car factories face closure, elsewhere assembly lines are being built at a rapid clip. So once again worldwide car production, at 82.8m, will exceed sales, at 81.9m. As the metal stacks up on dealers' forecourts, motorists can look forward to some great deals on wheels in 2013.

Fig. 1 An example of two references in an MSNV document, each consisting of a sentence in the body of narrative text and corresponding data points within the visualization. *Source*: The Economist—Dec 22, 2012

in MSNVs often there is more than one visual task specified throughout the narrative text. Multiple visual tasks in MSNVs are captured by *references*, namely segments of text that specifies a visual task on an accompanying visualization. Typically, references are used to support arguments or statements being made in the document text by providing added details or interpretations on a subset of data shown in the accompanying visualization. Figure 1 provides an example of two references in an MSNV. One reference is the sentence "*India and China will have further strong rises*," and it *refers* to the bars marked by the solid red arrows in the accompanying bar chart. The second reference is the sentence "*Brazil and Britain will suffer reverses*," and it refers to the bars pointed to by the dashed green arrows. As a user reads through an MSNV, they will often encounter a variety of references in the text, each soliciting attention to different aspects of the accompanying visualization. Visualizations in an MSNV cannot be designed to favor the visual task of any particular reference, because favoring one task may hinder the others; thus, Carenini et al. (2013) proposed to facilitate MSNV processing by interactively highlighting relevant aspects of the visualization depending on what part of the text the user is reading and possibly on the user's characteristics that may impact MSNV processing. This guidance is a form of cuing, which has been investigated to support learning from multimodal material in instructional settings; see van Gog (2014) for an overview.

The long-term goal of our work is to design and implement such user-adaptive support to MSNV processing, based on the following methodology.

- Conduct exploratory user studies in order to identify which user characteristics can impact MSNV processing and therefore may warrant adaptive support.
- Leverage eye tracking data to investigate where users who are low on the relevant abilities identified are struggling during MSNV processing.
- Design adaptive support mechanisms to alleviate these difficulties.

This paper presents results related to the first two steps of this methodology, as well as guidelines on how to address the third step, grounded in these results. In the previous work (Toker et al. 2018), we reported a preliminary analysis on a user study we conducted with MSNVs. In that study, we measured a battery of nine different user characteristics in order to identify which ones play a significant role during MSNV processing, and found indications of user characteristics specifically impacting performance. In this paper, we expand that analysis and include eye tracking data that were collected during the study. Here, we first identify four user characteristics (*Verbal Working Memory*, *Reading Proficiency*, *Need for Cognition*, and *Verbal IQ*) for which users low in either of these abilities are at a disadvantage, in terms of either longer time on task or low accuracy. Next, we perform an analysis of gaze data aimed at identifying where significant differences in MSNV performance are occurring for each of these four user characteristics in terms of how the documents are visually processed. To accomplish this task, we generate numerous gaze metrics over distinct regions (i.e., *Areas of Interest,* or *AOIs* for short) of the MSNV documents, and then leverage Linear Mixed-Effects Models to identify significant relationships between user characteristics and gaze metrics that relate to low task performance.

The overall methodology adopted in this paper is inspired by previous work on designing user-adaptive support for visualization processing (Toker et al. 2013; Toker and Conati 2014), and allows us to clearly identify sub-optimal gaze processing behaviors of users with *low abilities* in user characteristics which contributes to their decreased task performance. Specifically, we identified several sub-optimal gaze processing behaviors shown by users with low measures of *Reading Proficiency* when they process the MSNV visualizations. These behaviors are captured by different elements of the visualizations (e.g., transitions between relevant and non-relevant bars) suggesting that low *Reading Proficiency* users could benefit from guidance specific to the multimodal nature of the MSNV as proposed in Folker et al. (2005).

The remainder of this paper is structured as follows; first we discuss related work, followed by a description of the user study. Next, we conduct an analysis of user experience with MSNVs to identify relevant user characteristics. We then describe gaze metrics that we computed from eye tracking data collected during the study, followed by an analysis of gaze metrics, relevant user characteristics, and MSNV performance. Lastly, we wrap up with a discussion and conclusion.

2 Related work

2.1 Relevant findings from psychology

There has been extensive work in psychology on investigating how people process combinations of textual and graphical information, mainly related to instructional text, with several findings supporting the intuitions that two media are better than one and that user characteristics should impact MSNV processing. For instance, Hegarty and Just (1993) showed that students who studied instructional material on pulley systems that contained both text and diagrams scored better on kinematic comprehension questions than those who studied an informationally equivalent version with only text. More tellingly for our work, the study also looked at the impact of two students' abilities (aptitude for reasoning with mechanical principles and reading ability) on both learning and gaze patterns when studying with the text and diagram material. Interestingly, no effect was found for reading ability, possibly because the participants were students at one of the top American universities and thus had all high reading abilities. Mechanical aptitude had no effect on learning outcomes, but a marginal effect on time taken to study the material (higher for low-ability students), which could be explained by the significant differences found in two specific gaze patterns: Low-mechanical-ability students re-read more clauses in the text and inspected the diagram more often. More recently, Wiley et al. (2014) present evidence that working memory capacity (WMC—a trait of individuals in relation to their ability to use their working memory system) can predict learning from illustrated text. They argue that lower WMC reduces a reader's ability to select specific information and integrate it to develop overall understanding, and they suggest various forms of personalized support for learners with low WMC. In this paper, we also consider two user characteristics related to working memory and study their impact on MSNV processing. Focusing on a different user trait, Kalyuga et al. (2013) argue that whether delivering instructional material by integrating two modalities increases comprehension or creates overload depends on the viewer's expertise. For instance, in a much earlier seminal work (Kalyuga et al.

1998) found that inexperienced electrical trainees learned better from textual explanations integrated into the diagrams of electrical circuits, whereas more experienced trainees performed better with the diagram only. In our study, we do not look at the user characteristic of domain expertise, but this could be a venue for future work.

2.2 User characteristics in visualization research

An accumulating amount of work has linked several user characteristics[1] to performance and preference with various types of information visualizations. For instance, the cognitive ability *perceptual speed* has been shown to correlate negatively with time on task while working with static grouped bar charts (Carenini et al. 2014), three-dimensional representations (Velez et al. 2005), as well as interactive stacked bar charts (Conati et al. 2014), and it can also influence visualization suitability among available alternatives (Allen 2000; Conati and Maclaren 2008). For the cognitive ability *visual working memory*, users with high levels of this ability were found to have a stronger preference for radar charts over bar charts (Toker et al. 2012), and were shown to prefer deviation charts over maps (Lallé et al. 2017). Findings linking other cognitive traits to visualization performance include: *disembedding* on task accuracy (Velez et al. 2005), *verbal working memory* on response time (Carenini et al. 2014; Conati et al. 2014), *spatial memory* on both task performance (Conati and Maclaren 2008; Velez et al. 2005) and visualization usability (Lallé et al. 2017), and *need for cognition* on task accuracy (Conati and Maclaren 2008). Even some personality traits, such as *locus of control*, have been shown to play a significant role in determining which layout of tree visualizations a user performs best with (Green and Fisher 2010; Ottley et al. 2015; Ziemkiewicz et al. 2011). All of these findings provide strong motivation for developing visualizations that are user-adaptive, i.e., visualizations that can support individual users by tailoring the interaction according to their relevant user characteristics. Generally speaking, the work presented in this paper is essentially broadening all this previous work on visualizations only to scenarios where users interact with visualizations embedded in narrative text.

2.3 User adaptation

Cuing, namely adding visual prompts that guide learners' attention to relevant elements in multimodal material, has been extensively investigated as a means to provide support (see van Gog 2014 for an overview) and has generated positive results for written text with graphics. For instance, Folker et al. (2005) and (Ozcelik et al. 2010) show that color-coding matching parts of the text and the graphics can increase comprehension. Yet, this approach can raise the issue of not having a sufficient number of easily distinguishable colors for color matching. Kalyuga (2009) sidestepped this problem by color matching corresponding parts of text and graphics dynamically. They gave to novice learners instructional material on an electric circuit, including both a diagram and a textual description. Attentional guidance was provided dynamically when student clicked on a specific paragraph by color-coding all the electrical elements

[1] Definitions of the user characteristics discussed in this sub-section are provided in Table 3 (Sect. 3.4).

mentioned, both in the text as well as in the diagram. Results showed that novices who received this guidance learned significantly more than those who studied the same material without it. Similarly, Carenini et al. (2013) proposes the concept of dynamic cuing for helping users process MSNVs, by guiding user attention to relevant parts of a graph as users read the corresponding textual reference (as detected via eye tracking), but they did not consider the impact to user characteristics as we do in this paper.

Carenini et al. (2014) evaluated several forms of dynamic highlighting to guide attention to relevant data points within grouped bar charts (stand-alone, i.e., not included in MSNVs) and showed a significant improvement in task performance compared to using no interventions, paving the way to the idea of effective cuing in MSNVs. As discussed in the introduction, Toker et al. (2018) conducted a preliminary investigation on whether user experience while processing MSNVs (comprehension, time on task, and subjective measures of satisfaction) depends on specific user cognitive abilities or traits. Here, we extend that work by further investigating the impact of user characteristics on MSNV processing, including an in-depth analysis of gaze patterns.

Several works have also shown the value of providing dynamic personalized guidance in processing visualizations systems. Guidance is provided either by proposing different visualizations based on detected user needs such as suboptimal behaviors (Gotz and Wen 2009) and evolving knowledge (Grawemeyer 2006), or by changing aspects on the current visualization (Nazemi et al. 2013). There is also initial research on providing dynamic guidance to reading. For instance, Loboda et al. (2011) leveraged eye tracking to ascertain the feasibility of inferring word relevance during reading tasks, to assess the informational needs of users and provided personalized content. Our work can be seen as building the foundations for extending personalized guidance to MSNVs reading.

In narrative visualization, the previous work has looked at automating the generation of new text and graphical presentations (Green et al. 2004), as well as identifying sentences in the narrative text to corresponding datapoints in the accompanying visualization(s) of existing documents via either crowdsourcing or natural language processing techniques (Metoyer et al. 2018). For now, in our work, we are assuming that the MSNVs are given with all the references annotated. Developing robust methods for generating novel MSNVs, or automatically extracting references from existing MSNVs, for adaptation is left as future work.

Other researchers have looked at supporting users while reading instructional texts by detecting instances of mind wandering and intervening to refocus user attention (D'Mello et al. 2017). However, to the best of our knowledge, no one has focused on the next step of designing user-adaptive support to help users process them.

2.4 Eye tracking in user modeling for information visualizations

Existing research has leveraged eye tracking data to perform a variety of user modeling tasks to facilitate the development of user-adaptive interfaces. Here, we focus on research examining users' gaze in order to understand the relationship between user characteristics and information visualization processing.

