Exploring students' approach to factual texts in different presentation media

Jens Nirme* [1], Olga Fredriksson [2], Magnus Haake [1] and Agneta Gulz [1]

[1] Lund University Cognitive Science
[2] Special Education Department, McLean Lower Secondary School, Skurup, Sweden

*jens.nirme@lucs.lu.se

Alternative ways to present factual texts are becoming increasingly common in primary education. Previous research has shown possible benefits to learners with poor reading skills of simultaneous presentation of synthesized speech and visual text. Others findings have suggested increased engagement with material by visual presentation of an animated speaker. However, it is often difficult to compare results across different media and different studies, since presentation interfaces often vary in information transience and navigation possibilities. We explored how different media affect comprehension and metacognitive strategies when differences in these factors are minimized, in a study with 119 thirteen- to fourteen- year old participants. We replicated some previous findings: that reading promotes better comprehension compared to only listening, and found some support that reading-while-listening decreases the demands on reading skill for some students. We also found that visual presentation of an animated speaker improved comprehension compared to synthetic speech alone. Overall, we found little repetition or non-sequential navigation even though the experimental navigation interface and time limits allowed for this. We discuss possible reasons for this, result.

Keywords – learning; comprehension; reading; listening; multimodality; meta-cognition

Introduction

In typical educational settings, students frequently acquire knowledge through teachers' oral presentations, or by reading textbooks. Teachers generally try to make sure students attend and understand the material they present orally, whereas, when it comes to reading, students have to manage their comprehension and attention for themselves. Students vary with respect to how they access and process learning materials. Some prefer and/or learn best by reading on their own; others by listening to a teacher or a peer. However, learning to independently acquire knowledge from factual texts is an important skill to learn and it is worthwhile to meet the individual challenges students might face. The integration of digital devices in primary education offers various ways to address this challenge.

Digital media can, compared to printed media, be more readily adapted both in content and presentation. In the current paper we will focus on the latter, describing a study where we explored how different presentation modalities affected comprehension and metacognitive approaches to factual texts, when minimizing de-facto differences in constraints such as time limits, information transience and navigation possibilities. We were interested in the replicability of some previous findings – namely that comprehension improves when reading compared to hearing texts read by a synthetic voice, and that reading-while-listening can mitigate poor baseline reading comprehension – when controlling for these factors. Moreover, we wanted to compare the mentioned media (visual- and/or audio- presentation of text) to an alternative way of presenting material in digital media: by a digitally animated character with synthesized speech and movement.

Reading while listening

The simultaneous presentation of text and speech, either prerecorded or generated in real-time as synthesized speech, is today commonly used to support students who for different reasons have problems with reading comprehension (Drager, Reichle, & Pinkoski, 2010). While there is no documented origin of teachers' (or peers') reading aloud in a classroom while students follow along with the text, Schneeberg (1977) presented the first systematic study of a reading-while-listening

paradigm that we are aware of. The study included both teachers reading aloud and audiotapes, and results indicated that reading-while-listening could have long term benefits on students' reading ability. The advantages of presenting both text and speech in a digital format, whether synthesized- or digitized natural speech, is the potential to individualize support by adapting variables such as speech rate or voice gender to abilities or preferences of different students. It also allows students to study material individually where and when they want to, and serves to eliminate distractions associated with reading in a group. In addition, digital presentation allows for automatic highlighting of text as it is voiced (Montali & Lewandowski, 1996) and is a potential time saver for educators or producers of educational material.

Some studies suggest that reading while listening to synthesized speech can be beneficial for students with attention disorders (Hecker, Burns, Katz, Elkind, & Elkind, 2002) and dyslexia (Elkind, Cohen, & Murray, 1993) as well as for foreign language learners (Handley, 2009). Listening to synthesized or digitized speech along with text is recommended by the Swedish Agency for Accessible Media (https://www.mtm.se) as a method for reinforcing memory and comprehension.

A possible explanation for improved memory by reading while listening and comprehension can be found in the 'redundant coactiviation effect' demonstrated by (Miller, 1982): that congruent (nonverbal) audiovisual stimuli improved reaction times in simple decision tasks beyond what could be explained by processing of the two channels separately. Audiovisual presentation has also been shown to enhance recall of single words compared to presentation in only one channel (Penney, 1989), attributable to dual coding in (verbal and visual) working memory (Baddeley, 1992). However, it is worth pointing out that reading in silence already involves activation of phonological word representations (Van Orden, Johnston, & Hale, 1988), although the activation might be perhaps less efficient for poor readers (Unsworth & Pexman, 2003).

Comprehension is a complex task, involving more than simply recognizing and encoding words; the words' meaning have to be semantically interpreted and integrated with previous knowledge. According to cognitive load theory (Sweller, 1988) there are three types of load roughly corresponding to processes involved in comprehension: extracting information from presentation media ('extraneous load'), understanding it ('intrinsic load') and constructing knowledge schemas ('germane load'). All three processes share a limited cognitive resource, and the combined load of the three determines the load on working memory and the mental effort required by a task. Parallel or intrinsic processing of visual information may interfere with dual channel processing of text auditory information, which may result in readers inhibiting the auditory channel (further increasing load) and counteracting any possible redundant co-activation effect (Hilbert, Nakagawa, Puci, Zech, & Bühner, 2015; Kalyuga, Chandler, & Sweller, 1999; Moreno & Mayer, 2002; Moussa-Inaty, Ayres, & Sweller, 2012).

Given the complexity of comprehension, it is hardly surprising that studies of how text comprehension is affected by 'reading-while-listening' have had heterogeneous results. Some have found improved comprehension compared to only listening (Dowell & Shmueli, 2008; Taake, 2009), but no difference compared to only reading.

It is however worth pointing out that the participants in these studies were university students and presumably quite skilled readers. Higgins & Zvi (1995) tested adult students with learning disabilities of different kinds, and found that those who had a poor baseline reading ability particularly benefited from hearing a synthesized voice along with their reading compared to reading without the voice. Montali and Lewandowski (1996) tested 13-15 year old students and found that voice recordings played along with text improved comprehension in general, and that the effect was more pronounced for students with weak baseline reading comprehension.

Other researchers have specifically investigated differences in effects between natural and synthesized voices, generally finding the latter to be somewhat more difficult to comprehend. Moreover, Winters and Pisoni (2006) also found that previous exposure and phonetic variation in the speech signal were mitigating factors to the detrimental effect of synthesized voices. On the other hand, Taake (2009) found no difference in college students' comprehension for natural and synthetic voices, neither when listening nor reading-while-listening. It is worth pointing out that in recent years, speech synthesizers have improved considerably with regards to the latter aspect and sound less flat and 'robotic' (while still in most cases being distinguishable from human voices). Drager et al. (2010) reviewed ten studies that tested intelligibility or comprehension (response latencies) of synthesized speech of children, concluding that children perform similarly to adults with synthesized speech, but sometimes at lower levels.

As for long-term learning effects, results are even less clear. Gisterå (1995) found that voice recordings of factual texts could not by themselves meet the demands of dyslexic students. Reed, Swanson, Petscher and Vaughn (2013) found no improved learning or retention of social studies material in high school seniors' (17-18 years of age) from having teachers read texts aloud while students also had access to the texts, compared to individual reading only.

Video lectures and digitally animated speakers

For students who prefer listening over reading and/or perform better after listening compared to after reading, video recorded lectures or instructional

videos offer an alternative way to approach a learning material independently. A survey by Allison (2015) found that 85% of teachers regularly used videos as part of primary and secondary education in the US. Also, videos are often an important component of online learning platforms, which are increasingly used in university level education (Allen & Seaman, 2009). Scagnoli, Choo, & Tian (2019) surveyed university students taking online courses and found that 63% self-reported a benefit of video lectures, with arguments that they increased independence and control. It is however unclear if video material has any measurable effect on learning outcome (Vagula & Liu, 2016).

The added sensory input of a visually presented speaker has in itself demonstrated effects on speech recognition (Sumby & Pollack, 1954) as well as comprehension, particularly in noisy listening environments (Nirme, Haake, Lyberg Åhlander, Brännström & Sahlén, 2018). The phenomenon that the visual presentation of incongruent lip movements modulates auditory perception of articulated syllables is called the McGurk effect (McGurk & MacDonald, 1976). The fact that this effect occurs despite awareness that there is incongruity of the speech and lip movements shows that integration of visual speech happens at an early processing state, however not necessarily without increased (extraneous) load (Jansen, Chaparro, Downs, Palmer & Keebler, 2013; Mishra, Lunner, Stenfelt, Rönnberg & Rudner, 2013).

