Journal of Research in Reading
UKLA
The United Kingdom Literacy Association

Journal of Research in Reading, ISSN 0141-0423
Volume 00, Issue 00, 2019, pp 1–38
DOI:10.1111/1467-9817.12269

Reading from paper compared to screens: A systematic review and meta-analysis

Virginia Clinton

University of North Dakota, Grand Forks, ND, USA

Background: Given the increasing popularity of reading from screens, it is not surprising that numerous studies have been conducted comparing reading from paper and electronic sources. The purpose of this systematic review and meta-analysis is to consolidate the findings on reading performance, reading times and calibration of performance (metacognition) between reading text from paper compared to screens.
Methods: A systematic literature search of reports of studies comparing reading from paper and screens was conducted in seven databases. Additional studies were identified by contacting researchers who have published on the topic, by a backwards search of the references of found reports and by a snowball search of reports citing what was initially found. Only studies that were experiments with random assignment and with participants who had fundamental reading skills and disseminated between 2008 and 2018 were included. Twenty-nine reports with 33 identified studies met inclusion criteria experimentally comparing reading performance (k = 33; n = 2,799), reading time (k = 14; n = 1,233) and/or calibration (k = 11; n = 698) from paper and screens.
Results: Based on random effects models, reading from screens had a negative effect on reading performance relative to paper ($g = -.25$). Based on moderator analyses, this may have been limited to expository texts ($g = -.32$) as there was no difference with narrative texts ($g = -.04$). The findings were similar when analysing literal and inferential reading performance separately ($g = -.33$ and $g = -.26$, respectively). No reliable differences were found for reading time ($g = .08$). Readers had better calibrated (more accurate) judgement of their performance from paper compared to screens ($g = .20$).
Conclusions: Readers may be more efficient and aware of their performance when reading from paper compared to screens.

Keywords: paper reading, screen reading, digital reading, systematic review, meta-analysis

Highlights

What is already known about this topic

- Reading from screens is common.
- Many concerns exist about effects of medium on reading.
- Numerous studies on the topic have been conducted.

What this paper adds

- Small benefit of paper on reading performance.
- No difference in reading times.
- Small benefit of paper on metacognition.

Implications for theory, policy or practice

- Reading from paper may be more efficient.
- Better understanding of why paper has reading benefits is needed.

Reading has primarily been from paper until relatively recent advances in technology have brought about a number of electronic sources with screens for reading, such as e-readers, computers and tablets (Shenoy & Aithal, 2016). Reading from these screens has become increasingly prevalent for both educational and recreational reading (Hyman, Moser, & Segala, 2014), but is that desirable? A review conducted in 2008 concluded that reading from paper was superior to reading from screens, but the authors noted that technological advances could change that pattern (Noyes & Garland, 2008). Many studies have examined the issues related to reading from electronic versus paper sources in terms of performance on reading assessments and how the text is read (i.e., the process of reading). The purpose of this systematic literature review and meta-analysis is to examine differences in performance and processes between reading from paper and screens.

Reading text from screens, also called digital reading or reading electronic text, has been controversial. Advocates have emphasised the lower costs of reading from screens. Indeed, cost is a driving force in the use and development of electronic texts (Hancock, Schmidt-Daly, Fanfarelli, Wolfe, & Szalma, 2016). For example, electronic books are typically more cost-effective than paper books (Bando, Gallego, Gertler, & Romero, 2016). In addition, text can be easily accessed electronically from various devices (e.g., e-readers, computers and smartphones), and a single device can hold a large number of books or other reading material. These characteristics give electronic medium some advantages over paper in terms of convenience in access and transport of reading material.

Despite these advantages, there are staunch critics who have concerns about reading from screens (e.g., Herold, 2014). The experience of reading from screens is frequently described as less pleasant and less engaging than that of reading from paper (Mangen & Kuiken, 2014). For example, college students tend to prefer reading from paper rather than screens (Kazanci, 2015) and will usually opt for paper textbooks rather than less expensive electronic versions (Rockinson-Szapkiw, Courduff, Carter, & Bennett, 2013). One longstanding concern is that of eye strain from reading text from screens (Ziefle, 1998), although this is becoming less of an issue with advances in screen technology (Rosenfield, Jahan, Nunez, & Chan, 2015). There are also concerns that reading time is longer from screens than paper, but without any relative benefit to comprehension (Daniel & Woody, 2013). Some critics argue that there is *screen inferiority*, in which readers have weaker performance and metacognitive awareness of their performance, on assessments based on reading from screens compared to paper (Ackerman & Lauterman, 2012). Given these issues, it is not surprising that some argue that reading from screens may only be suitable for easy light reading (Baron, 2015).

Given these controversies, a cohesive understanding of findings comparing reading from paper and screens is necessary. In this systematic literature review and meta-analysis, research findings on performance on reading assessments, as well as two reading processes (reading time and metacognition), are examined.

Performance

Performance on reading comprehension assessments is considered to measure how well a text was understood. This is often referred to as the products of reading that are based on the mental representation of the text after it is read (Rapp & van den Broek, 2005). Reading comprehension assessments take a variety of forms such as multiple-choice questions, open-ended questions and free recalls (van den Broek, White, Kendeou, & Carlson, 2009). Performance on reading comprehension assessments involves a variety of interdependent skills (Kintsch, 2012). This variety of skills are due to the multidimensional nature of reading comprehension in which identifying words, developing ideas from words and connecting those ideas together are all necessary to create a coherent mental representation of a text (Magliano, Millis, Ozuru, & McNamara, 2007). Some skills, such as word identification, may not be specifically assessed, but reading performance is dependent upon them (Perfetti & Adlof, 2012).