Several studies have shown significant differences in gaze patterns of experts and novices during visualization tasks in a variety of domains, including chemistry (e.g., Tai et al. 2006; Tang et al. 2012), general information search (Kules and Capra 2012), and geography (Çöltekin et al. 2010; Ooms et al. 2014). However, little work has been done to formally connect significant differences in gaze behaviors due to user characteristics, to objective measures of task performance. Building this connection is key in order to understand whether differences in users' gaze behaviors even have an impact on performance with the visualization (otherwise there is little guidance on how to provide meaningful adaptive support), and if they do, which ones help or hinder performance (so that the gaze behaviors can be encouraged or discouraged accordingly). To the best of our knowledge, there are only three recent works that have begun to address this research gap. Firstly, Ooms et al. (2012) examined performance differences between experts and novices in cartography, for search tasks with map visualizations. Using basic fixation data, they identified that experts had shorter fixations and higher fixation rate than novices, suggesting that experts' shorter response times were due, respectively, to their ability to interpret individual elements within the maps more efficiently and were able to scan the maps overall more efficiently. Secondly, Toker et al. (2013) carried out an analysis of gaze data to explore why performance differences occurred between users while carrying out low-level analysis tasks on bar and radar graph visualizations. In this work, they identified that users with low *perceptual speed,* who were slower on task, spent significantly more time looking at the legend and transitioned to it more frequently, indicating that these users were having difficulty processing and/or remembering the visualization's legend. They also found that users with low *verbal working memory,* who were slower on task, generated more fixations and spent significantly more time reading the textual description of each visualization task to be performed. These findings thus offer preliminary guidance on where user-adaptive support could be provided, namely by devising ways to help users with low *perceptual speed* process the legend, and similarly helping users with low *verbal working memory* process the textual description of each task. In the third recent work, Toker and Conati (2014) collected eye tracking data from a study using bar graphs and two types of low-level analysis tasks (simple and complex). In that work, they found that for complex tasks, users with low *perceptual speed* (who were slower with these tasks), spent significantly more of their time looking at the bar labels along the x-axis. They also identified that for complex tasks, users with low *visual working memory* (who were slower with these tasks) spent significantly more of their time looking at the list of possible answers for each task (multiple choice radio buttons), and also transitioned more frequently to them. These findings demonstrate the value of using eye tracking data to identify where potential adaptive support could be provided within the visualization interface. With the same goal in mind, the aim of the work we present in this paper is to utilize eye tracking data to carry out a similar investigation on how user characteristics that impact task performance are influencing MSNV processing.

3 MSNV user study

We have conducted an exploratory user study to collect data on how users process MSNVs. First, we present the study procedure, followed by a description of the MSNV documents that were generated for the study. Next, we explain the dependent variables measured in the study, and after we present details on the set of user characteristics that were collected.

3.1 Study procedure

The experiment was a within-subjects repeated measures design, lasting at most 115 min. Fifty-six subjects (32 female), ranging in age from 19 to 69, participated in the study. Sixty percent of participants were university students, and the others were from a variety of backgrounds (e.g., retail manager, restaurant server, retired). Raw gaze data were captured during our study using a Tobii T-120 eye tracker with the IV-T fixation filter (Olsen 2012), and was calibrated at the beginning of the study to each user. The computer screen display was 1280 × 1024 pixels.

Participants were given the task of reading over an MSNV document on the computer screen, and would signal they were done by clicking “next.” They were then presented with a set of questions on the screen designed to elicit their opinion of the document and to test their comprehension of relevant concepts discussed in it (see Sect. 3.3). Participants were required to carry out this task for 15 different MSNVs (described in Sect. 3.2). The ordering of the 15 MSNVs was randomized for each participant. Users were not given a time limit to read the MSNVs. However, to ensure that participants dedicated adequate effort to the task, they were told that there would be a $50 bonus for the three participants with the best performance, evaluated in terms of both speed and accuracy. The bonus was given in addition to the $45 we paid participants as compensation for the study.

Standard tests were used to assess the target battery of *nine* user characteristics (described in Sect. 3.4). The tests were split up so as to not fatigue users with too many tests all at the same time. Three of these tests (*Visualization Literacy, Need for Cognition, Verbal Working Memory*), which are computer-based and do not require an invigilator, were done at home prior to the experiment. A simple web-server was used to administer and record the test results accordingly. The other six user characteristic tests were administered in the laboratory: two before and four after the set of 15 MSNV tasks. The first two (*Visual Working Memory* and *Verbal IQ*) consisted of a computer test and a spoken test that both required specialized software. The last four (*Perceptual Speed, Reading Proficiency, Spatial Memory, Disembedding*) were all paper-based tests, and were completed consecutively at the end. The order of administration of tests was identical for all users.

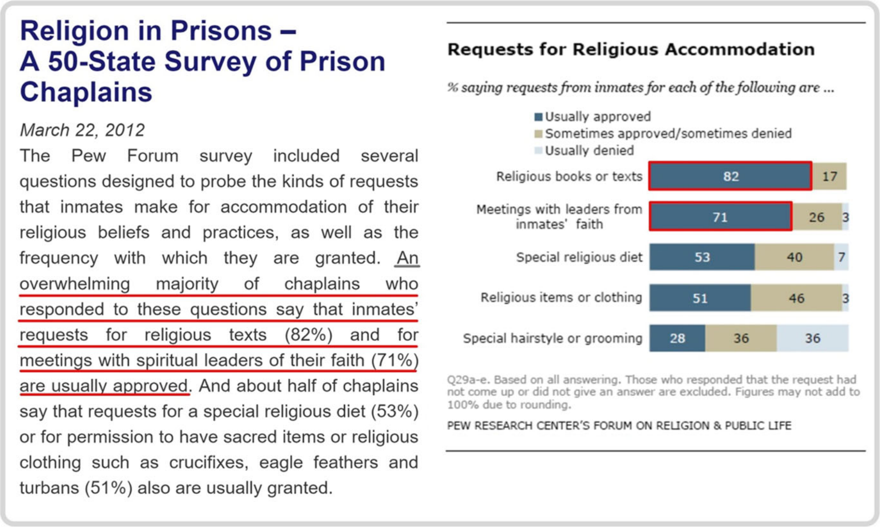
3.2 MSNVs used in the study

As we mentioned in the introduction, salient processing points in an MSNV are solicited by references, namely segments of text that specify a visual task on an accompanying visualization. The MSNVs we used for the study tasks were derived from an existing dataset of 40 magazine style documents extracted from real-world sources (e.g., *Pew Research*, *The Guardian*, and *The Economist*) where the references in each document had been previously identified via a rigorous coding process, indicating which data points in each visualization correspond to each reference sentence (Kong et al. 2014). Despite the obvious value of this dataset for our research, there were some issues with the format of the documents that we had to address. Each document in this dataset consisted of “snippets” of larger source documents, whereby each snippet included exactly one paragraph of text and one accompanying visualization. This simple document format was required to support the research purposes of Kong et al. (2014) to automate the extraction of references in each document utilizing crowdsourcing and clustering. Regrettably for our purposes, many of these document snippets were *fragmented*, i.e., the document or individual sentences within the document were difficult to comprehend because some of the required details were expressed in sentences from prior paragraphs in the original source material that were not included. We solved this problem by retrieving and adding the missing text from the source articles, to which we have access. In cases where fragmentation issues could not be resolved, the document snippet was removed. To provide more realism, we also added the original date and title to each document. We also identified several document snippets that had been derived from the same source article, and merged them into a single MSNV, respectively. Lastly, among the documents remaining after applying all of the above-mentioned changes, we selected a subset so as to have a varied number of words and references, to account for the potential influence that these factors of complexity might have on MSNV processing. We also selected documents to include a balanced variety of three bar chart types (i.e., simple, stacked, grouped (Munzner 2014)). We focused only on one class of visualizations to keep the complexity of the study manageable, and we chose bar charts because they are one of the most popular and effective visualizations for the common tasks of looking up and comparing values in simple tabular data (Munzner 2014). The end result of our work yielded a set of 15 self-contained MSNV documents, consisting of one visualization each, and one body of narrative text (see Fig. 2). Summary statistics on the composition of the 15 MSNV documents is provided in Table 1.

3.3 Dependent measures

The aim of our study was to evaluate the impact of user characteristics on users’ experience with MSNVs, where experience comprises of objective performance (time on task and comprehension) as well as subjective measures (MSNV ease-of-understanding and interest). Comprehension and subjective measures were assessed for each MSNV via a set of questions that we designed (see Fig. 3), which were shown to the user after they read each document.

Given that the MSNV documents are fairly short in length, we wanted to ensure that the number and types of questions we asked were not too long and would not be more difficult to process than the MSNVs themselves. First, we asked two *subjective* questions using a 5-point Likert scale to measure, respectively, perceived ease-of-understanding and interest (top two questions in Fig. 3), based on the work by Waddell et al. (2016), where they used a similar question format to capture users’ subjective attitudes toward *End User License Agreements*. Next, we asked *objective* questions to measure document comprehension, based on the work by Dyson and Haselgrove (2001), where they employed five different types of multiple choice questions for evaluating users’ comprehension of *National Geographic* articles. We designed questions



Religion in Prisons – A 50-State Survey of Prison Chaplains

March 22, 2012

The Pew Forum survey included several questions designed to probe the kinds of requests that inmates make for accommodation of their religious beliefs and practices, as well as the frequency with which they are granted. An overwhelming majority of chaplains who responded to these questions say that inmates' requests for religious texts (82%) and for meetings with spiritual leaders of their faith (71%) are usually approved. And about half of chaplains say that requests for a special religious diet (53%) or for permission to have sacred items or religious clothing such as crucifixes, eagle feathers and turbans (51%) also are usually granted.

Requests for Religious Accommodation
% saying requests from inmates for each of the following are ...
Usually approved
Sometimes approved/sometimes denied
Usually denied
Religious books or texts 82 17
Meetings with leaders from inmates' faith 71 26 3
Special religious diet 53 40 7
Religious items or clothing 51 46 3
Special hairstyle or grooming 28 36 36
Q29a-e. Based on all answering. Those who responded that the request had not come up or did not give an answer are excluded. Figures may not add to 100% due to rounding.
PEW RESEARCH CENTER'S FORUM ON RELIGION & PUBLIC LIFE

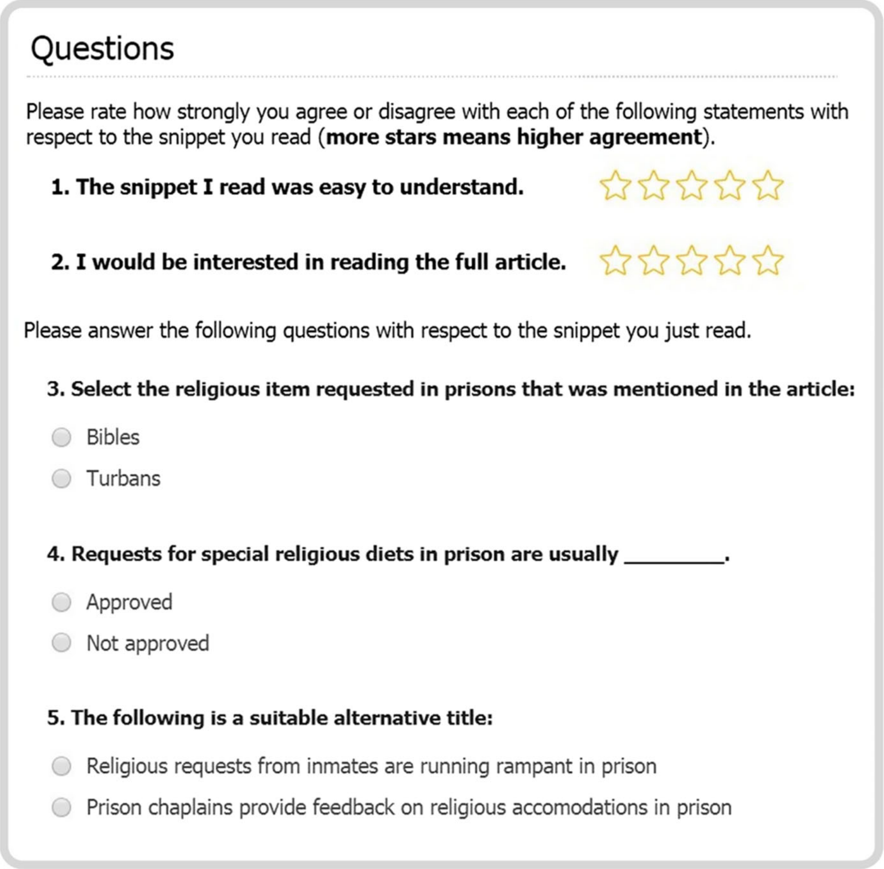
Fig. 2 One of the 15 MSNVs administered in the user study. *Note: Red highlighting is shown to illustrate the concept of a *reference*. Highlighting was not provided to users in the study. (Color figure online)

Table 1 Summary statistics illustrating the variety of document characteristics across the 15 MSNVs administered in the user study

MSNV property	Min	Max	Median	Mean	SD
Words: *Total number of words in the body of narrative text.*	43	228	75	90.8	49.7
Sentences: *Total number of sentences in the body of narrative text.*	2	14	4	4.9	3.0
Reference sentences: *Total number of sentences in the body of narrative text that specify a visual task on the visualization.*	1	7	2	2.6	1.8
Datapoints in viz: *Total number of data points in the visualization.*	4	63	14	22.1	19.7
Reference targets in viz: *Total number of data points in the visualization mentioned by any reference sentences.*	2	24	6	10.1	7.8

based on two of their question types, chosen because both types of questions could be asked for all the MSNVs in our dataset. The two question types we selected were:

- One *title question* which asks to select a suitable alternative title for the MSNV (see question 5, bottom of Fig. 3), and provides a simple way to ensure that the user had a grasp of the general document narrative.
- One or two (depending on document length) *recognition questions* asking to recall specific information from the MSNV: identifying a named entity discussed in the text (e.g., question 3 in Fig. 3), or identifying the magnitude/directionality of a named entity discussed in the text (e.g., question 4 in Fig. 3). For most documents, two recognition questions were asked. When the document was too short to provide



Questions

Please rate how strongly you agree or disagree with each of the following statements with respect to the snippet you read (**more stars means higher agreement**).