An alternative way of presenting a learning material is to combine speech – synthesized or digitized – with a digital animated representation of the speaker (Nirme et al., 2019). Different types of animated speakers have (depending on application domain) been called Virtual Humans (Garau, Slater, Pertaub, & Razzaque, 2005), Animated Pedagogical Agents (Clark & Choi, 2005) or Embodied Conversational Agents (Cassell, Sullivan, Churchill, & Prevost, 2000). Implementations may vary in aspects: the degree of scripted versus autonomous behavior; the social (or pedagogical) role of the animated speaker; the naturalism and expressiveness (some communicate strictly by text; Gulz & Haake, 2006a). Mattheyses and Verhelst (2015) give an overview of techniques for generating audiovisual speech. It has been shown that speech recognition can be facilitated by seeing digitally animated faces with procedurally generated lip movements that match the phonemes (Cohen, Walker, & Massaro, 1995) but usually to a lesser degree than seeing a real speaker (Grant & Seitz, 2000; Ross, Saint-Amour, Leavitt, Javitt & Foxe, 2007).

Animated speakers add further possibilities for customization as visual appearance can be modified independently from speech and other behaviors (Gulz & Haake, 2006b). Lester et al. (1997) found that animated pedagogical agents promoted positive experiences of learning activities; a phenomenon that they called 'the persona effect'. Dunsworth and Atkinson (2007) found that a visually presented agent narrating slides explaining the human circulatory system improved retention compared to audio-only as well as text-only narration. Moreno, Mayer, Spires and Lester (2001) proposed that pedagogical agents that exhibit 'social agency' promote motivation and engagement with educational material by making students (on some level) relate in social terms to the agent. However, some studies contradict the general validity of such claims, finding great variation in school children's responses to animated agents in pedagogical roles (Gulz, 2005). Also, both students' perception of the agent and its effect on learning may depend on to what extent it exhibits realistic and appealing speech and movement (Domagk, 2010; Mayer & DaPra, 2012). Craig, Gholson, & Driscoll (2002) found no effect on retention from a visually presented agent combined with audio narration, regardless of whether the agent made animated gestures or not. Nirme et al. (2019) found no effect on comprehension from the presence of a realistically animated speaker, in the absence of background babble noise – but there was a marginal effect with background babble. Clark & Choi (2005) suggested that animated speakers increase extraneous load since they present nonessential or distracting information such as facial expressions or cosmetic visual features.

Information transience

In a preliminary study we investigated how presentation media affected 14-15 year-olds (N = 76) comprehension of four factual texts (Fredriksson, 2015). More specifically, we compared their scores on multiple choice questions in four different conditions: *r*, reading; *rl*, reading while listening to synthesized speech; *l*, listening to synthesized speech (without text visually present) and *v*, listening and watching a video of a digitally 3D-animated character with synthesized speech and movements. The order and pairing of texts and conditions were counterbalanced. The main result revealed significantly stronger comprehension in *r* and *rl* (the conditions where text was visually available), compared to conditions *l* and *v*. Poor readers, however, seemed to benefit from reading while listening (*rl*) compared to reading text alone. We found no main effect of seeing the animated speaker, in contrast to some previous studies (Dunsworth & Atkinson, 2007).

One explanation for the main result is that visual text is permanent whereas audio and animation (in the way it was presented in the study) is transient. Texts were printed on paper in the *r* and *rl* conditions and although there was a time limit set for the participants, it was possible for fast readers to read through the text, or parts of it, more than once, whereas in *l* and *v* the time was determined by the rate of the synthesized speech which could only be listened to one time. Students thus had no possibility to control their own pace or re-read things that they failed to understand or needed to remind themselves of. That information transience can be detrimental to comprehension is in line with previous findings. Singh, Marcus, & Ayres (2012) found that students' information uptake from the same material (about passing a bill in the US parliamentary system) was better when the material was presented as text compared to as recorded speech. Learning from speech, however, improved when the speech was divided into smaller segments separated by 5 s pauses and participants were instructed to use the pauses to think about what they just heard. This instruction likely helped to decrease students' working memory load while they listened. There was no clear indication that segmentation improved results (uptake of information) when material was presented as text. Also, the difference in comprehension between listening and reading or reading-while-listening reported by Dowell and Shmueli (2008) was not observed for short sentences.

Wong, Leahy, Marcus and Sweller (2012) similarly found that segmenting long and complex animations into shorter segments led to improved learning. They also investigated the well documented 'modality effect' (Ginns, 2005), stating that graphical information is better understood when accompanied by speech than by text, explained by reduced load when information is distributed over parallel working memory systems corresponding to the two modalities (Baddeley, 1992). Wong et al. (2012) found that the effect was reversed for longer (non-segmented) material, indicating that it was counteracted by an added load associated with maintaining the transient speech information in working memory. Another study also found modality effects only with long sections of spoken text (Leahy & Sweller, 2011).

Self-regulation and meta-cognition

Speech is by its nature transient. A possible explanation of the superiority of text over speech, measured as uptake and comprehension (Dowell & Shmueli, 2008; Fredriksson, 2015; Taake, 2009), is that a text – as long as it is permanently visually present and accessible to students, makes it possible for students to take on new material at their own pace and to repeat material. Self-pacing and repetition are components of self-regulated learning, which is held to be central in order to acquire knowledge on one's own (not being supervised or instructed step-by-step), perhaps particularly so in today's prevalence of nonlinear 'hypermedia' as information sources (Azevedo, 2005).

Self-regulation requires both strong meta-cognition and executive control, where the latter includes the ability to sustain attention to a given material. 'Mind wandering' is a common and well-documented phenomenon that interferes with reading comprehension, particularly when the text is challenging (Feng, D'Mello, & Graesser, 2013). A social response to an animated speaker, could potentially decrease the occurrence of 'mind wandering' by generating stronger engagement with the material and/or by implicitly triggering a social convention not to ignore someone speaking (Moreno et al., 2001). However, a study by Risko, Anderson, Sarwal, Engelhardt and Kingstone (2012) revealed that 'mind wandering' was common both during lectures given in a classroom setting and lectures presented as video recordings, and increasingly so during the second half of the lectures.

Self-regulated learning also depends on strong meta-cognitive ability (Zimmerman & Moylan, 2009). Nelson (1990) proposed a theory of 'metamemory' that emphasizes the role of monitoring and self-assessing one's own knowledge both for acquisition and for retrieval of knowledge. 'Calibration' is a concept from psychology that refers to consistency between self-assessed and actual performance, either predicted or post-hoc (Bol & Hacker, 2012). It is often proposed that the more accurate calibration is, the greater is the potential for self-regulation (Alexander, 2013). Assessments of one's knowledge seem causally linked to study behavior, such as choosing whether or not to repeat items to be recalled later (Metcalfe, 2009). However, both children and adults are often weak when it comes to assessing their own understanding. This can be explained by learners lacking the particular knowledge necessary to become aware of their own limitations (Kruger & Dunning, 1999), or by them having an incorrect mental model of their own learning and memory processes (Bjork, Dunlosky & Kornell, 2013). Salomon (1984) found that most school children estimated that it was easier to learn from a TV-program than from a printed text. As a consequence they chose the TV-program before the printed text, which, according to Salomon, led them to spend less effort which, in turn, resulted in weaker learning outcomes.

More accurate self-assessments can be promoted through instructions telling students to focus on predetermined or self-chosen keywords while reading (Gillström & Rönnberg, 1995) or by providing intermittent quizzes on the material presented in video recorded lecture segments (Schacter & Szpunar, 2015). Such improvements

cannot be explained by greater exposure to the learning material – instead they are likely due to the students' *approach* to the material. Schacter & Szpunar (2015) found no increased accuracy for a control group that were presented the questions and answers included in the intermittent quizzes without the requirement to answer them.

In most of the mentioned studies, including our own preliminary study (Fredriksson, 2015), that investigate learning from different media, there were also what could be called implicit instructions at work. For material not presented at text, but orally, possibilities to navigate in the material, to decide on the pace of presentation and/or to repeat material are limited or absent. Some researchers have specifically studied differences between 'user-paced' or 'system-paced' learning, where the former means that the learner controls the rate by which information is new presented and the latter means that the rate is set by the system and beyond the learner's control. Findings include that the 'modality effect' (improved comprehension when information is distributed over visual and auditory channels) is weakened or reversed in user-paced studying conditions (Ginns, 2005; Witteman & Segers, 2010).

What if the possibilities for user-control with respect to navigation and repetition are not constrained by the media? Do certain media in themselves – without constraints as time-limits and information transience – afford certain metacognitive strategies? List & Ballenger (2019) compared learning from text sources and learning from video sources with no time limits or constraints on navigation or repetition in either condition, and found differences in what strategies participants used when approaching the material. Text yielded more engagement in information accumulation, more comparison of different information sources and more time spent on sources. One explanation for apparent lack of metacognitive strategy in the video condition provided by the authors rests on video's "linear structure and lack of organizational markers" although participants also engaged in a fair amount of non-sequential navigation in the video sources such as pausing and restarting.

Others have studied reading of printed versus digital media. Singer & Alexander (2016) found that undergraduate students preferred a digital format, which made them read faster and self-assess their comprehension as higher. Their actual comprehension was, however stronger for printed text. Others have found similar effects (Ackerman & Goldsmith, 2011), however Singer Trakhman, Alexander, & Berkowitz (2017) found that the difference in comprehension outcome between the two media disappeared in the absence of time-pressure (given by a narrow time limit). Others have compared different navigation paradigms for visual text presentation on digital media and found that page-by-page navigation impairs comprehension compared to gradual scrolling (Sanchez & Wiley, 2009), by possibly allowing readers to better mentally represent content and structure (Piolat, Roussey, & Thunin, 1997).