1. The snippet I read was easy to understand. ☆☆☆☆☆

2. I would be interested in reading the full article. ☆☆☆☆☆

Please answer the following questions with respect to the snippet you just read.

3. Select the religious item requested in prisons that was mentioned in the article:

○ Bibles
○ Turbans

4. Requests for special religious diets in prison are usually _________.

○ Approved
○ Not approved

5. The following is a suitable alternative title:

○ Religious requests from inmates are running rampant in prison
○ Prison chaplains provide feedback on religious accomodations in prison

Fig. 3 Subjective and comprehension questions presented to users after reading each MSNV document. Note: Users were not allowed to proceed without answering all of the questions

enough content for generating two questions, only one recognition question was asked.

In total, we generated four dependent measures (two subjective and two objective) that capture user experience with each MSNV document. The first three dependent measures are calculated from the set of questions described above, and include:

- *MSNV Ease-of-Understanding* Subjective rating on a 5-point Likert scale.
- *MSNV Interest* Subjective rating on a 5-point Likert scale.
- *MSNV Accuracy* Accuracy (% correct) of the comprehension questions.

The fourth dependent measure was logged during MSNV processing, and consists of:

- *MSNV Time on Task* Time (seconds) spent on the MSNV document.

Table 2 provides summary statistics on each of the four measures of user experience collected during the study.

Table 2 Summary statistics of the four measures of MSNV user experience obtained from the study

Measure	Min	Max	Median	Mean	SD
Time on Task (sec)	7.8	296.5	49.75	57.91	33.2
Comprehension Accuracy (%)	0	1	0.67	0.69	0.30
Ease-of-Understanding (1–5)	1	5	4	3.9	1.03
Document Interest (1–5)	1	5	3	3.7	1.26

3.4 User characteristics

We measured nine different user characteristics in the study using standard tests in psychology, defined in Table 3. The first seven characteristics consist of cognitive abilities and traits that we selected because previous research has shown that they play a significant role in user experience with visualizations. For instance, *Perceptual Speed, Visual Working Memory, Verbal Working Memory, Visualization Literacy, and Spatial Memory* were chosen because previous studies have shown that they can impact visualization preference and task performance with bar chart visualizations (Toker et al. 2012; Carenini et al. 2014; Conati et al. 2014; Lallé et al. 2017), which are also the types of visualizations in our MSNVs. We also included the user characteristic *Need for Cognition* because previous work has shown that it can influence accuracy with visualization search tasks (Conati and Maclaren 2008), and also because we hypothesized that it may play a role in how much effort users were willing to invest in reading the MSNV documents given that a minimum and maximum time limit was not enforced in the study. Although previous research examining the link between *Disembedding* and visualization performance is limited (e.g., Velez et al. 2005), we opted to include this user characteristic in our study because of the *references* contained in the MSNV documents. Specifically, we hypothesized that the act of processing any of the reference sentences may require some level of disembedding, namely identifying groups of bars of interest amidst the full set of datapoints contained in the visualization.

In addition, we included two user characteristics relating to reading comprehension ability, to account for potential performance differences arising due to reading the body of narrative text contained in each MSNV. Unfortunately, assessing reading comprehension ability can be a very time-consuming endeavor. For instance, standard tests such as the *ESOL*, *IELTS*, and *TOEFL iBT* require more than an hour to administer, which was not feasible for our user study. Hence, we selected two tests that could each be administered in under 5 min and have been shown to reliably approximate two different constructs relating to reading comprehension ability: *Reading Proficiency* (Meara and Jones 1990) and *Verbal IQ* (Blair and Spreen 1989).

Lastly, we report in Table 4 several statistics on the nine user characteristics test results, collected from the 56 users in our study. We also report in Table 5 pairwise correlation scores among the user characteristics to provide a sense of how well they are each capturing complementary or non-overlapping dimensions of user abilities. Since Shapiro–Wilk normality tests revealed that none of the nine user characteristics were normally distributed ($p < .001$), we used a nonparametric test, Kendall's tau (τ)

Table 3 The set of nine user characteristics measured in the study

User characteristic	Definition	Instrument
Need for Cognition	Extent to which individuals are inclined toward effortful cognitive activities (Cacioppo et al. 1984)	*Need for Cognition Scale* (Cacioppo et al. 1984), a questionnaire asking users to rate their agreement (5-point Likert scale) with 18 statements about the satisfaction they gain from thinking in various scenarios. Final scores range from −36 to 36
Visualization Literacy	Ability to confidently use a visualization to translate questions specified in the data domain into visual queries in the visual domain, as well as interpreting visual patterns in the visual domain as properties in the data domain (Boy et al. 2014)	*Visualization Literacy 101 – Bar Chart Test* (Boy et al. 2014), a computer-based test where users perform a series of standard benchmark visualization tasks (e.g., finding min/max, estimating average, detecting trends) with bar chart visualizations. Final scores range from range from −2.0 to 1.0, and are computed based on accuracy and time taken
Visual Working Memory	Measures the quantity of visual information (e.g., shapes and colors) that can be temporarily maintained or manipulated in working memory (Logie 2009)	*Colored Squares Sequential Comparison Task (uncued)* (Vogel et al. 2001), a computer-based test where users are briefly shown a sample array of *n* colored squares, then after a short blink delay, a single colored square appears and participants indicate (yes/no) if its color matches one in the sample array. This task repeats 120 times over three different array sizes ($n = 4, 6, 8$). Final scores range from 0 to 6 by averaging the scores obtained from each array size
Spatial Memory	Ability to remember the configuration, location, and orientation of figural material (Ekstrom et al. 1976)	*MV-1 Shape Memory Test* (Ekstrom et al. 1976), a timed paper-based test that requires users to first study a page filled with abstract shapes, and afterward recall the relative positions of specific subsets of shapes. Final scores range from 0 to 16
Verbal Working Memory	Measures the quantity of verbal information (e.g., words) that can be temporarily maintained and manipulated in working memory (Baddeley 1986)	*OSPAN (Operation-word span)* (Turner and Engle 1989), a short computer-based test where users are briefly shown a list of 1–6 words, then respond to a basic arithmetic operation, and afterward are asked to recall the list of words. Final scores range from 0 to 6

Table 3 continued

User characteristic	Definition	Instrument
Perceptual Speed	Speed in scanning/comparing figures or symbols, or carrying out other very simple tasks involving visual perception (Ekstrom et al. 1976)	*P-3 Identical Pictures Test* (Ekstrom et al. 1976), a timed paper-based test that measures how quickly users can locate matching objects amidst a set of distractors. Final scores range from 0 to 72
Disembedding	Ability to hold a given visual percept or configuration in mind so as to disembed it from other well-defined perceptual material (Ekstrom et al. 1976)	*CF-2 Hidden Patterns Test* (Ekstrom et al. 1976), a timed paper-based test that requires users to identify (i.e., disembed) if a given figure is hidden among other lines. Final scores range from 0 to 300
Reading Proficiency	Vocabulary size and reading comprehension ability in English (Meara 2010)	*X_Lex Vocabulary Test* (Meara 2010), an untimed paper-based test. Users indicate on a vocabulary list (yes/no) if they know the meaning of each word. Some words are fake, and users are not told this. Final scores range from 0 to 100, based on the # of hits (word exists and the user indicates they know it) and false alarms (user indicates they know the meaning of a fake word)
Verbal IQ	Overall verbal intellectual abilities that measure acquired knowledge, verbal reasoning, and attention to verbal materials (Blair and Spreen 1989)	*North American Adult Reading Test (NAART)* (Strauss et al. 2006), an untimed spoken test where users are asked to read aloud a series of increasingly difficult words in English. The total number of incorrectly pronounced words is then used to compute the user's *VerbalIQ*, with possible scores ranging from 74.41 to 128.7

Table 4 Summary statistics showing the range of user characteristics scores measured in the study

User characteristic	Min	Max	Median	Mean	SD
Need for Cognition	−20	26	12.5	10.6	10.2
Visualization Literacy	−2.1	1.0	0.47	0.30	0.71
Visual Working Memory	0	5	2.5	3.8	1.0
Spatial Memory	1	14	8	7.6	3.4
Verbal Working Memory	2	6	5	5.0	1.1
Perceptual Speed	25	66	45	45.2	8.9
Disembedding	12	84	61.5	57.6	15.3
Reading Proficiency	54.7	96.3	84.9	83.4	9.7
Verbal IQ	84.2	122.5	101.1	101.6	8.9

(as opposed to a standard Pearson's r). In general, the pairwise correlations are small or medium (i.e., $\tau \sim 0.19$ or smaller).[2] There are three exceptions. Two involve perceptual abilities, namely *Visual Working Memory* and *Visualization Literacy* ($\tau = 0.37$) and *Perceptual Speed* and *Disembedding* ($\tau = 0.38$). These higher correlations are likely due to the fact that some parts of each test for *Visualization Literacy* and *Disembedding* reply on lower-level perceptual abilities. In particular, the test for *Visualization Literacy* requires processing different colored bars to create mappings to their corresponding entities, a sub-task that leverages users' *Visual Working Memory* which measures the quantity of colors that can be temporarily maintained or manipulated in working memory; and similarly because the test for *Disembedding* requires users to repeatedly match embedded shapes, a task that leverages users' *Perceptual Speed* which measures how quickly users can scan figures or symbols. We opted to retain all of the above user characteristics because, despite the partial overlaps, none can be removed without losing information relating to specific scope of the perceptual abilities each test is designed to capture. The third exception relates to the two characteristics that we used to measure users' reading comprehension abilities: *Reading Proficiency* and *Verbal IQ* ($\tau = 0.27$). Since reading comprehension ability is comprised of and can be assessed according to several different measurable factors (Grabe and Jiang 2013) (in our case vocabulary size and pronunciation ability, respectively), it is not surprising that there is some overlap between these two measures, but the correlation is only partial, indicating they are each still capturing distinct information. As with the previous two correlations with perceptual abilities, we opt to keep both reading ability measures to retain as much information as possible relating to the specific factors each measure captures.

4 Effects of user characteristics on MSNV user experience

In this section, the goal is to perform the first step toward designing user-adaptive support for MSNV processing, by identifying which user characteristics are impacting

[2] Using the guidelines from Field (2003) that $r = 0.10$ is a small correlation, $r = 0.30$ is medium, and $r = 0.50$ is large, we computed the Kendall's τ equivalent according to Walker (2003), yielding: $\tau = 0.06$ small association, $\tau = 0.19$ medium, and $\tau = 0.33$ large.