Research questions

In the study presented in this paper we set out to explore a number of topics.

1) Do results from previous studies on comprehension of factual information still hold up when differences in information transience, navigation constraints and time limits are reduced between media? Specifically, we explore the following previous results: a) comprehension of factual information is generally improved whenever presentation media includes printed text (Dowell & Shmueli, 2008; Fredriksson, 2015; Taake, 2009), b) comprehension of factual information is less dependent on strong reading ability when reading while listening (Higgins & Zvi, 1995; Montali & Lewandowski, 1996) and c) for comprehension of factual information there is no benefit from seeing an animated character delivering synthesized speech compared to only listening to the speech (Fredriksson, 2015).

2) Do different presentation media elicit metacognitive strategies - such as repetition and non-sequential navigation - differently?

3) Are metacognitive strategies (such as revisiting material) associated with stronger comprehension in different types of media?

Methods

Participants

To explore our research questions, we designed and performed a study to test students' comprehension of factual material that we had adapted from texts on social science topics. In all, 119 (of which 58 female) 13- to- 14-year old students in their first year of Swedish secondary school participated in the study. All participants went to the same school; however 20% of them were also enrolled in a soccer academy under the school's administration. Two participants were excluded due to malfunctioning data logging. All students spoke Swedish fluently and had been enrolled in the Swedish primary educations system for at least 3 years.

Materials and conditions

The texts used in the study were taken from a text book on social science targeted to the actual age group. The texts in the book were also freely available as (audio) recorded readings. Three

sections of similar difficulty and length (each just below 400 words) were selected. They covered the topics 'communication infrastructure', 'global markets and competition', and 'currency and inflation'. We verified that the texts were similar in terms of readability using the LIX readability measure (Björnsson, 1983). The selected topics were part of the participating students' current curriculum but had not yet been introduced to them, and assumed (by their social studies teacher) to be unfamiliar to them.

The texts were converted into five different presentation media that corresponded to the experimental conditions we defined for the study: reading (R), reading-while-listening (RL), listening (L), listening while watching video of speaker (V) and reading-with-manual-scrolling (RS). The realization of the conditions will be described below. To minimize effects of information transience (i.e. how much information was perceptually available at any time and for how long) between our conditions, the texts were divided into short segments (between 12 and 16 segments per text). Each segment consisted of one or two sentences, and the mean number of words per segment was 28.5 (SD = 7.0).

We developed a web application with a tool to enter the texts and corresponding media files (descriptions below) and a frontend interface to present and navigate the material using the Django framework (version 1.8.2) and JavaScript with jQuery (version 1.9.1). The frontend user interface, which participants used during the study phases of the experiment, was designed to minimize differences in navigation opportunities between presentation conditions. The user interface consisted of a main frame in which visual material (text and video) was presented (figure 1, A).

In the R, RS and RL conditions (that all presented text visually), three text segments were visible at the screen at the same time within the main frame, however only one of them, the 'current' segment was presented at the middle of the screen and was readable without considerable effort. This was achieved by drastically lowering the contrast between the text of the surrounding segments (presented above and below the current segment) against the white background (figure 1, A). The purpose of this layout was to present text in a familiar format, with respect to how longer digital text is normally laid out and navigated (vertically), while limiting the range of information visually accessible at any given time. The main frame containing text (in conditions R, RL and RS) spanned approximately 120 mm (height) by 180 mm (width) on the screens used for the experiment.

For the L, V and RL conditions, synthetic speech was generated from the texts using the 'Acapela Box' online tool developed by Acapela Group SA (https://acapela-box.com). The synthetic voice replicated an adult female speaking in a standard Swedish dialect. The average speech rate of the generated output was around 127 words per minute. The mean duration of the segments of speech (corresponding to the text segments) was 13.5 s (SD = 3.1)

The video of the speaker used in the V condition was rendered in Autodesk Maya (version 2014), showing an animated digital 3D character in a frontal view from the torso and up (figure 2). While only the upper arms were visible, both the torso and arms were slightly but visibly animated to match the speech. The character model was created in

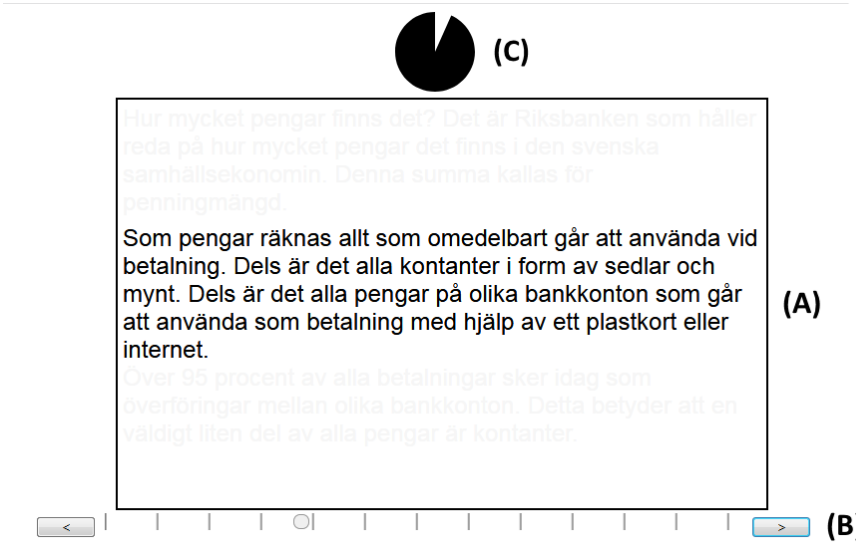

Figure 1. The user interface in the R (reading) and RL (reading while listening) conditions during the study phases of the experiment

Figure 2. The user interface in the V (listening while watching video of speaker) condition during the study phases of the experiment.

Autodesk Character Generator (charactergenerator.autodesk.com, version 2015.2, 2015) matching the gender and approximate age of the voice of the synthesized speech. The character model's hair, clothes and facial features were configured to have a plain appearance and not be distracting. The animation including lip movement and 'visual prosody' (movements of the head and eyebrows following speech prosody; Munhall, Jones, Callan, Kuratate & Vatikiotis-Bateson, 2004) of the character was generated using the FaceFX proprietary software (version 2015.2, 2015). Lip movements were based on a model accounting for coarticulation (Cohen & Massaro, 1993). In condition L the synthetic speech was presented without any video or text. The main frame of the user interface instead contained an animated icon depicting a speaker emitting sound. The main frame spanned approximately 135 mm (height) by 180 mm (width) on the screens used for the experiment for conditions L and V.

The user interface also consisted of a navigation bar (figure 1, B). The navigation bar consisted of a progress indicator, indicating approximately where in the extent of the material the currently presented information appeared. The limits between segments were marked with ticks (lines perpendicular to the span of the navigation bar), that were equally spaced. The time for the progress indicator to move from one tick to the next was determined by the duration of the generated speech for that segment. Once the tick indicating the next segment was reached, the text displayed in the main frame in the R and RL conditions was updated. Thus, the presentation in all the conditions R, RL, L and V was system paced by default, however the user interface allowed participants to freely navigate the material between segments, either by clicking the left and right arrow buttons (indicating a step to the preceding or following segment) or by clicking or dragging the progress indicator to a specific segment on the navigation bar. This navigation and interaction scheme is typical for presenting linear digital material such as audio and video. To reiterate; the level of granularity was determined to minimize differences between conditions and to make the same navigation actions possible for the different media formats.

In the RS (reading with manual scroll) the interface was fully user paced, i.e. the progress indicator text is not changed unless participants performed an explicit navigation. The navigation bar was also displayed vertically to the right of the main frame in the RS condition, to more closely match the typical presentation format for digital text. This condition was added as a control, mainly to the R condition which presented exactly the same information (visual text only) but with a less typical – and by default system paced – navigation interface. In all conditions, participants' behavior while interacting with the material was recorded by logging their navigation interactions to a SQLite database with details about origin- and destination segment as well as timestamps.

The final fixed component of the interface was a pie-chart like countdown timer indicating the remaining time to navigate and study the topic (figure 1, C). The maximum time was 4 minutes per topic. All the texts (when read as synthesized speech) had durations of around three minutes.

In all conditions, participants had the option to terminate the study phase and move on to the questions after having traversed the material at least once.

For each of the three texts, six or seven content questions were formulated by the authors in collaboration with an experienced social studies teacher. These were multiple-choice questions, each with four alternative answers. More than one alternative could be correct and participants could chose as many of the alternative answers they wanted.

As a measure of baseline reading level, we used the participants' summed word- and sentence- level comprehension scores on the standardized Swedish reading comprehension test 'Reading chains' ('Läskedjor'; Jacobson, 2011) taken earlier during the semester. After participating in the experiment participants individually completed a short questionnaire consisting of two parts. The first part concerned their preferred learning media and consisted of eight ratings on 5-point Likert scales ranging from "agree" to "disagree" with no labels for intermediate levels. The statements to be rated all had the form "I learn well when …" followed by a description of a study activity using different types of media. See box 1 for a complete list of statements. The second part of the questionnaire concerned general media habits and consisted of estimations of how many hours per week students spent with different kinds of media: 0-2 hours, 2-4 hours. 4-6 hours, 6-8 hours or more than 8 hours. See box 2 for a complete list of items.

Proceedure

The experimental sessions included three or four participants performing the test in parallel. However in all conditions performance was individual with each participant assigned their own PC with monitor, keyboard, mouse and headphones. The participants were supervised by an experimenter who assured they did not look at each other's' screens or interact with each other during the experiment. The experiment was performed in the Mozilla Firefox browser (version 45) in full screen mode.

Participants in each session were assigned one condition (all participants receiving the same condition) according to a predefined balanced order repeating order. One experimenter had previously assigned each participant an ID number which was used by the other experimenter to assign them to the session in which they would perform the experiment. Participants were given login information which included a unique username and a password on a printed ticket before being seated. They were then given a short introduction to the test and were told that it was important that the remained focused and performed the test individually. Then they were instructed to put on their headphones and log into the system.

Box 1

Items from the 'preferred learning media' part of the questionnaire, to be rated on Likert scales (Translated from Swedish).

I learn well when …

… I watch informative TV / videos

… I read books or printed handouts

… I read on a screen

… I research topics on the web

… someone verbally explains a topic to me

… I listen to someone reading out loud

… I listen to a computer reading out loud

… I read while listening to the same text

After having logged in, the participants received more detailed instructions via text and images on the screen. These instructions were partly customized for each condition, and explained that they would be presented material (in the media corresponding to their assigned conditions) on three different social studies topics on which they would then be tested. The instructions also explained how the navigation interface worked and the time limits they had for each topic.

Box 2

Items from 'general media habits' part of questionnaire, given with instructions to estimate hours spent per week (Translated from Swedish).

Reading books or magazines / newspapers

Listening to radio, podcasts or audiobooks

Watching videos online

Researching topics on the web

Using text-to-speech feature on websites

When they had read the instructions and indicated that they were ready by pressing an onscreen button, participants started the first topic's study phase, followed by the first test phase where they answered

questions related to the first topic, then moved on to the second topic's study phase and so on. The topics (study phase followed by test phase) appeared in the same order for all participants, and each participant was presented with all three topics in the same condition.

Data treatment and analysis

The comprehension score per text was calculated as a sum of all correct answers to the associated comprehension questions divided by the maximum possible score (i.e. the number of correct answers on all questions). These scores were used to analyze effects on comprehension.

We used time spent studying a topic as a measure of repetition of the presented material related to the topic (more specifically the time spent relative the time one sequential read-through takes given the default pacing). As mentioned (see *Materials and conditions*), the default pacing is based on the speech rate of the synthesized speech and determines the rate of progression over segments without explicit participant interaction in all conditions except RS.

All intentional navigations between text segments – made by clicking the back or forward buttons or by clicking or dragging the progress indicator to a specific segment on the navigation bar – were logged. Navigations that were performed stepwise, by a sequence of actions temporally separated by less than 1s, were logged as one "complex" navigation as starting with the first navigation step's start segment and ending with the final navigation step's destination segment. For the current study we were interested in non-sequential navigation, and extracted only those navigations where the start and destination segments were separated by at least two steps. This means that we ignored for example navigations intended to skip to the next segment after finishing reading the current segment or to repeat a segment that had been passed automatically by the default rate of progression.

All statistical analyses were performed in R studio (version 3.5.2, 2018). All p-values were evaluated at an alpha-level of .05.

Results

Comprehension

The overall comprehension test scores (M = .653, SD = .178) indicated that answering the multiple choice questions had been fairly challenging for the participants, and that there was a substantial variation in these scores (one data point excluded due to missing data). 24 participants had been assigned to the R condition, 22 to the RS condition, 24 to the L condition, 24 to the RL condition and 23 to the V condition. Out of the total 117 participants, 12 were excluded from the analysis due to missing baseline reading level scores. The mean baseline reading level score ('Reading chains') of the remaining 105 participants was 9.64 (SD = 2.32) out of a maximum 18.Main effects of condition and baseline reading comprehension, and their interactions were analyzed with ANOVA. The analysis revealed significant main effects of condition: $F(4, 304) = 3.24$, $p = .013$ and baseline reading level: $F(1, 304) = 25.40$, $p < .001$, and a significant interaction effect of condition and reading level, $F(4, 304) = 8.43$, $p < .001$. A post-hoc analysis to examine differences between conditions was performed by pairwise comparisons using Tukey's range test (designed to not inflate probabilities for significance with multiple comparisons, Tukey, 1949). Table 1.a summarizes the results. We found significantly higher comprehension scores for both the R and V conditions compared to the L condition. Also, there were differences in the mean comprehension score between the RS and RL conditions and the L condition, but these were not statistically significant (p = .068 and .091 respectively). There was no difference between the R and RS conditions, or between the R and RL conditions. A post-hoc analysis to examine the differences in influence of baseline reading level between conditions was performed by pairwise comparisons of estimated marginal means of linear trends (Lenth, 2018) with Tukey corrections for multiple comparisons. Table 1.b summarizes the results. We found significantly stronger linear relationships between baseline reading level and comprehension scores for the L condition compared to the RL and RS conditions, indicating that comprehension was more strongly determined by baseline reading level when only listening. There were no other significant differences in influence by baseline reading level on comprehension, neither between R and RL. The influence of baseline reading level in RL was weaker than V, but the difference was not statistically significant (p = .098). Figure 3 shows an overview of the comprehension scores and the effects of baseline reading level for the different conditions.

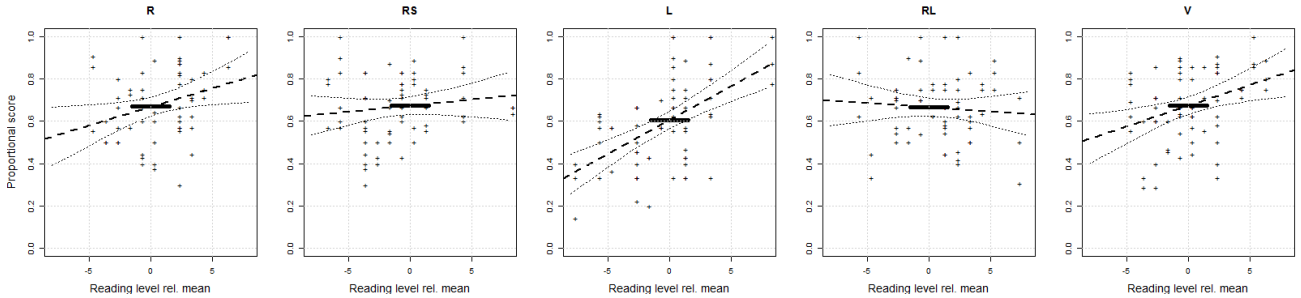

Figure 3. Comprehension scores (proportional, y-axes) plotted against baseline reading level relative global mean for the 5 conditions: reading (R), reading-while-listening (RL), listening (L), listening while watching video of speaker (V) and reading-with-manual-scrolling (RS). Thick solid lines indicate mean comprehension scores per condition, dashed lines the linear relationship between baseline reading level and comprehension scores and dotted lines the 95% confidence intervals.

Table 1.a. Pairwise comparisons between the different conditions' comprehension test scores using Tukey's range test. Highlighted effects are significant according to an alpha-level of .05.

contrast	difference	lower	upper	p
L-RS	-0.077	-0.158	0.003	0.068
RL-RS	-0.004	-0.085	0.078	1.000
W-RS	0.009	-0.072	0.090	0.998
L-R	**-0.090**	**-0.173**	**-0.008**	**0.024**
RL-R	-0.017	-0.100	0.067	0.982
W-R	-0.004	-0.087	0.078	1.000
RL-L	0.074	-0.007	0.154	0.091
W-L	**0.086**	**0.007**	**0.166**	**0.026**
W-RL	0.012	-0.068	0.093	0.993

Table 2.b. Pairwise comparisons differences in influence of baseline reading level on comprehension scores between the different conditions using estimated marginal means of linear trends with Tukey corrections. Highlighted effects are significant according to an alpha-level of .05.

contrast	estimate	SE	t.ratio	p
RS-R	-0.012	0.010	-1.234	0.731
RS-L	**-0.026**	**0.008**	**-3.197**	**0.013**
RS-RL	0.010	0.009	1.084	0.815
RS-V	-0.014	0.009	-1.549	0.532
R-L	-0.014	0.010	-1.481	0.575
R-RL	0.022	0.010	2.125	0.212
R-V	-0.002	0.010	-0.197	0.999
L-RL	**0.036**	**0.009**	**4.113**	**< 0.001**
L-V	0.012	0.009	1.363	0.652
RL-V	-0.024	0.010	-2.479	0.098

Repetition (time spent per topic)

As a measure for repetition, we used the time spent on each topic by the participants, relative to the time one sequential read-through takes with the "default pacing" based on the speech rate of the synthesized speech. The default pacing determines the rate of progression over segments without explicit participant interaction in all conditions except RS. Figure 4 shows an overview of the measured times. We observed a qualitative difference between those conditions that included the synthesized speech (L, RL and V) and the two that included only text (R and RS); the former having a large peak after 0 (i.e. the time one sequential playthrough takes) and the latter measures of time spent having greater variance and approximately normal distributions. A Wilcoxon rank sum test also revealed a statistically significant difference in time spent between the (aggregated) conditions with or without speech (W = 8326, p < .001).