Table 5 Kendall's Tau Correlation scores between the all of the user characteristics

	Need for Cognition	*Visualization Literacy*	*Visual Working Memory*	*Spatial Memory*	*Verbal Working Memory*	*Perceptual Speed*	*Disembedding*	*Reading Proficiency*	*Verbal IQ*
Need for Cognition									
Visualization Literacy	0.16								
Visual Working Memory	0.21	0.37							
Spatial Memory	– 0.01	0.14	0.09						
Verbal Working Memory	0.11	0.02	0.04	0.10					
Perceptual Speed	0.08	0.03	0.18	0.26	0.13				
Disembedding	0.13	0.12	0.14	0.21	0.15	0.38			
Reading Proficiency	– 0.07	0.02	0.05	– 0.18	0.06	– 0.03	– 0.13		
Verbal IQ	– 0.11	0.06	0.04	0.01	– 0.02	0.10	0.10	0.27	

which measures of MSNV user experience and therefore may warrant further investigation for providing adaptive support.[3] We first describe the statistical analysis used and then summarize the obtained results. After, we explain based on the results which user characteristics we select for further investigation.

4.1 Analysis and results

To carry out the analysis, we use Linear Mixed-Effects Models; an alternative to using a traditional repeated measures ANCOVA. We opted for Mixed Models, since they can model multiple random effects at once. For our purposes, the specification of two random effects is required since our study was a repeated measures design where all users were exposed to the same set of 15 documents. The first random effect *user_id* accounts for a within-subject correlation (i.e., due to non-independence), since multiple measurements are collected from the same user. The second random effect *MSNV_id* accounts for a within-document correlation, since repeated measurements are collected from the same MSNV document. We used the lmerTest software package in R (Kuznetsova et al. 2017) and constructed one Mixed Model for each measure of MSNV user experience (described in Sect. 3.3) as the dependent measure, along with the nine user characteristics as covariates (described in Sect. 3.4), and *user_id* and *MSNV_id* as random effects. For each model, we run a bi-directional stepwise algorithm for model selection defined by Akaike Information Criteria (AIC) (Akaike 1974). The two subjective dependent measures (*Ease-of-Understanding* and *Document Interest*) were collected using a standard 5-point Likert scale. Shapiro–Wilk normality tests revealed that these two measures were not normally distributed ($p < .001$), therefore we applied the Aligned Rank Transformation (Wobbrock et al. 2011), to convert them to a normal distribution. Significant results obtained are reported in Table 6.

We identified main effects for five user characteristics on objective measures of MSNV performance, but no main effects of user characteristics were found on the two subjective measures of MSNV experience (see Table 6). Based on these results, we group our findings into three different categories of users:

- The first group includes two user characteristics (V*erbal Working Memory* and *Reading Proficiency*) that display a **negative** directionality with Time on Task (as shown by the negative slope of the model coefficients *b*). Since no significant results were found for these two user characteristics on Comprehension Accuracy, it indicates that users **low** in these abilities spend more time looking at the MSNV to achieve comparable accuracy as their counterparts. The most straightforward explanation as to why these users are struggling is precisely because of their low abilities in either of these two user characteristics.
- The second group of main effects includes user characteristics (*Need for Cognition* and *Verbal IQ*) that display a **positive** directionality with Comprehension Accuracy (as shown by the positive slope of the model coefficients *b*). Since no significant results were found for these two user characteristics on Time on Task, it indicates that users **low** in these abilities are spending comparable time as their counterparts on

[3] The current analysis overrides the analysis described in Toker et al. (2018) that contained a statistical flaw.

Table 6 Results indicating which user characteristics have a significant effect on measures of MSNV user experience. The normalized model coefficient b indicates the size and directionality of the relationship

Main Effect of User Characteristic	Time on Task	Comprehension Accuracy	Ease-of-Understanding	Document Interest
Verbal Working Memory	$p < .05$ $b = -0.08$	*not sig.*	*not sig.*	*not sig.*
Reading Proficiency	$p < .05$ $b = -0.09$	*not sig.*	*not sig.*	*not sig.*
Visualization Literacy	$p < .01$ $b = 0.13$	$p < .01$ $b = 0.13$	*not sig.*	*not sig.*
Need for Cognition	*not sig.*	$p < .05$ $b = 0.08$	*not sig.*	*not sig.*
Verbal IQ	*not sig.*	$p < .05$ $b = 0.07$	*not sig.*	*not sig.*
Visual Working Memory	*not sig.*	*not sig.*	*not sig.*	*not sig.*
Spatial Memory	*not sig.*	*not sig.*	*not sig.*	*not sig.*
Perceptual Speed	*not sig.*	*not sig.*	*not sig.*	*not sig.*
Disembedding	*not sig.*	*not sig.*	*not sig.*	*not sig.*

the task, but end up achieving lower accuracy. Once again, the most straightforward explanation as to why these users are struggling is precisely because of their low abilities in either of these two user characteristics.

- The third group only includes *Visualization Literacy*, for which there is a **positive** directionality with both Time on Task and Accuracy (as shown by the positive slopes of the model coefficients b). This result indicates that users with **high** *Visualization Literacy* are slower on task, but the accuracy is improved, which is a standard tradeoff.

In the next section, we present an analysis of gaze data aimed at explaining our findings in terms of how the documents are visually processed. We focus on the first and second categories, which include (*Verbal Working Memory* and *Reading Proficiency)* and *(Need for Cognition* and *Verbal IQ).* For these four characteristics, there was a clear indication that users with low abilities are not using their time effectively (i.e., they either needed more time to achieve comparable accuracy as the high ability users, or given the same time they achieved lower accuracy). For such users, eye tracking data could arguably explain the sources of these inefficiencies, revealing sub-optimal behaviors and then providing ideas for adaptation. The analysis of the findings about *Visualization Literacy,* for which a similar relation to eye tracking data appears to be less straightforward, is left as future work.

5 Eye tracking analysis and results

In this section, we leverage eye tracking data to investigate where users with low user characteristics are struggling during MSNV processing. First, we present in Sect. 5.1

details on how the eye tracking data collected during the user study were utilized to generate numerous gaze metrics that capture various MSNV processing behaviors. Next, in Sect. 5.2 we carry out an exploratory analysis on the set of generated gaze metrics, to identify which among them are relevant to performance with the MSNVs (i.e., *time on task* followed by *comprehension accuracy*). After, we conduct a further exploratory analysis in Sect. 5.3 to identify which of the gaze metrics relevant to MSNV performance are significantly influenced by user characteristics. Based on these results, we further refine our analysis by looking at finer-grained gaze metrics (Sect. 5.4) and specific performance metrics and user characteristics (Sects. 5.5 and 5.6).

5.1 Generating 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 (Toker and Conati 2014; Toker et al. 2013; D'Mello et al. 2012; Martínez-Gómez and Aizawa 2014). We processed users' raw gaze data using EMDAT (www.github.com/ATUAV/EMDAT), an open source library written in Python and developed in our research laboratory. EMDAT 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 is listed in Table 7. These metrics are defined over four AOIs that capture users' gaze activity within different regions of the MSNV documents (see Fig. 4). These AOIs were defined to gain a general sense of MSNV document processing according to the two primary forms of information contained in the MSNV documents, namely two AOIs for the *textual information* (block of text on the left), and to two AOIs for the *visual information* (the area including the visualization on the right). The four AOIs 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).

5.2 Identifying gaze metrics relevant to MSNV performance

Our goal here is to identify which gaze metrics have a significant relationship with MSNV *time on task* and MSNV *comprehension accuracy.* For the purposes of our research, gaze metrics that do not have any significant relationship to task performance are non-relevant and do not warrant further consideration. Non-relevant gaze metrics

Table 7 Set of 17 gaze metrics generated for each of the four AOIs. These metrics are generated by EMDAT for each user and each task

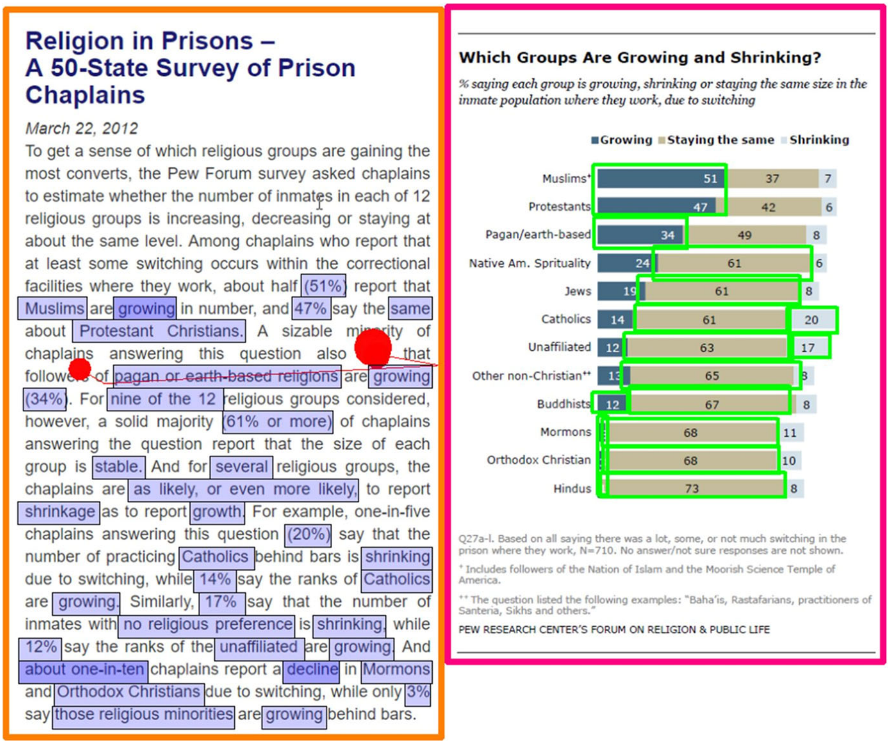
No.	Metric	Description
1	• *fixation_rate*	Fixation rate in AOI
2	• *number_of_fixations*	Total number of fixations in AOI
3	• *longest_fixation*	Longest fixation in AOI
4–6	• *sum_fix_durations* • *mean_fix_durations* • *stddev_fix_durations*	Sum, Mean, and Std. Deviation of fixation durations in AOI
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 fixations in all AOIs)

offer no concrete indication on how the captured processing behavior translates to MSNV performance, and thus provide little guidance toward designing meaningful adaptive support.

First, we checked for correlations among gaze metrics within each AOI. Recall there are a total of 68 gaze metrics (17 gaze metrics × 4 AOIs, c.f. Table 7). Shapiro–Wilk normality tests revealed that *time_to_first_fix* on the Text AOI was not normally distributed ($p < .001$), and therefore we removed it. This measure was skewed heavily to the right and captured very little variability likely because the Text AOI was usually the first place users looked at the outset of each MSNV task. Pearson correlations on gaze metrics within each of the 4 AOI groups revealed very high correlations ($r > 0.9$) in all four AOIs among: *sum_fix_durations, number_of_fixations*, and *transitions_to_self;* as well as *longest_fixation* and *stddev_fix_durations*. Based on these high correlations, we removed three gaze metrics for each AOI: *number_of_fixations*, *transitions_to_self*, and also *stddev_fix_durations* from further analysis. Correlations of gaze metrics were not checked between AOIs because our goal is to investigate how different areas of the MSNV documents are processed, and we wanted to preserve the ability to report and discuss results at the granularity of each AOI. Therefore, the total number of gaze metrics we retain for further investigation is 55: (14 metrics × 4 AOIs) − 1 metric in the Text AOI.