We interpret this general observation as having three implications. First, when the synthetic speech was presented participants seem to have generally followed the default pacing given by speech rate, and not skipped ahead, also when they had visual text available (in the RL condition). Second, given that the distribution of time spent for the R condition is more similar to the RS than the ones with speech, participants in the R condition seem to generally have determined their own pace and not followed the default system pace. Third, at least in the conditions with speech, participants seem to have spent little time repeating material. In fact, the median time spent relative one default-paced sequential read-through, was only ten seconds.

Given the qualitative differences, the time spent per topic was analyzed for L, RL and V separately. We also excluded data points where the time spent relative the default sequential read-through was less than -30 s. Due to the non-normal distributions we categorized the sampled time measures as 'repetition' or 'no repetition' by a median split, and analyzed the effects of condition as a logistic

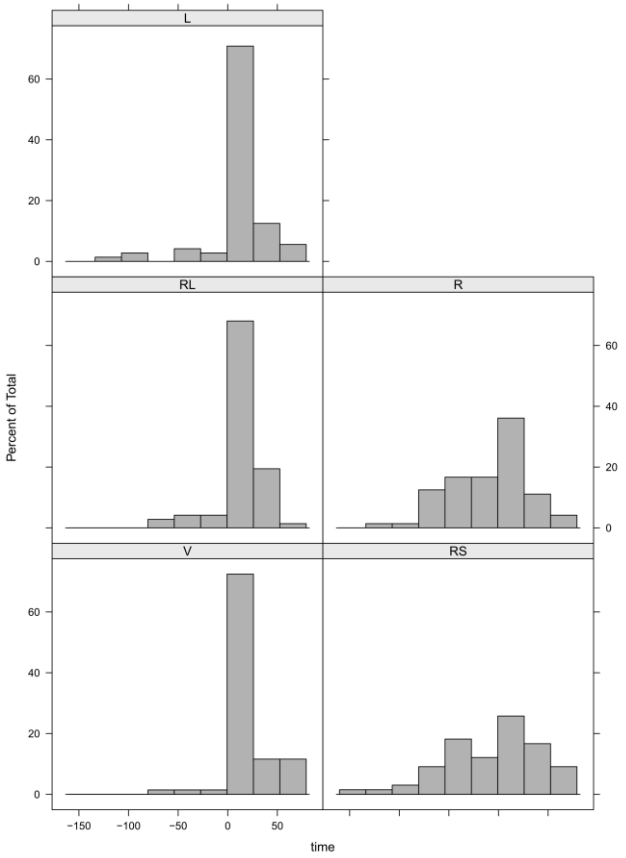

Figure 4. The distributions of time spent on each topic relative the time one read-through of the topic takes for the synthetic voice (the "default pacing") for the different conditions.

binomial regression model with L as base level. The analysis showed a significant difference between L and V conditions (β = .865, z = 2.421, p = .016), indicating increasing odds that the study phases for topics presented with an animated speaker delivering the synthetic speech were more likely to fall into the 'repetition' (more than 10 seconds) category. There was no significant difference between the L and RL conditions (β = -0.434, z = -1.227, p = .220). A Cox-Snell calculation of pseudo R^2 (Cox & Snell, 1989; Signorell, 2016) indicated that the logistic model only accounted for about 6.7 % of the variance in the measured time data. A t-test revealed no difference (t = .116, p = .908) between comprehension scores tested after study phases categorized as 'repetition' (M = .651, SD = .178) and 'no-repetition' (M = .648, SD = .196).

A T-test revealed no difference (t = -0.390, p = .698) in time spent between the two conditions without speech, R (M = -8.79 seconds, SD = 39.21 seconds) and RS (M = -5.91 seconds, SD = 46.27 seconds). Again, there was no significant relationship between time spent and comprehension score (Pearson's R = .070, p = .419).

Non-sequential navigation

To examine non-sequential navigation strategies, we extracted the logged explicit interactions with the navigation interface where participants had moved at least two segments ahead or backwards. It turned out that such interactions were rare. Summing all the navigations for all participants and topics for each group gave us the following numbers: 33 navigations by 24 participants in the R condition, 37 navigations by 22 participants in the RS condition, 22 navigations by 24 participants in the RL condition, 35 navigations by 24 participants in the L condition and 28 navigations by 23 participants in the V condition. While these numbers hint at a slightly lower prevalence of non-sequential in the conditions including the synthetic speech (whose time spent per topic suggested qualitatively different navigation strategies), they are too low to make any meaningful quantitative analysis. Note that the numbers of non-sequential navigations were summed over the study phases of the three topics, the average number of non-sequential navigations were thus only about .5 per topic. Some participants never performed any non-sequential navigation.

Learning- and general media habits

Questionnaire responses were obtained from all except one of the 117 participants. Figure 5 shows the distribution of Likert ratings of statements about the preferred study media (prefixed by "I learn the best when …"). "Someone verbally explaining a topic" had the highest agreement and "listening to a computer reading" had the lowest. A linear regression modelling comprehension score as the sum of the Likert ratings revealed no significant effects, F(8,338) = .755, p = .642, adjusted R^2 = -0.006.

Figure 6 shows the hours per week spent with different media according to participants assessments in the 'general media habits' part of the questionnaire. All except one participant reported using text-to-speech on websites for less than two hours per week, due to the lack of variance this item was excluded from any analysis. Generally, participants reported spending substantially more time watching videos online than consuming any other media. A linear regression modelling comprehension score as the sum of self-assessed hours spent in each media revealed significant positive effects of hours spent reading (β = .030, t = 2.055, p = .041) and watching video (β = .016, z = 2.156, p = .032). The model however only accounted for around 2.4% of the variance in the score data, F(2,338) = 5.106, p = .007, adjusted R^2 = -0.024.

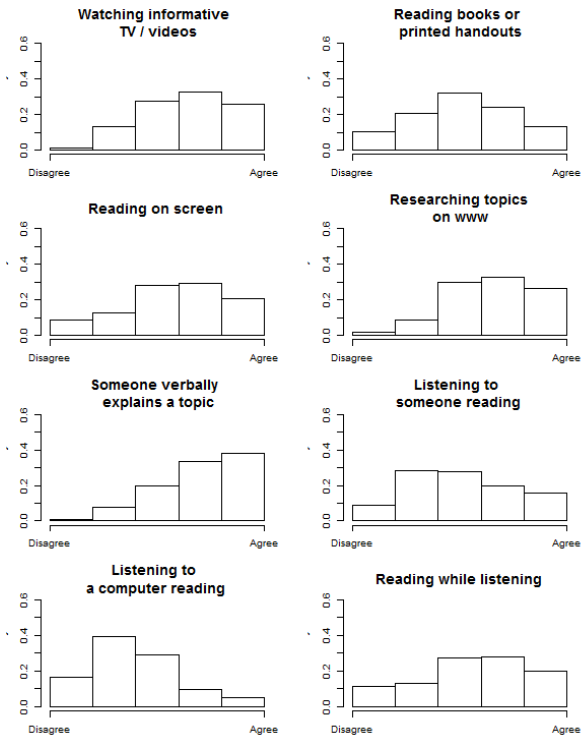

Figure 5. Likert ratings on the 8 statements of the 'preferred learning media' part of the questionnaire.

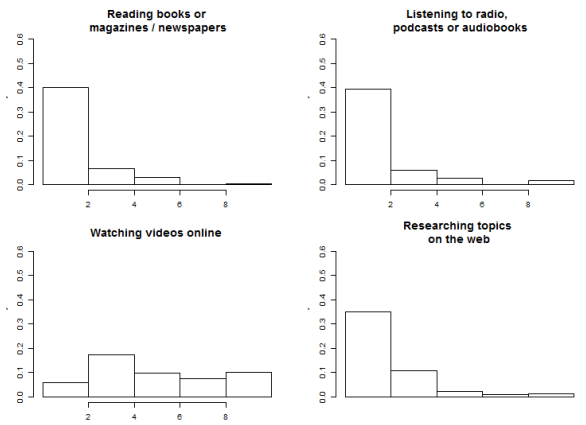

Figure 6. Self-assessed hours per week spent consuming different media from the 'general media habits' part of the questionnaire.