5.2.1 Gaze metrics relevant to time on task

In order to identify gaze metrics that have a significant relationship to *time on task*, we conduct an analysis using Mixed Models (see description in Sect. 4.1). For each



Religion in Prisons – A 50-State Survey of Prison Chaplains

March 22, 2012

To get a sense of which religious groups are gaining the most converts, the Pew Forum survey asked chaplains to estimate whether the number of inmates in each of 12 religious groups is increasing, decreasing or staying at about the same level. Among chaplains who report that at least some switching occurs within the correctional facilities where they work, about half (51%) report that Muslims are growing in number, and 47% say the same about Protestant Christians. A sizable minority of chaplains answering this question also say that followers of ~~pagan or earth-based religions~~ are growing (34%). For nine of the 12 religious groups considered, however, a solid majority (61% or more) of chaplains answering the question report that the size of each group is stable. And for several religious groups, the chaplains are as likely, or even more likely, to report shrinkage as to report growth. For example, one-in-five chaplains answering this question (20%) say that the number of practicing Catholics behind bars is shrinking due to switching, while 14% say the ranks of Catholics are growing. Similarly, 17% say that the number of inmates with no religious preference is shrinking, while 12% say the ranks of the unaffiliated are growing. And about one-in-ten chaplains report a decline in Mormons and Orthodox Christians due to switching, while only 3% say those religious minorities are growing behind bars.

Which Groups Are Growing and Shrinking?

% saying each group is growing, shrinking or staying the same size in the inmate population where they work, due to switching

Q27a-l. Based on all saying there was a lot, some, or not much switching in the prison where they work, N=710. No answer/not sure responses are not shown.

* Includes followers of the Nation of Islam and the Moorish Science Temple of America.

** The question listed the following examples: "Baha'is, Rastafarians, practitioners of Santeria, Sikhs and others."

PEW RESEARCH CENTER'S FORUM ON RELIGION & PUBLIC LIFE

Fig. 4 The four AOIs we defined to capture MSNV processing, shown here for one of the documents administered in the user study

of the 55 gaze metrics, we construct one Mixed Model, using gaze metric as the independent measure, with *time on task* as the dependent measure, and *user_id* and *MSNV_id* as random effects (i.e., repeated measures). Due to the exploratory nature of our analysis, we account for multiple comparisons using the Benjamini–Hochberg procedure to control for the false discovery rate (Benjamini and Hochberg 1995). The obtained p-values from our models are ordered from smallest to largest, such that the smallest p value has a rank of $i = 1$, the next smallest has $i = 2$, etc. Then we compare each individual p-value to its Benjamini–Hochberg critical threshold of $q = (i/m)\alpha$, where i is the rank, m is the total number of models, and α is set to 0.05. Next, we find the largest p-value that has $p < q$ given its rank r, and then, all p-values at rank i $\leq r$ are also considered significant. Applying the Benjamini–Hochberg procedure to our results yielded a critical threshold of $q = 0.0273$, obtained at rank $r = 30$. Thus, our analysis revealed 30 relevant gaze metrics that have a significant relationship with time on task, listed in Table 8, and in all cases (as indicated by slope of the model coefficient b) have positive correlation with time on task, except for one metric: *fixation_rate* in the Text AOI. 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

Table 8 For each AOI, we report gaze metrics that were found to be significant with *time on task*. The normalized model coefficient b indicates the size and directionality of the relationship. Endash (–) cells indicate metrics excluded from the analysis due to either lack of normality, i.e., *time_to_first_fix* for Text AOI; or high correlation, i.e., transitions to self (the same AOI)

Gaze Metric	Text AOI	Refs AOI	Viz AOI	R-Bars AOI
sum_fix_durations	$p < .001$ $b = 0.76$	$p < .001$ $b = 0.67$	$p < .001$ $b = 0.55$	$p < .001$ $b = 0.30$
longest_fixation	$p < .001$ $b = 0.24$	$p < .001$ $b = 0.23$	$p < .001$ $b = 0.19$	$p < .001$ $b = 0.19$
time_to_first_fix	–	$p < .001$ $b = 0.20$	$p = .016$ $b = 0.06$	$p < .001$ $b = 0.20$
time_to_last_fix	$p < .001$ $b = 0.86$	$p < .001$ $b = 0.76$	$p < .001$ $b = 0.79$	$p < .001$ $b = 0.58$
mean_fix_durations	$p = .001$ $b = 0.18$	$p < .001$ $b = 0.14$	*not sig.*	$p < .001$ $b = 0.18$
fixation_rate	$p < .001$ $b = -0.22$	*not sig.*	*not sig.*	*not sig.*
prop_num_fixations	*not sig.*	*not sig.*	*not sig.*	*not sig.*
transitions_to_Text	–	$p < .001$ $b = 0.87$	$p < .001$ $b = 0.31$	$p < .001$ $b = 0.11$
transitions_to_Refs	$p < .001$ $b = 0.86$	–	$p < .001$ $b = 0.15$	*not sig.*
transitions_to_Viz	$p < .001$ $b = 0.29$	$p < .001$ $b = 0.11$	–	$p < .001$ $b = 0.30$
transitions_to_R-Bars	$p < .001$ $b = 0.15$	$p = .001$ $b = 0.07$	$p < .001$ $b = 0.29$	–

what extent any of these relationships are qualified by the user characteristics *Verbal Working Memory* and *Reading Proficiency*.

5.2.2 Gaze metrics relevant to comprehension accuracy

For each of the 55 relevant gaze metrics, we construct one Mixed Model, using gaze metric as the independent measure, with *comprehension accuracy* as the dependent measure, and *user_id* and *MSNV_id* as random effects. Our analysis revealed only three gaze metrics with $p < .05$ on task accuracy, listed in Table 9. However, after applying the Benjamini–Hochberg procedure to adjust for multiple comparisons, neither of these three results were found to be significant (at best they could be considered marginally significant). It is surprising to see that unlike *time on task* (reported in the previous subsection), the collection of gaze metrics we evaluated has very little or no relationship with *comprehension accuracy*, with only a marginal indication that some processing behaviors captured in the R-Bars AOI of the MSNVs may play a role toward users' comprehension. Ultimately, since we were unable to identify statistically significant relationships for any of the gaze metrics with *comprehension accuracy*, no subsequent

Table 9 No gaze metrics were found to have a significant relationship with *comprehension accuracy*. Three metrics in the R-Bars AOI yielded *p*-values < .05; however, none remained statistically significant after correcting for multiple comparisons. Endash (–) cells indicate metrics that were excluded from the analysis due to either lack of normality, i.e., *time_to_first_fix* for Text AOI; or high correlation, i.e., transitions to self (the same AOI)

Gaze Metric	Text AOI	Refs AOI	Viz AOI	R-Bars AOI
sum_fix_durations	*not sig.*	*not sig.*	*not sig.*	*not sig.*
longest_fixation	*not sig.*	*not sig.*	*not sig.*	*not sig.*
time_to_first_fix	–	*not sig.*	*not sig.*	*not sig.*
time_to_last_fix	*not sig.*	*not sig.*	*not sig.*	$p = .026$ $b = 0.15$
mean_fix_durations	*not sig.*	*not sig.*	*not sig.*	$p = .036$ $b = 0.20$
fixation_rate	*not sig.*	*not sig.*	*not sig.*	$p = .039$ $b = 0.14$
prop_num_fixations	*not sig.*	*not sig.*	*not sig.*	*not sig.*
transitions_to_Text	–	*not sig.*	*not sig.*	*not sig.*
transitions_to_Refs	*not sig.*	–	*not sig.*	*not sig.*
transitions_to_Viz	*not sig.*	*not sig.*	–	*not sig.*
transitions_to_R-Bars	*not sig.*	*not sig.*	*not sig.*	–

analysis of gaze metrics will be carried out for *Need for Cognition* and *Verbal IQ*, since these two user characteristics were only found to impact *comprehension accuracy*.

5.3 Impact of user characteristics on gaze metrics relevant to time on task

As discussed in Sect. 4, we found two user characteristics (*VerbalWM* and *ReadingP*) which impact performance with MSNVs in a manner that may call for personalized support, namely users with low values of either of these two characteristics were spending significantly more time on task to achieve comparable accuracy compared to users with higher values. Here, our goal is to see whether any of these user characteristics (UC) impacts any of the 30 gaze metrics relevant to *time on task* (identified in Sect. 5.2.1), so as to detect possible sub-optimal gaze processing behaviors of users with low abilities in these user characteristics. We construct one Mixed Model for each of the 30 relevant gaze metrics as the dependent measure, with both UCs as covariates, and *user_id* and *MSNV_id* as random effects. We apply the Benjamini–Hochberg procedure to our results, yielding a critical threshold of $q = 0.0166$, obtained at rank $r = 10$. Significant results are reported in Table 10. The structure of Table 10 is designed to facilitate understanding which user characteristics have a significant effect on gaze metrics that belong to the *same AOI* (looking by column), as well as which user characteristics have a significant effect on the same type of gaze metrics *across all four AOIs* (looking by row). Thus, the rows in Table 10 list the type of gaze metric, the columns list the four AOIs on which these metrics are generated,

Table 10 Results showing in which AOIs a significant effect of user characteristics was found on the corresponding gaze metric. The normalized model coefficient b indicates the size and directionality of the relationship. Endash (–) cells indicate gaze metrics non-relevant to time on task, and were thus not evaluated

Gaze Metric	Text AOI	Refs AOI	Viz AOI	R-Bars AOI
sum_fix_durations	***VerbalWM*** $p = .012$ $b = -0.11$	*not sig.*	***ReadingP*** $p = .002$ $b = -0.15$	***ReadingP*** $p = .003$ $b = -0.12$
longest_fixation	*not sig.*	*not sig.*	***ReadingP*** $p = .01$ $b = -0.11$	***ReadingP*** $p = .003$ $b = -0.15$
time_to_first_fix	–	***VerbalWM*** $p = .003$ $b = -0.07$	*not sig.*	*not sig.*
time_to_last_fix	*not sig.*	*not sig.*	***ReadingP*** $p = .01$ $b = -0.13$	***ReadingP*** $p = .005$ $b = -0.16$
mean_fix_durations	*not sig.*	*not sig.*	–	*not sig.*
fixation_rate	*not sig.*	–	–	–
prop_num_fixations	–	–	–	–
transitions_to_Text	–	*not sig.*	*not sig.*	*not sig.*
transitions_to_Refs	*not sig.*	–	*not sig.*	–
transitions_to_Viz	*not sig.*	*not sig.*	–	***ReadingP*** $p = .007$ $b = -0.13$
transitions_to_R-Bars	*not sig.*	*not sig.*	***ReadingP*** $p = .006$ $b = -0.14$	–

and a cell (*i,j*) lists all (if any) UCs mediated by the gaze metric in row *i* generated over the AOI in column *j*. The model coefficient *b* listed under each UC indicates the directionality of the effect that this UC has on the corresponding gaze metric. For instance, the negative *b* in the first cell of Table 10 indicates a negative directionality, namely that users with low *VerbalWM* spend more time looking at the Text AOI compared to users with high *VerbalWM*.

Table 10 shows several results for both *VerbalWM* and *ReadingP*. It is interesting to see the distinct roles that *VerbalWM* and *ReadingP* each play when examining the table by column (recall too that these two UCs are virtually uncorrelated, c.f. Table 5 in Sect. 3.4). First, there are no main effects of *VerbalWM* on visualization processing; it only appears for textual AOIs (i.e., Text AOI and Refs AOI, first two columns of Table 10). Specifically, users with low *VerbalWM* are spending significantly more time (*sum_fix_durations*) processing the Text AOI and their first encounter with the textual references (*time_to_first_fix*) on Refs AOI are significantly later in the task compared to users with high *VerbalWM*, likely because they are having issues reading the text. Our findings regarding this connection between *VerbalWM* and textual processing mirror results found in the previous work (Toker and Conati 2014; Toker et al. 2013) where textual elements consisted of text in the visualization's legend, as well as sentences below the visualization eliciting the task to be carried out. Our results thus extend these previous findings on *VerbalWM* to include accompanying bodies of narrative text.

In contrast, results for *ReadingP* are entirely related to visualization processing (i.e., Viz AOI and Bars AOI, shown in the last two columns of Table 10). For instance, users with low *ReadingP* are spending significantly more time (*sum_fix_durations*) processing the visualization and relevant bars, have higher values for their longest fixations (*longest_fixation)*, and their last fixations (*time_to_last_fix*) are significantly later in the task compared to users with high *ReadingP*. Increased processing of the visualization by users with low *ReadingP* is also captured by additional back-and-forth transitions between the visualization and relevant bars (last two rows of Table 10). These results provide strong evidence that users with low *ReadingP* are having difficulty with visualization processing. As far as we are aware, our results are the first to show a significant impact of reading proficiency on visualization processing.

As reported in Sect. 4, we found that users with low *VerbalWM* and low *ReadingP* spend significantly more time on task to achieve comparable comprehension accuracy as their counterparts. Our findings here shed light on exactly where these users are likely having difficulty, thus providing insights into how they could be helped in processing the MSNVs more efficiently. Our results show that, for users with low *VerbalWM*, this help should target the textual region of the MSNV, whereas users with low *ReadingP* would likely benefit from help in processing the visualization. Based on these findings, we choose to carry out an additional analysis of gaze metrics for users with low *ReadingP*, to ascertain whether we can identify specific aspects of the visualizations they need help with. Specifically, we will further examine the role that *ReadingP* has on visualization processing by defining a new set of *finer-grained* AOIs within the MSNV visualization only. We opt to focus only on *ReadingP* as a first step, because the primary goal of our current work is to target user-adaptive support on the visualization, and because previous research has indicated there are many candidate elements that could come into play during visualization processing (e.g., legend, labels, and bars relevant to the task).

5.4 Specifying finer-grained AOIs on the visualization

In the previous sub-section, we identified that users with low *ReadingP* were having difficulty processing the visualization part of the MSNVs, and these behaviors contribute to longer overall task completion time. In order to see whether we can identify *where* within the visualization these users are having difficulty, we specify a new set of *finer-grained* AOIs defined explicitly over key features of the visualizations. The new AOI definitions are as follows (an example is shown in Fig. 5):

- *Legend AOI*: Area surrounding the visualization legend.
- *Labels AOI*: Region along x-axis or y-axis (depending on orientation of the visualization) where textual bar labels are shown.

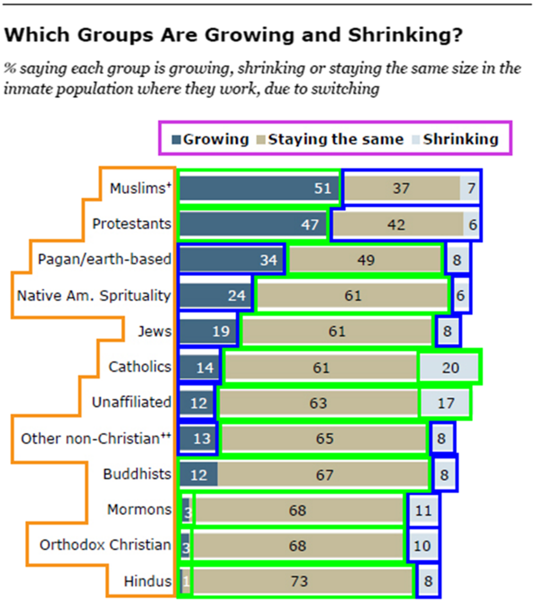


Fig. 5 A visualization in our MSNV dataset illustrating the four *finer-grained* AOIs we defined: *Legend AOI* (purple box), *Labels AOI* (orange region), *R-Bars AOI* (green boxes), and *NR-Bars AOI* (blue boxes). (Color figure online)

- *Referenced Bars (R-Bars) AOI*: Area covering the set of all bars mentioned by any reference (this AOI is identical to the *R-Bars AOI* in the previous section).
- *Non-Referenced Bars (NR-Bars) AOI*: Area covering all the other bars in the visualization, not mentioned by any reference.

We then re-compute the same collection of gaze metrics as before (see Table 7 in Sect. 5.1) using these four finer-grained AOIs, yielding a total of 68 new gaze metrics (i.e., 17 gaze metrics × 4 AOIs) to be used in the next analysis.

5.5 *Finer-Grained AOI* gaze metrics relevant to time on task

As done before (see Sect. 5.2), we first check for correlations among the 68 gaze metrics (i.e., 17 gaze metrics × 4 AOIs, c.f. Table 7) within each AOI. Shapiro–Wilk normality tests revealed that all of the gaze metrics were normally distributed ($p > .05$). Pearson correlations revealed very high correlations ($r > 0.9$) in all four finer-grained AOIs among: *sum_fixation_durations, number_of_fixations*, and *transitions_to_self;* and *longest_fixation* and *stddev_fixation_durations*. Based on these high correlations, we removed for each AOI: *number_of_fixations*, *transitions_to_self*, and *stddev_fixation_durations* from further analysis. Therefore, the total number of finer-grained AOI gaze metrics we retain for further investigation is 56: (14 metrics × 4 AOIs).

Next, to identify finer-grained AOI gaze metrics that have a significant relationship to *time on task*, we construct one Mixed Model for each of the 56 gaze metric as the independent measure, with *time on task* as the dependent measure, and *user_id* and *MSNV_id* as random effects. We apply the Benjamini–Hochberg procedure to our results, yielding a critical threshold of $q = 0.0286$, obtained at rank $r = 32$. Thus, our

Table 11 Results indicating which gaze metrics using *finer-grained* AOIs were found to be significant with time on task. Endash (–) cells indicate metrics that were excluded from the analysis due to high correlation, i.e., transitions to self (the same AOI). The normalized model coefficient b indicates the size and directionality of the relationship

Gaze Metric	Legend AOI	Labels AOI	R-Bars AOI	NR-Bars AOI
sum_fix_durations	$p < .001$ $b = 0.25$	$p < .001$ $b = 0.25$	$p < .001$ $b = 0.30$	$p < .001$ $b = 0.27$
longest_fixation	$p < .001$ $b = 0.16$	$p < .001$ $b = 0.13$	$p < .001$ $b = 0.16$	$p < .001$ $b = 0.15$
time_to_first_fix	$p < .001$ $b = 0.17$	$p < .001$ $b = 0.17$	$p < .001$ $b = 0.20$	$p < .001$ $b = 0.15$
time_to_last_fix	$p < .001$ $b = 0.38$	$p < .001$ $b = 0.63$	$p < .001$ $b = 0.58$	$p < .001$ $b = 0.62$
mean_fix_durations	$p = .002$ $b = 0.13$	*not sig.*	$p < .001$ $b = 0.18$	$p < .001$ $b = 0.11$
fixation_rate	*not sig.*	*not sig.*	*not sig.*	*not sig.*
prop_num_fixations	*not sig.*	*not sig.*	*not sig.*	$p = .008$ $b = 0.07$
transitions_to_Legend	–	$p < .001$ $b = 0.11$	$p < .001$ $b = 0.09$	$p < .001$ $b = 0.22$
transitions_to_Labels	$p = .005$ $b = 0.11$	–	$p < .001$ $b = 0.16$	$p < .001$ $b = 0.17$
transitions_to_R-Bars	$p < .001$ $b = 0.10$	$p < .001$ $b = 0.17$	–	$p < .001$ $b = 0.22$
transitions_to_NR-Bars	$p < .001$ $b = 0.18$	$p < .001$ $b = 0.15$	$p < .001$ $b = 0.25$	–

analysis revealed 32 finer-grained AOI gaze metrics that have a significant relationship with time on task, listed in Table 11, and in all cases (as indicated by b) have positive directionality with time on task (i.e., higher values of these gaze metrics indicate longer times on task).

5.6 Effects of *ReadingP* on finer-grained AOI gaze metrics

Here, our goal is to see where there is an effect of *ReadingP* on any of the finer-grained AOI gaze metrics identified in the previous sub-section. Using the same methodology as before, we construct one Mixed Model for each of the 32 relevant gaze metrics as the dependent measure, with *ReadingP* as a covariate, and user_id and MSNV_id as random effects. We apply the Benjamini–Hochberg procedure to our results, yielding a critical threshold of $q = 0.0297$, obtained at rank $r = 19$. Significant results from this analysis are reported in Table 12.

Starting with an examination of the first, second, and fourth rows in Table 12 (i.e., *sum_fix_durations*, *longest_fixation*, *and time_to_last_fix*), we can see that for all three of these gaze metrics, *ReadingP* appears with a negative directionality across all four of the AOI regions we examined (i.e., users with low *ReadingP* are generating

Table 12 Results showing significant effects of *ReadingP* were found on *finer-grained* AOI gaze metrics. The normalized model coefficient *b* indicates the size and directionality of the relationship. Endash (–) cells indicate gaze metrics that are non-relevant to time on task, and were not evaluated

Gaze Metric	Legend AOI	Labels AOI	R-Bars AOI	NR-Bars AOI
sum_fix_durations	***ReadingP*** $p = .007$ $b = -0.10$	***ReadingP*** $p = .003$ $b = -0.14$	***ReadingP*** $p = .003$ $b = -0.12$	***ReadingP*** $p = .002$ $b = -0.14$
longest_fixation	***ReadingP*** $p = .003$ $b = -0.12$	***ReadingP*** $p = .018$ $b = -0.08$	***ReadingP*** $p = .003$ $b = -0.15$	***ReadingP*** $p = .01$ $b = -0.12$
time_to_first_fix	*not sig.*	*not sig.*	*not sig.*	*not sig.*
time_to_last_fix	***ReadingP*** $p = .009$ $b = -0.14$	***ReadingP*** $p = .01$ $b = -0.14$	***ReadingP*** $p = .006$ $b = -0.16$	***ReadingP*** $p = .013$ $b = -0.13$
mean_fix_durations	***ReadingP*** $p = .025$ $b = -0.06$	–	*not sig.*	*not sig.*
fixation_rate	–	–	–	–
prop_num_fixations	–	–	–	*not sig.*
transitions_to_Legend	–	***ReadingP*** $p = .004$ $b = -0.07$	*not sig.*	*not sig.*
transitions_to_Labels	*not sig.*	–	***ReadingP*** $p = .019$ $b = -0.09$	***ReadingP*** $p = .007$ $b = -0.14$
transitions_to_R-Bars	*not sig.*	*not sig.*	–	***ReadingP*** $p = .002$ $b = -0.13$
transitions_to_NR-Bars	*not sig.*	***ReadingP*** $p = .001$ $b = -0.15$	***ReadingP*** $p = .004$ $b = -0.10$	–

higher values of these gaze metrics). As such, there is no clear indication on where to begin providing support to users with low *ReadingP*, since all of the visualization regions are possible candidates. However, for the gaze metric *mean_fix_durations* (fifth row in Table 12), *ReadingP* appears for the Legend AOI only with a negative directionality, indicating that users with low *ReadingP* were generating longer fixations on average while processing the visualization's Legend. Prior gaze research has shown that longer fixations are an indication that users are having difficulty extracting information, or at best are capturing some form of increased engagement (Just and Carpenter 1976). Therefore, our findings suggest that users with low *ReadingP* require extra time and effort to process the bar-group mappings elicited by the legend. We also found significant main effects of *ReadingP* on several transition-based gaze metrics (see the last four rows of Table 12).