Discussion

The exploratory study did not produce all the data that we expected, and some of the results were inconclusive. There were however some relevant findings. Reading a factual text produced better comprehension than only listening to the text being read. This is in line with previous findings in studies using both natural (Dowell & Shmueli, 2008; Taake, 2009), and synthetic speech (Fredriksson, 2015; Taake, 2009). Our results show that reading still produces better comprehension than listening when differences in information transience and repetition and navigation possibilities are minimized. Also, in line with previous results (Dowell & Shmueli, 2008; Taake, 2009; Fredriksson, 2015), we found no general improvement in comprehension by reading-while-listening (RL) compared to simply reading (R).

More surprisingly, the difference in comprehension between reading-while-listening and only listening was non-significant in our study in contrast to previous findings (Fredriksson, 2015; Taake, 2009; Dowell & Shmueli, 2008). One possible explanation for the divergent results can be the extended navigation possibilities in the current study compared to previous work. Our participants could repeat what they heard if and when they chose to. Perhaps this allowed participants in the listening condition to adopt strategies that compensate for a reduction in working memory load by having the text available, similarly to how the modality effect is eliminated under user-paced studying conditions (Ginns, 2005; Witteman & Segers, 2010). We did however observe a trend in the expected direction ($p = .09$) and it is possible that a larger sample would have revealed a significant difference.

We found no difference in comprehension when comparing the R and RS condition, where the only difference is that in the R condition has a "default" pacing (determined by the speech rate in the L and RL conditions: 127 words per minute) by which the presentation of the text moves on to next segment, whereas the RS condition is completely user paced. This is hardly surprising considering the time spent per text indicates that participants in the R condition have ignored the default pacing and moved on as soon as they finished reading a segment, which they generally seem to have done faster than the default pacing (fig 4). In contrast, the RL group seem to have followed the default pacing, indicating both that they did in fact listen to the speech playing along with the text and that they let it determine the pacing. It is possible that a positive effect of the added modality (synthetic speech) was counteracted by the lack of user pacing, or time left for repetition after one 'read-through', compared to the R condition. Reitsma (1988) found that first-graders improved their reading accuracy more when trained when they themselves could control when and what to hear read together with the text compared to reading-while-listening.

The ANOVA revealed a strong main effect of baseline reading level on comprehension on condition test score which, taken together with the significant positive linear relationship between self-assessed time spent reading per week, suggests that reading skill is generally useful for learning in other media. Reading comprehension has been linked to working memory capacity in children (Seigneuric, Ehrlich, Oakhill & Yuill, 2000) however in most cases specific to the language domain (Nation, Adams, Bowyer-Crane & Snowling, 1999; Nation,

Clarke & Snowling, 2002). Baseline reading level had the strongest influence on comprehension score in the listening condition (L) – significantly stronger compared to RS and RL) – and the listening (without visible text) is also arguably the condition which places more demands on (verbal) working memory. Our results indicate that this is also the case when allowing for similar possibilities for repetition and navigation.

The comprehension scores in the reading-while-listening condition was the least influenced by baseline reading level, however we found no significant differences in influence by baseline reading level on comprehension, between R and RL ($p = 0.212$). Previous findings would suggest a smaller influence for RL, since poor reader benefit from listening while reading (Fredriksson, 2015; Higgins & Zvi, 1995; Montali & Lewandowski, 1996). We did however see a trend in this direction, with baseline reading comprehension basically having no influence on comprehension score in the RL group (figure 3). A larger sample and comparisons within the group of students with poor baseline reading comprehension would have given us a stronger basis for any conclusive statements.

Another surprising finding was that the video condition (V), combining synthesized speech with synthesized animation of a digital character, resulted in better comprehension scores compared to the listening-only condition (L). A seemingly straightforward explanation for this would be that visual cues such as lip movements help speech recognition and thereby facilitates comprehension. However, it is not obvious that such an effect is expected when it comes to comprehension. Nirme et al. (2019), found that a similar digitally animated speaker – however animated and voiced based on recordings of a real speaker – only improved comprehension when presented with background babble noise. It is however possible that the synthetic voice constitutes a comparably challenging listening condition (Drager et al., 2010; Mattys et al., 2012). The observed difference contrasts the results of our preliminary study (Fredriksson, 2015), in which no repetition or non-sequential navigation was possible. The higher odds of repetition (measured by time spent per topic) in the video condition is therefore another feasible explanation. However, we found no significant relationship between repetition and comprehension test outcome. Yet another - but perhaps related – explanation could be that the video condition promoted engagement and attention to the material by attribution of 'social agency' to the animated character (Moreno et al., 2001). Also, the high number of self-reported hours spent watching videos in the general media habits questionnaire (fig 6) indicates that our participants were quite familiar with the video format compared to strictly audio-based media, and - especially given the positive relationship between hours spent watching videos - we cannot rule out that this might have influenced the result.

Even though the participants self-reported spending virtually no time "using text-to-speech feature on websites" we did not control for other kinds of exposure and familiarity with synthetic speech, which has previously been shown to affect word recognition (McNaughton, Fallon, Tod, Weiner, & Neisworth, 1994).

Apart from previous exposure serving perceptual learning with regards to synthetic speech (Schwab, Nusbaum & Pisoni, 1985), it may also affect the level of trust in the artificial speaker. Craig, Chiou & Schroeder (2019) found that a 'virtual human' coupled with a human voice produced higher "trust score" (obtained by questionnaire described in Jian, Bisantz & Drury, 2000) compared to both high and low quality synthetic voices. These results contrast those of a previous study where Craig & Schroeder (2017) had found that both virtual characters presented with a real voice and a high-quality synthetic voice were perceived as more 'credible' (measured by the Agent Persona Inventory; Ryu & Baylor, 2005) than a character with a low quality synthetic voice. Authors speculate that the different results might be due to the measure of credibility drawing specific attention to the character and point out that the results might not be generalizable to situations where the speaker is not presented visually. Torre, Goslin, White and Zanatto (2018) found indications that trust is shaped by the initial interactions with an artificial speaker (digital character or robot).

The results of the general media habits questionnaire results should be interpreted with some general observations about the sampled population in mind. The students at this age have often disagreements with their parent about amount of use of internet, for games and such. This could have biased their answers to the general media habits questionnaire, as the students could have deliberately minimized the time spent on internet activities. The positive side of it (if this assumption is correct), is that the students are aware of the drawbacks of spending too much time on the computer, playing games. It can also be so that the scale in the questionnaire was too optimistic. For students at this age, with small or none homework, zero hours or two hours makes a big difference. Homework takes them about 15 minutes to do. The intervals should have been 1 hour instead of 2. It can also be so that the students do all the five activities during a week, a little bit of everything every day. So they tried to estimate an approximate distribution between these activities. For example, those who have the lowest amount of time for all the activities maybe spend up to 10 hours a week all together, but doing different things.

Concerning the results of the 'preferred learning media' questionnaire, the fact that we did not find any relationship between learning media preferences and comprehension scores is not surprising given previous research showing that media preferences do not predict or reflect improved comprehension outcome, be it on-screen digital text (Singer & Alexander, 2016) or video (Salomon, 1984) over printed text as also observed in the responses we collected. This also relates to the absence of any clear benefit from reading-while-listening compared to reading, given the increasingly common practice to give learners the option to listen to text being read instead of reading themselves. While we do not rule out that it might be beneficial some (e.g. poor readers as reading-while-listening seems to be less dependent on strong reading skills), previous research suggests that learners often lack the metacognitive skills to themselves make optimal decisions about their learning (Bjork, Dunlosky & Kornell, 2013; Kruger & Dunning, 1999). Moreover, a preference for alternative media over independent reading might in fact be a missed opportunity to practice reading skills, which would be unfortunate. Both our results, and previous research (Garner, 1987), indicate that reading skills are linked to successful learning and metacognition.

It is worth mentioning that "listening to someone verbally explain a topic" had a stronger preference compared to both "listening to someone reading" or "listening to a computer reading". Perhaps it is possible that a visually presented 'agent', if developed to a broader and more interactive behavior repertoire, could replicate sense of being personally addressed with is attractive to the listener. This again relates to the idea of social agency promoting engagement (Moreno et al., 2001). It might also better support comprehension to combine an animated speaker with on screen text, as suggested by meta-analysis by Schroeder, Adesope, and Gilbert (2013), which however is possibly at odds with the 'modality effect' (Moreno & Mayer, 1999). Regarding the visualization of the speaker, a less realistic and more stylized design arguably might be preferable in educational context (Gulz & Haake, 2006a).The specific quality of both the synthetic speech and the visual presentation and animation of the speaker likely had some effect on the outcome in the current study. Technological solutions in these areas are continuously developing both in terms of quality and availability. For our implementations we used proprietary software), however open/free alternatives are available (example for audiovisual speech synthesis: Cudeiro, Bolkart, Laidlaw, Ranjan & Black, 2019).