First, users with low *ReadingP* transitioned more often between the R-Bars AOI (i.e., bars in the visualization mentioned by references) and NR-Bars AOI (i.e., bars

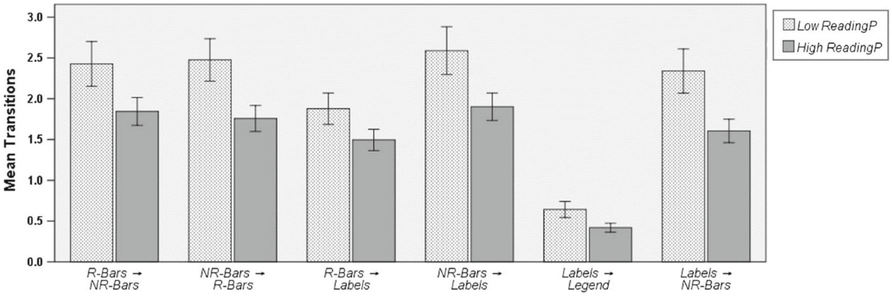


Fig. 6 Differences in transition behaviors within the visualization between users with low vs high *ReadingP* (median split) that result in slower performance. Bars are shown with 95% confidence intervals

in the visualization not referenced and not relevant for comprehension), indicating that they might have problems identifying the referenced bars. Figure 6 (left) illustrates the observed differences in these two transitions between users with low and high *ReadingP*, reported via a median split. As mentioned above, this extra transitioning might be a consequence of the fact that these users have difficulty establishing the mapping encoded by the legend compared to their high *ReadingP* counterparts. Thus, in order to alleviate the time that users with low *ReadingP* are wasting scanning for the referenced bars, highlighting could be provided in real time to guide their attention there, by using, for instance, the examples of effective bar chart highlighting techniques (e.g., bolding, de-emphasizing) for guiding users' attention presented in Carenini et al. (2014).

Second, always looking at the last four rows of Table 10 we identified increased transitioning relating to the Labels AOI. Users with low *ReadingP* transitioned more often from R-Bars and NR-Bars AOI to the Labels AOI, and transitioned more often from the Labels AOI to the Legend and NR-Bars AOI. Figure 6 (center and right) illustrates the observed differences in these four transitions between users with low and high *ReadingP*, reported via a median split. This extra processing implies these users are likely having difficulty with the mappings between the bars and their textual labels, and/or may be spending this extra time double-checking to ensure they are looking at the right bars once identified. As such, further guidance could be useful to help emphasize these mappings for these users, e.g., by providing highlighting on the labels, along with the highlighting of corresponding relevant bars as discussed above.

6 Conclusion and future work

In this paper, we conducted several analyses using Linear Mixed-Effects Models to uncover processing behaviors that are negatively impacting user experience (i.e., time on task, comprehension accuracy) with Magazine Style Narrative Visualizations (MSNVs) for users with low abilities in several user characteristics. First, we identified two groups of users for which low abilities in four user characteristics are at a disadvantage and thus could potentially benefit from adaptive support to aid them in processing MSNVs: low *Need for Cognition* or *Verbal IQ* achieved worse comprehension accuracy despite spending comparable time on task as their counterparts; and low *Verbal Working Memory* or *Reading Proficiency* require more time on task to achieve comparable accuracy as their counterparts. Next, we performed an analysis of gaze data aimed at identifying where significant differences in MSNV performance are occurring for these four user characteristics in terms of how the documents are visually processed. Our analysis did not uncover any significant relationships between MSNV processing and comprehension accuracy, and as a result, we were not able to provide any insights into why users low in *Need for Cognition* or *Verbal IQ* were less accurate. However, our analysis did reveal numerous MSNV processing behaviors that related to time on task, and as a result, we were able to identify main effects of *Verbal Working Memory* and *Reading Proficiency* on several of them. First, we found that users with low *Verbal Working Memory* spent more time processing the main body of text contained in the MSNVs, and took longer to locate for the first time the textual references that discuss specific bars/datapoints within the visualization. Second, we found that processing behaviors, indicative of where users with low *Reading Proficiency* were struggling, were all exclusively related to the visualizations contained in each of the MSNV documents, which included difficulty processing relevant bars elicited by the references. Therefore, as a first step toward better understanding how meaningful adaptive support *within* the visualization could be devised for users with low *Reading Proficiency*, we conducted a follow-up analysis of *finer-grained* gaze processing behaviors within the visualization only. This follow-up analysis revealed several MSNV processing behaviors that capture where users with low *Reading Proficiency* were struggling. Here, we summarize our findings and include preliminary suggestions on ways to provide help during MSNV processing:

i. Users with low *Reading Proficiency* transition back and forth significantly more often between the relevant bars and the non-relevant bars in the visualization. This extra transitioning is likely a consequence of difficulty establishing the mappings encoded by the Legend, given we also found that users with low *Reading Proficiency* spend significantly more time looking at the legend with longer average fixation durations. To alleviate the time, these users are wasting scanning for the referenced bars, highlighting could be provided in real time to guide their attention there (e.g., bolding, de-emphasizing) as was effectively demonstrated on bar chat visualizations in Carenini et al. (2014).
ii. For several gaze transitions relating to the labels, users with low *Reading Proficiency* had significantly more transitions. Specifically, users with low *Reading Proficiency* transitioned more often from the relevant bars and non-relevant bars to the labels, and they transitioned more often from the labels to the legend and non-relevant bars. These users are likely having difficulty processing the mappings between the bars and their textual labels, and as such further highlighting on the labels, along with the highlighting of corresponding relevant bars as discussed in the previous bullet, could be provided concurrently to help reinforce these mappings.

The research presented in this paper is extending previous work on user-adaptive information visualizations to MSNVs, which are arguably more complex and challenging for the reader. Because of this additional complexity and being the first attempt, both the user study and the data analysis were limited along several dimensions. In terms of visualizations embedded in the MSNVs, we have only considered bar charts.

In future work, we will investigate whether our findings generalize to MSNVs containing other possibly more complex visualizations. Presumably, users will experience even more serious difficulties with such MSNVs, as references in the accompanying text will likely be longer and more complicated; and the same will be the case for visualization's legends and labels. Still, considering the visualizations embedded within the MSNVs, the bar charts we examined were not all the same. Our MSNVs are contained in equal proportion simple, stacked, and grouped bar charts. So an interesting question is whether bar chart styles influence user gaze behavior and ultimately user experience. Since the focus of this paper is exclusively on the impact of user characteristics, answering this question is also left as future work. With respect to user traits, to keep the study manageable, we had to exclude some promising candidates, like *locus of control* and *domain expertise.* Further studies could explore the impact of these and other traits. Admittedly, the number of subjects in our user study was rather small, and several insignificant or marginally significant findings could turn into strong statistically significant results just by collecting more data. Based on this observation, future user studies should involve many more participants. However, because such studies are time-consuming and resource intensive they would be extremely challenging for a single research group. In principle, one way forward could be to leverage resources in multiple institutions. Notice that more data could also support exploring more specific research questions. For instance, as noted above, since our MSNVs are contained in equal proportion simple, stacked, and grouped bar charts, it would be quite interesting to verify whether the influence of user characteristics on user experience is mediated by the type of bar chart. However, this would require at least three times the amount of data we have collected so far.

The ultimate objective of our work is to provide real-time, user-adaptive support to MSNV processing. In this direction, we are currently designing and implementing two important functionalities: first, capturing users' fixations in real time by interfacing with the eye tracking hardware, so that that adaptations can be triggered for users with low characteristics based on when and where they are looking (e.g., triggering an intervention when a user looks at the visualization); second, implementing the functionality to dynamically highlight specified regions of the visualization (e.g., highlighting the labels or legend), including the ability to control various properties of the highlighting (e.g., duration, fade-in time, color, shape, etc.). Once implemented, we are planning to conduct a follow-up user study to design and test the effectiveness of the adaptive strategies identified above for users with low *Reading Proficiency*. An interesting question to explore in such study is whether users with high *Reading Proficiency* would also benefit from these strategies and to what extent. Lastly, we are also planning to implement and evaluate detecting relevant user characteristics (e.g., *Reading Proficiency*) without the need to administer tests prior to the study. We plan to do this noninvasively and in real time, while users process the MSNVs, by feeding their gaze data into machine learning models to generate predictions of their desired user characteristics. The feasibility of this approach has been previously demonstrated with several information visualizations and tasks (Steichen et al. 2014; Gingerich and Conati 2015; Lallé et al. 2015, 2016; Toker et al. 2017; Conati et al. 2017) via logistic regression and random forests, and we plan to leverage similar techniques and extend them to MSNVs.

References

Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)

Allen, B.: Individual differences and the conundrums of user-centered design: two experiments. J. Am. Soc. Inf. Sci. **51**(6), 508–520 (2000)

Baddeley, A.: Oxford Psychology Series, No. 11. Working Memory. Clarendon Press/Oxford University Press, New York (1986)

Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate—a practical and powerful approach to multiple testing. J. R. Stat. Soc. **57**(1), 289–300 (1995)

Blair, J.R., Spreen, O.: Predicting premorbid IQ: a revision of the national adult reading test. Clin. Neuropsychol. **3**(2), 129–136 (1989)

Boy, J., Rensink, R.A., Bertini, E., et al.: A principled way of assessing visualization literacy. IEEE Trans. Vis. Comput. Graph. **20**(12), 1963–1972 (2014)

Cacioppo, J.T., Petty, R.E., Kao, C.F.: The efficient assessment of need for cognition. J. Pers. Assess. **48**(3), 306–307 (1984)

Carenini, G., Conati, C., Hoque, E., et al.: User task adaptation in multimedia presentations. In: Proceedings of the 1st International Workshop on User-Adaptive Information Visualization (WUAV 2013), in Conjunction with the 21st Conference on User Modeling, Adaptation and Personalization (UMAP 2013) (2013)

Carenini, G., Conati, C., Hoque, E., et al.: Highlighting interventions and user differences: informing adaptive information visualization support. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 1835–1844 (2014)

Çöltekin, A., Fabrikant, S.I., Lacayo, M.: Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. Int. J. Geogr. Inf. Sci. **24**(10), 1559–1575 (2010)

Conati, C., Maclaren, H.: Exploring the role of individual differences in information visualization. In: Proceedings of the Working Conference on Advanced visual Interfaces, New York, NY, USA. ACM, pp. 199–206 (2008)

Conati, C., Carenini, G., Hoque, E., et al.: Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. In: Computer Graphics Forum, Wiley Online Library, pp. 371–380 (2014)

Conati, C., Lallé, S., Rahman, M.A., et al.: Further results on predicting cognitive abilities for adaptive visualizations. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia. AAAI Press (2017)

D'Mello, S., Olney, A., Williams, C., et al.: Gaze tutor: a gaze-reactive intelligent tutoring system. Int. J. Hum. Comput. Stud. **70**(5), 377–398 (2012)

D'Mello, S., Mills, C., Bixler, R., et al.: Zone out no more: mitigating mind wandering during computerized reading. In: Proceedings of the 10th International Conference on Educational Data Mining, Wuhan, China, pp. 8–15 (2017)

Dyson, M.C., Haselgrove, M.: The influence of reading speed and line length on the effectiveness of reading from screen. Int. J. Hum. Comput. Stud. **54**(4), 585–612 (2001)

Ekstrom, R.B., French, J.W., Harman, H.H., et al.: Manual for Kit of Factor Referenced Cognitive Tests. Educational Testing Service, Princeton (1976)

Field, A.P.: How to Design and Report Experiments. Sage, London (2003)

Folker, S., Ritter, H., Sichelschmidt, L.: Processing and integrating multimodal material—the influence of color-coding. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 27(27), pp. 690–695 (2005)

Gingerich, M.J., Conati, C.: Constructing models of user and task characteristics from eye gaze data for user-adaptive information highlighting. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)

Gotz, D., Wen, Z.: Behavior-driven visualization recommendation. In: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI'09, New York, NY, USA. ACM, pp. 315–324 (2009)

Grabe, W., Jiang, X.: Assessing reading. In: Kunnan, A.J. (ed.) The Companion to Language Assessment, pp. 185–200. Wiley, Hoboken (2013)

Grawemeyer, B.: Evaluation of ERST: an external representation selection tutor. In: Proceedings of the 4th International Conference on Diagrammatic Representation and Inference, Diagrams'06. Springer, Berlin, Heidelberg, pp. 154–167 (2006)

Green, T.M., Fisher, B.: Towards the personal equation of interaction: the impact of personality factors on visual analytics interface interaction. In: 2010 IEEE Symposium on Visual Analytics Science and Technology (VAST), pp. 203–210 (2010)

Green, N.L., Carenini, G., Kerpedjiev, S., et al.: AutoBrief: an experimental system for the automatic generation of briefings in integrated text and information graphics. Int. J. Hum. Comput. Stud. **61**(1), 32–70 (2004)

Hegarty, M., Just, M.A.: Constructing mental models of machines from text and diagrams. J. Mem. Lang. **32**(6), 717–742 (1993)

Huang, D., Tory, M., Aseniero, B.A., et al.: Personal visualization and personal visual analytics. IEEE Trans. Vis. Comput. Graph. **21**(3), 420–433 (2015)

Human Factors (HF): Guidelines on the Multimodality of Icons, Symbols and Pictograms. European Telecommunications Standards Institute, Sophia Antipolis (2008)

Just, M.A., Carpenter, P.A.: Eye fixations and cognitive processes. Cogn. Psychol. **8**(4), 441–480 (1976)

Kalyuga, S.: Managing Cognitive Load in Adaptive Multimedia Learning. Information Science Reference, Hershey (2009)

Kalyuga, S., Chandler, P., Sweller, J.: Levels of expertise and instructional design. Hum. Factors **40**(1), 1–17 (1998)

Kalyuga, S., Law, Y.K., Lee, C.H.: Expertise reversal effect in reading Chinese texts with added causal words. Instr. Sci. **41**(3), 481–497 (2013)

Kong, N., Hearst, M.A., Agrawala, M.: Extracting references between text and charts via crowdsourcing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, Canada. ACM, pp. 31–40 (2014)

Kules, B., Capra, R.: Influence of training and stage of search on gaze behavior in a library catalog faceted search interface. J. Am. Soc. Inform. Sci. Technol. **63**(1), 114–138 (2012)

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B.: lmerTest package: tests in linear mixed effects models. J. Stat. Softw. **82**(13) (2017). http://www.jstatsoft.org/v82/i13/. Accessed 12 October 2018

Lallé, S., Toker, D., Conati, C., et al.: Prediction of users' learning curves for adaptation while using an information visualization. In: Proceedings of the 20th International Conference on Intelligent User Interfaces. ACM, pp. 357–368 (2015)

Lallé, S., Conati, C., Carenini, G.: Predicting confusion in information visualization from eye tracking and interaction data. In: Proceedings on the 25th International Joint Conference on Artificial Intelligence, pp. 2529–2535 (2016)

Lallé, S., Conati, C., Carenini, G.: Impact of individual differences on user experience with a visualization interface for public engagement. In: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP'17, New York, NY, USA. ACM, pp. 247–252 (2017)

Lankow, J., Ritchie, J., Crooks, R.: Infographics: The Power of Visual Storytelling. Wiley, Hoboken (2012)

Loboda, T.D., Brusilovsky, P., Brunstein, J.: Inferring word relevance from eye-movements of readers. In: ACM Press, pp. 175 (2011)

Logie, R.H.: Visuo-spatial Working Memory. Nachdr. Psychology Press, Hove (2009)

Martínez-Gómez, P., Aizawa, A.: Recognition of understanding level and language skill using measurements of reading behavior. In: ACM Press, pp. 95–104 (2014)

Mayer, R.E.: Multimedia Learning, 2nd edn. Cambridge University Press, Cambridge (2009)

Meara, P.: EFL Vocabulary Tests, 2nd edn. Lognostics, Swansea (2010)

Meara, P., Jones, G.: Eurocentres Vocabulary Size Test 10KA. Eurocentres Learning Service, Zurich (1990)

Metoyer, R., Zhi, Q., Janczuk, B., et al.: Coupling story to visualization: using textual analysis as a bridge between data and interpretation. In: ACM Press, pp. 503–507 (2018)

Munzner, T.: Visualization Analysis and Design. CRC Press, Taylor & Francis Group (2014)

Nazemi, K., Retz, R., Bernard, J., et al.: Adaptive semantic visualization for bibliographic entries. In: Advances in Visual Computing, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 13–24 (2013)

Olsen, A.: The tobii i-vt fixation filter. Tobii Technology (2012). http://www.tobii.com/global/analysis/training/whitepapers/tobii_whitepaper_tobiiivtfixationfilter.pdf. Accessed 13 September 2015

Ooms, K., De Maeyer, P., Fack, V., et al.: Interpreting maps through the eyes of expert and novice users. Int. J. Geogr. Inf. Sci. **26**(10), 1773–1788 (2012)

Ooms, K., De Maeyer, P., Fack, V.: Study of the attentive behavior of novice and expert map users using eye tracking. Cartogr. Geogr. Inf. Sci. **41**(1), 37–54 (2014)

Ottley, A., Yang, H., Chang, R.: 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, Seoul, Korea. ACM, pp. 3251–3254 (2015)

Ozcelik, E., Arslan-Ari, I., Cagiltay, K.: Why does signaling enhance multimedia learning? Evidence from eye movements. Comput. Hum. Behav. **26**(1), 110–117 (2010)

Scheiter, K., Wiebe, E., Holsanova, J.: Theoretical and Instructional Aspects of Learning with Visualizations. Instructional Design: Concepts, Methodologies, Tools and Applications. IGI Global, Hershey (2011)

Segel, E., Heer, J.: Narrative visualization: telling stories with data. IEEE Trans. Vis. Comput. Graph. **16**(6), 1139–1148 (2010)

Steichen, B., Conati, C., Carenini, G.: Inferring visualization task properties, user performance, and user cognitive abilities from eye gaze data. ACM Trans. Interact. Intell. Syst. **4**, 11 (2014)

Strauss, E., Sherman, E.M.S., Spreen, O., et al.: A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary, 3rd edn. Oxford University Press, Oxford (2006)

Tai, R.H., Loehr, J.F., Brigham, F.J.: An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. Int. J. Res. Method Educ. **29**(2), 185–208 (2006)

Tang, H., Topczewski, J.J., Topczewski, A.M., et al.: Permutation test for groups of scanpaths using normalized Levenshtein distances and application in NMR questions. In: ACM Press, pp. 169 (2012)

Toker, D., Conati, C.: 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. Springer, Aalborg, Denmark (2014)

Toker, D., Conati, C., Carenini, G., et al.: 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. Springer, Berlin, Heidelberg, pp. 274–285 (2012)

Toker, D., Conati, C., Steichen, B., et al.: Individual user characteristics and information visualization: connecting the dots through eye tracking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'13. ACM, New York, NY, USA, pp. 295–304 (2013)

Toker, D., Lallé, S., Conati, C.: Pupillometry and head distance to the screen to predict skill acquisition during information visualization tasks. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. ACM Press, pp. 221–231 (2017)

Toker, D., Conati, C., Carenini, G.: User-adaptive support for processing magazine style narrative visualizations: identifying user characteristics that matter. In: ACM Press, pp. 199–204 (2018)

Tufte, E.R.: Visual Explanations: Images and Quantities, Evidence and Narrative. Graphics Press, Cheshire (1997)

Turner, M.L., Engle, R.W.: Is working memory capacity task dependent? J. Mem. Lang. **28**(2), 127–154 (1989)

van Gog, T.: The signaling (or cueing) principle in multimedia learning. In: Mayer, Richard (ed.) The Cambridge Handbook of Multimedia Learning, pp. 263–278. Cambridge University Press, Cambridge (2014)

Velez, M.C., Silver, D., Tremaine, M.: Understanding visualization through spatial ability differences. In: Proceedings of the IEEE Conference on Visualization, Minneapolis, MN, USA. IEEE, pp. 511–518 (2005)

Vogel, E.K., Woodman, G.F., Luck, S.J.: Storage of features, conjunctions, and objects in visual working memory. J. Exp. Psychol. Hum. Percept. Perform. **27**(1), 92–114 (2001)

Waddell, T.F., Auriemma, J.R., Sundar, S.S.: Make it simple, or force users to read? Paraphrased design improves comprehension of end user license agreements. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'16. ACM Press, pp. 5252–5256 (2016)

Walker, D.A.: Converting Kendall's Tau for correlational or meta-analytic analyses. J. Mod. Appl. Stat. Methods **2**(2), 525–530 (2003)

Wiley, J., Sanchez, C.A., Jaeger, A.J.: The individual differences in working memory capacity principle in multimedia learning. The Cambridge Handbook of Multimedia Learning, pp. 598–620. Cambridge University Press, Cambridge (2014)

Wobbrock, J.O., Findlater, L., Gergle, D., et al.: The aligned rank transform for nonparametric factorial analyses using only anova procedures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 143–146 (2011)

Ziemkiewicz, C., Crouser, R.J., Yauilla, A.R., et al.: How locus of control influences compatibility with visualization style. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology, Providence, RI, USA. IEEE, pp. 81–90 (2011)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dereck Toker is a Ph.D. candidate at the University of British Columbia (Canada), and received his M.Sc. degree there in 2013. His areas of interest include User Modeling, Information Visualization, Eye-Tracking, and Machine Learning. His peer-reviewed publications in these fields have won the best paper award at UMAP'14 (User Modeling, Adaptation, and Personalization) and include a best paper award nomination at IUI'14. He is currently completing his Ph.D. and plans to enter industry as a statistics consultant and machine learning engineer.

Cristina Conati, Ph.D is a Professor of Computer Science at the University of British Columbia (Canada). Her research goal is to integrate research in Artificial Intelligence (AI), Cognitive Science, and Human Computer Interaction (HCI) with the goal to create intelligent interactive systems that can capture relevant user's properties (states, skills, needs) and personalize the interaction accordingly. Her areas of interest include User-Adaptive Interaction, User Modeling, Intelligent Tutoring Systems, and Affective Computing. She has over 100 peer-reviewed publications in these fields, and her research has received awards from a variety of venues, including UMUAI (2002), IUI (2007), UMAP (2013, 2014), TiiS (2014), and IVA 2016. Dr. Conati is an associate editor for UMUAI, ACM TiiS, Iand IJAIED. She has served as Program or Conference Chair for several international conferences including UMAP, IUI, and AI in Education. She is a Senior Member of AAAI, and serves on the Executive Committee of the Association.

Giuseppe Carenini, Ph.D is a Professor in Computer Science at UBC (Vancouver, Canada). Giuseppe has broad interdisciplinary interests. His work on natural language processing and information visualization to support decision making has been published in over 100 peer-reviewed papers (including best paper at UMAP-14 and ACM-TiiS-14). Dr. Carenini was the area chair for "Sentiment Analysis, and Text Classification" of ACL 2009, for "Summarization and Generation" of NAACL 2012, and or "Discourse Analysis" of ACL 2019. He was the Program Co-Chair for IUI 2015, and the Program Co-Chair for SigDial 2016. He has also co-edited in 2012 an ACM-TIST Special Issue on "Intelligent Visual Interfaces for Text Analysis." In 2011, he published a co-authored book on "Methods for Mining and Summarizing Text Conversations." In his work, Dr. Carenini has also extensively collaborated with industrial partners, including Microsoft and IBM. Giuseppe was awarded a Google Research Award, an IBM CASCON Best Exhibit Award, and a Yahoo Faculty Research Award in 2007, 2010, and 2016, respectively.