Overall, there was less repetition and non-sequential navigation than we expected in the collected data. It would therefore be dubious to draw any hard conclusions from those results. We can however speculate as to why there was so little repetition and non-sequential navigation with the current experimental setup. It is possible that the navigation interface, with a focus on the navigation bar, affords (in Norman's sense, Norman, 1988) a sequential conceptualization of the material's structure. In all conditions except reading-with-manual-scrolling (RS) the navigation bar was horizontally distributed, which is typical for video and audio presentation of linear media. Still, although the RS condition generated the highest number of non-sequential navigations (37 in total), they were not substantially more numerous than in the other conditions and a lot fewer than we had expected. An extended interface allowing for navigation more typical for 'hypermedia' could potentially be better to study metacognitive and self-regulation strategies (Azevedo, 2005; Gerjets et al., 2009). For example, one could visualize and enable keyword search to ease navigation in lecture videos (Tuna, Subhlok, & Shah, 2011).

Another possibility is that the maximum time given to study each topic (4 minutes) was insufficient to make revisiting information worthwhile or those participants down-prioritized due to fatigue or to strategically avoid fatigue. Repetition less likely for later topics, which Risko et al. (2012) has shown is associated with increased mind-wandering. However the order of the topics was identical and not counter-balanced in the current study, so the decrease in repetition could be a caused by the perceived difficulty of the individual topics. The measure we used to compare the readability of the topics (LIX) is based exclusively on a set of surface properties, which might be too narrow (Mühlenbock & Johansson Kokkinakito, 2009) to be able rule out that some topic is perceived as easier or more difficult to understand. The vocabulary used in the initial topic – communication infrastructure – might for example have been more familiar to the participants. The type of topic, e.g. narrative vs factual, might also influence metacognitive strategy and tendency for mind wandering (Szpunar, Moulton, & Schacter, 2013).

Whatever the reason, the low frequency of non-sequential navigation rendered a planned qualitative analysis of participants' metacognitive representation of the topics' content and structure meaningless. Nevertheless, we see this as an interesting area for future investigation, specifically how the type of media affects stronger representation of the structure and links of the material itself ("where to find things" cf. Sparrow, Liu & Wegner, 2011) or to the information on the topic that the material presents (Zwaan & Radvansky, 1998). The findings of McNamara & Kintsch (1996) suggest that the structure of texts interacts with previous knowledge, in that a less coherent structure can promote inference processes in learners with high previous knowledge.

Despite its limitations, our study contributes to the understanding of how comprehension works in different media by replicating some - and contrasting other – previous findings, under conditions where information transience and differences in navigation possibilities were minimized. A way to further control for these factors could be to test a condition with text scrolling at the same rate as the synthesized speech. In the current exploratory study we however prioritized conditions representing media that could have feasible applications in real learning situations. Also, the level of granularity of the segmentation enables quantitative comparisons between navigation across conditions. The platform has potential as a research tool to connect comprehension and learning to low-level behavior since all navigation actions are logged, (cf. Kim et al., 2014). It could be extended with, for example, eye-tracking data (Lai et al., 2013) or further navigation options (Tuna, Subhlok, & Shah, 2011).

In conclusion, visual presentation of text has a strong positive effect on comprehension compared to strictly audial presentation, also when minimizing differences in information transience and navigation possibilities. However, the possibilities for repetition and navigation afforded by the interface together with the type of media might influence comprehension strategies (e.g. being more likely to repeat material presented by a visual speaker) as well as demands on baseline reading skill. The clearest effect of the media themselves we found was the qualitative difference in navigation behavior between condition with and without speech: participants tended to follow the pace given by the speech also when text was visually available and "skippable". Seeing an animated speaker delivering synthetic speech improved comprehension to a similar degree as seeing text did. Previous research into under what conditions comprehension is aided by audiovisual presentation of speech is inconclusive and our results have several feasible explanations. There is a still lot of work to be done mapping out effects of different presentation media and navigation interfaces have on comprehension, learning and development of metacognitive strategies. Although the range of outcomes of controlled studies, suggest they should be applied with caution, new technologies such as real-time speech synthesis and pedagogical agents with ever expanding repertoire of behaviors hold promise as tools to adapt instruction to the varying needs of students and for controlled study of their effects.

References

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: on screen versus on paper. *Journal of Experimental Psychology: Applied, 17*(1), 18–32.

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24*, 1–3.

Allen, E., & Seaman, J. (2010).Learning on demand: Online education in the United States 2009. Needham, MA: Sloan Consortium.

Allison, C. (2015). The use of instructional videos in K-12 classrooms: A mixed-method study (Doctoral dissertation, Indiana University of Pennsylvania).

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational psychologist*, *40*(4), 199-209.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444.

Björnsson, C. H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*, 480-497.

Bol, L., & Hacker, D. J. (2012). Calibration research: where do we go from here? *Frontiers in Psychology, 3*, 229.

Cassell, J., Sullivan, J., Churchill, E., & Prevost, S. (2000). *Embodied Conversational Agents*. MIT Press.

Clark, R. E., & Choi, S. (2005). Five Design Principles for Experiments on the Effects of Animated Pedagogical Agents. *Journal of Educational Computing Research, 32*(3), 209–225.

Cohen, M. M., & Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In *Models and techniques in computer animation* (pp. 139-156). Springer, Tokyo.

Cohen, M. M., Walker, R. L., & Massaro, D. W. (1995). Perception of synthetic visual speech, speechreading by man and machine: Models, systems and applications. In *NATO Advanced Study Institute* (Vol. 940584).

Cox, D.R., Snell, E. J. (1989). *Analysis of Binary Data.* Second Edition. Chapman & Hall.

Craig, S. D., Gholson, B., & Driscoll, D. M. (2002). Animated pedagogical agents in multimedia educational environments: Effects

of agent properties, picture features and redundancy. Journal of Educational Psychology, 94(2), 428.

Craig, S. D., Chiou, E. K., & Schroeder, N. L. (2019). Impact of virtual human's voice on learner's trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (in press). Los Angeles, CA: SAGE Publications.

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education, 114*, 193-205.

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., & Black, M. J. (2019). Capture, Learning, and Synthesis of 3D Speaking Styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 10101-10111).

Domagk, S. (2010). Do Pedagogical Agents Facilitate Learner Motivation and Learning Outcomes? *Journal of Media Psychology, 22*(2), 84–97.

Dowell, J., & Shmueli, Y. (2008). Blending speech output and visual text in the multimodal interface. *Human Factors, 50*(5), 782–788.

Drager, K. D. R., Reichle, J., & Pinkoski, C. (2010). Synthesized speech output and children: a scoping review. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association, 19*(3), 259–273.

Dunsworth, Q., & Atkinson, R. K. (2007). Fostering multimedia learning of science: Exploring the role of an animated agent's image. *Computers & Education, 49*(3), 677-690.

Elkind, J., Cohen, K., & Murray, C. (1993). Using computer-based readers to improve reading comprehension of students with dyslexia. *Annals of Dyslexia, 43*(1), 238–259.

Feng, S., D'Mello, S., & Graesser, A. C. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review, 20*(3), 586–592.

Fredriksson, O. (2015). Looking for a more effective way of learning: reading versus listening – How different modes of presentation of factual texts using synthesized speech influence text learning for Swedish 8th-graders. Master's Thesis in Cognitive Science, Div. of Cognitive Science (LUCS), Lund University.

Garau, M., Slater, M., Pertaub, D. P., & Razzaque, S. (2005). The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators & Virtual Environments, 14*(1), 104-116.

Garner, R. (1987). *Metacognition and reading comprehension*. Norwood, NJ: Ablex Publishing.

Gerjets, P., Scheiter, K., Opfermann, M., Hesse, F. W., & Eysink, T. H. S. (2009). Learning with hypermedia: The influence of representational formats and different levels of learner control on performance and learning behavior. *Computers in Human Behavior, 25*(2), 360–370.

Gillström, Å., & Rönnberg, J. (1995). Comprehension calibration and recall prediction accuracy of texts: Reading skill, reading strategies, and effort. *Journal of Educational Psychology, 87*(4), 545–558.

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and instruction, 15*(4), 313-331.

Gisterå, E.M. (1995): Dyslexi och dyskalkyli: Utvärdering av läromedelskassetter för elever med läs- och skrivsvårigheter. Uppsala: Pedagogiska institutionen.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108* (3), 1197–1208.

Gulz, A. (2005). Social enrichment by virtual characters–differential benefits. *Journal of Computer Assisted Learning, 21*(6), 405-418.

Gulz, A., & Haake, M. (2006a). Virtual pedagogical agents: naturalism vs. stylization. In *International Workshop on Intelligent Virtual Agents* (pp. 455-455). Springer, Berlin, Heidelberg.

Gulz, A., & Haake, M. (2006). Design of animated pedagogical agents—A look at their look. *International Journal of Human-Computer Studies, 64*(4), 322-339.

Handley, Z. (2009). Is text-to-speech synthesis ready for use in computer-assisted language learning?. *Speech Communication, 51*(10), 906-919.

Hecker, L., Burns, L., Katz, L., Elkind, J., & Elkind, K. (2002). Benefits of assistive reading software for students with attention disorders. *Annals of dyslexia, 52*(1), 243-272.

Higgins, E. L., & Zvi, J. C. (1995). Assistive technology for postsecondary students with learning disabilities: From research to practice. *Annals of Dyslexia, 45*(1), 123–142.

Hilbert, S., Nakagawa, T. T., Puci, P., Zech, A., & Bühner, M. (2015). The Digit Span Backwards Task. *European Journal of Psychological Assessment, 31*(3), 174–180.

Jacobson, C. (2011). Läskedjor. Stockholm: Hogrefe Psykologiförlaget.

Jansen, S., Chaparro, A., Downs, D., Palmer, E., & Keebler, J. (2013, September). Visual and cognitive predictors of visual enhancement in noisy listening conditions. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 1199-1203). Sage CA: Los Angeles, CA: SAGE Publications.

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53-71.

Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, W.-Y.et al(2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012.*Educational Research Review,10*, 90–115.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*(4), 351–371.

Kim, J., Li, S. W., Cai, C. J., Gajos, K. Z., & Miller, R. C. (2014). Leveraging video interaction data and content analysis to improve video learning. In *Proceedings of the CHI 2014 Learning Innovation at Scale workshop*.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Leahy, W., & Sweller, J. (2011). Cognitive load theory, modality of presentation and the transient information effect. *Applied Cognitive Psychology*, *25*(6), 943-951.

Lenth, R. (2018). Emmeans: Estimated marginal means, aka least-squares means. R package version 1.1.

Lester, J. C., Converse, S. A., Kahler, S. E., Todd Barlow, S., Stone, B. A., & Bhogal, R. S. (1997). The persona effect. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '97.

List, A., & Ballenger, E. E. (2019). Comprehension across mediums: the case of text and video. Journal of Computing in Higher Education, 1-22.

Mattheyses, W., & Verhelst, W. (2015). Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication, 66*, 182-217.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7-8), 953-978.

Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied, 18*(3), 239–252.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes*, *22*(3), 247-288.

McNaughton, D., Fallon, K., Tod, J., Weiner, F., & Neisworth, J. (1994). Effect of repeated listening experiences on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 10*(3), 161-168.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*(3), 159-163.

Miller, J. (1982). Divided attention: evidence for coactivation with redundant signals. *Cognitive Psychology, 14*(2), 247–279.

Mishra, S., Lunner, T., Stenfelt, S., Rönnberg, J., & Rudner, M. (2013). Visual Information Can Hinder Working Memory Processing of Speech. Journal of Speech, Language, and Hearing Research, 56, 1–13.

Montali, J., & Lewandowski, L. (1996). Bimodal reading: benefits of a talking computer for average and less skilled readers. Journal of Learning Disabilities, 29(3), 271–279.

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. Journal of Educational Psychology.

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*(1), 156–163.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?. *Cognition and instruction*, *19*(2), 177-213.

Moussa-Inaty, J., Ayres, P., & Sweller, J. (2012). Improving listening skills in English as a foreign language by reading rather than listening: A cognitive load perspective. *Applied Cognitive Psychology*, *26*(3), 391-402.

Mühlenbock, K., & Kokkinakis, S. J. (2009). LIX 68 revisited-An extended readability measure. In *Proceedings of Corpus Linguistics*.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological Science, 15*(2), 133–137.

Nation, K., Adams, J. W., Bowyer-Crane, C. A., & Snowling, M. J. (1999). Working memory deficits in poor comprehenders reflect underlying language impairments. *Journal of experimental child psychology, 73*(2), 139-158.

Nation, K., Clarke, P., & Snowling, M. J. (2002). General cognitive ability in children with reading comprehension difficulties. *British Journal of Educational Psychology, 72*(4), 549-560.

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation, 26,* 125-173.

Nirme, J., Haake, M., Lyberg Åhlander, V., Brännström, J., & Sahlén, B. (2019). A virtual speaker in noisy classroom conditions: supporting or disrupting children's listening comprehension?. *Logopedics Phoniatrics Vocology, 44*(2), 79-86.

Norman, D. A. (1988).*The psychology of everyday things.* New York: Basic Books .

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition, 17*(4), 398–422.

Piolat, A., Roussey, J.-Y., & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies, 47*(4), 565–589.

Reed, D. K., Swanson, E., Petscher, Y., & Vaughn, S. (2013). The effects of teacher read-alouds and student silent reading on predominantly bilingual high school seniors' learning and retention of social studies content. *Reading and Writing, 27*(7), 1119–1140.

Reitsma, P. (1988). Reading Practice for Beginners: Effects of Guided Reading, Reading-While-Listening, and Independent Reading with Computer-Based Speech Feedback. *Reading Research Quarterly, 23*(2), 219.

Risko, E. F., Anderson, N., Sarwal, A., Engelhardt, M., & Kingstone, A. (2012). Everyday attention: Variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology, 26*(2), 234-242.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex , 17*(5), 1147–1153.

Ryu, J. E. & Baylor, A. L. (2005). The psychometric structure of pedagogical agent persona. *Technology Instruction Cognition and Learning, 2*(4), 291.

Salomon, G. (1984). Television is" easy" and print is" tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of educational psychology, 76*(4), 647.

Sanchez, C. A., & Wiley, J. (2009). To Scroll or Not to Scroll: Scrolling, Working Memory Capacity, and Comprehending Complex Texts. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 51*(5), 730–738.

Scagnoli, N. I., Choo, J., & Tian, J. (2019). Students' insights on the use of video lectures in online classes. *British Journal of Educational Technology, 50*(1), 399-414.

Schacter, D. L., & Szpunar, K. K. (2015). Enhancing attention and memory during video-recorded lectures. *Scholarship of Teaching and Learning in Psychology, 1*(1), 60–71.

Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research, 49*(1), 1-39.

Schneeberg, H. (1977). Listening while reading: A four year study. *The Reading Teacher, 30*(6), 629-635.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human factors, 27*(4), 395-408.

Seigneuric, A., Ehrlich, M. F., Oakhill, J. V., & Yuill, N. M. (2000). Working memory resources and children's reading comprehension. *Reading and writing, 13*(1-2), 81-103.

Signorell, A. (2016). DescTools: Tools for descriptive statistics. R package version 0.99, 17.

Singer, L. M., & Alexander, P. A. (2016). Reading Across Mediums: Effects of Reading Digital and Print Texts on Comprehension and Calibration. *Journal of Experimental Education, 85*(1), 155–172.

Singer Trakhman, L. M., Alexander, P. A., & Berkowitz, L. E. (2017). Effects of Processing Time on Comprehension and Calibration in Print and Digital Mediums. *Journal of Experimental Education,* 1–15.

Singh, A.-M., Marcus, N., & Ayres, P. (2012). The Transient Information Effect: Investigating the Impact of Segmentation on

Spoken and Written text. *Applied Cognitive Psychology, 26*(6), 848–853.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776-778.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america, 26*(2), 212-215.

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science, 12*(2), 257–285.

Szpunar, K. K., Moulton, S. T., & Schacter, D. L. (2013). Mind wandering and education: from the classroom to online learning. *Frontiers in Psychology, 4*, 495.

Taake, K. P. (2009). A comparison of natural and synthetic speech: With and without simultaneous reading. Thesis. Washington University.

Torre, I., Goslin, J., White, L., & Zanatto, D. (2018, April). Trust in artificial voices: A congruency effect of first impressions and behavioural experience. In *Proceedings of the Technology, Mind, and Society* (p. 40). ACM.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*(2), 99-114.

Tuna, T., Subhlok, J., & Shah, S. (2011). Indexing and keyword search to ease navigation in lecture videos. 2011 IEEE Applied Imagery Pattern Recognition Workshop (AIPR).

Unsworth, S. J., & Pexman, P. M. (2003). The impact of reader skill on phonological processing in visual word recognition. *The Quarterly Journal of Experimental Psychology, 56*(1), 63–81.

Vagula, M., & Liu, H. (2016). Enhancing the learning experiences of undergraduate students through pre-recorded video lectures. *HAPS Educator, 20*(3), 45-48.

Van Orden, G. C., Johnston, J. C., & Hale, B. L. (1988). Word identification in reading proceeds from spelling to sound to meaning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 14*(3), 371–386.

Winters, S., Pisoni, D. (2005). Speech synthesis:Perception and comprehension. In Brown, K.,(ed), *Encyclopedia of Language and Linguistics, 12*, 31–49.

Witteman, M. J., & Segers, E. (2010). The modality effect tested in children in a user-paced multimedia environment. *Journal of Computer Assisted Learning, 26*(2), 132-142.

Wong, A., Leahy, W., Marcus, N., & Sweller, J. (2012). Cognitive load theory, the transient information effect and e-learning. *Learning and Instruction, 22*(6), 449-457.

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In *Handbook of metacognition in education* (pp. 311-328). Routledge.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. Psychological Bulletin, 123(2), 162–185.