Generally speaking, reading assessments and their items can be categorised as involving literal or inferential understanding (Bowyer-Crane & Snowling, 2005; Elleman, 2017; MacGinitie, MacGinitie, Cooter, & Curry, 1989). Assessments that measure literal understanding do not require the reader go beyond what the text explicitly states and typically focus on memory of the text (Hosp & Suchey, 2014). Free recall, multiple-choice and open-ended questions that prompt readers to retrieve information explicitly stated in the text are methods that assess literal understanding (Tarchi, 2017). In contrast, assessment items that measure inferential comprehension require the reader to make connections within the text or between the text and background knowledge (Cain, Oakhill, & Bryant, 2004; Clinton & van den Broek, 2012). Questions that go beyond individual ideas stated in the text and require integration of ideas are thought to assess inferential understanding (Stevens et al., 2015). Measures of literal understanding are typically easier than measures of inferential understanding (Basaraba, Yovanoff, Alonzo, & Tindal, 2013). Given that some argue reading from screens is particularly detrimental for challenging reading tasks (Baron, 2015), it is possible that differences between media in performance would be noted for inferential, but not literal understanding measures.

Process

A critical area of inquiry in reading research involves how the text is read, in other words, the *process* of reading (van den Broek et al., 2009). The process of reading can be examined in many ways, but two commonly used in comparing reading text from screens and paper are reading time and metacognitive accuracy. Reading time is a commonly used process measure in which longer reading times are considered indicative of increased processing (Rapp & van den Broek, 2005) and increased effort (Chen & Catrambone, 2015). For example, reading times are longer for text that is inconsistent with previously read text or background knowledge, which indicates more processing and effort is involved relative to reading text that is consistent (Albrecht & O’Brien, 1993). In addition, examinations of reading times for processing text from screens and paper provide a helpful context for interpreting performance findings – one can infer the amount of processing or effort based on reading times invested to achieve performance. Moreover, reading time measures can also be used to infer the efficiency of reading, in which performance on assessments is examined in relation to the time spent reading (e.g., Ackerman & Lauterman, 2012). In this way, reading time can be used to compare the amount of processing or effort that went into the task and efficiency of that processing between reading text from paper compared to screens.

Likely because of the usefulness of reading time in examining the process and efficiency of reading, differences in reading time for text from paper and screens have been examined for decades (Reinking, 1988). A synthesis by Dillon in 1992 concluded that reading text from screens is slower than paper because of the awkwardness of reading from screens compared to paper. However, comparisons of reading times by medium have been inconsistent in empirical findings since that review, with longer reading times for paper in some (e.g., Chen & Catrambone, 2015; Singer Trakhman, Alexander, & Berkowitz, in press; Singer Trakhman, Alexander, & Silverman, 2018) and longer reading times for screens in others (e.g., Connell, Bayliss, & Farmer, 2012; Kim & Kim, 2013). Considering that differences in reading times by medium have been proposed as key to understanding reading performance from paper and screens (Singer Trakhman et al., in press), a comprehensive examination of the findings on reading time is needed.

Metacognitive processes, such as how well readers perceive their comprehension of a text or the accuracy of their predictions of performance on an assessment of the text, are an important area of research in reading (Cross & Paris, 1988; Muijselaar et al., 2017). Accurate metacognitive processes in which readers are aware of how well they are understanding text is positively associated with better reading performance (Thiede, Griffin, Wiley, & Redford, 2009). An important metacognitive measure is the relation between a reader's confidence or prediction in performance and actual performance (i.e., calibration; Lin & Zabrucky, 1998). Readers often have inaccurate calibration in that they are overconfident in their performance (Vidal-Abarca, Mañá, & Gil, 2010). A number of researchers have examined calibration and have generally found that readers tend to be more biased (i.e., overestimate their reading performance) when reading from screens compared to paper (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014; Singer Trakhman et al., in press). However, not all studies have found that reading from screens is detrimental to calibration (Singer Trakhman et al., 2018) and this specific measure has not been examined in previous systematic reviews or meta-analyses (e.g., Kong, Seo, & Zhai, 2018; Nichols, 2016; Singer & Alexander, 2017b). Given concerns regarding screen inferiority with metacognition, this topic is critical to consider when comparing reading from screens and paper (Lauterman & Ackerman, 2014).

Characteristics of readers and texts

Genre and age are two characteristics of readers and texts that may be influential in comparing reading performance and processes between medium. A key characteristic of text is genre with the broad categorisations of narrative or expository. Narrative and expository texts are constructed differently, with narrative texts generally being easier to read than expository texts (Graesser, McNamara, & Kulikowich, 2011). The background knowledge necessary to fully comprehend a text tends to involve everyday common experiences for narratives, but more specialised and less common knowledge for expository texts (Graesser & McNamara, 2011). Given this, it is not surprising that performance on comprehension assessments is generally better for narrative than expository texts (Best, Floyd, & McNamara, 2008). Genre is particularly salient to consider due to arguments that reading from screens is better suited for light narrative reading than for serious study of expository texts (Baron, 2015).

Age is an important characteristic of the reader to be considered. Because adults, by definition, are older than children, they almost always have more experience with reading. In addition, it has become increasingly more common for elementary school students to read from screens (Picton, 2014).Therefore, it is likely that readers who are children may be more accustomed to reading from screens and readers who are adults likely learned how to read from paper. In contrast, there are reasons to expect adult readers would be less affected than children by medium. Ackerman and Lauterman (2012) reported that there is screen inferiority when reading from screens and readers have better awareness of performance (metacognition) with paper. Because metacognition improves with age (Schneider & Lockl, 2002), there may be more notable medium differences with children than adults.

Previous reviews

Dillon (1992) conducted a critical review of empirical literature on differences in speed and accuracy between reading from screens and paper. It was concluded that reading from a screen was slower and more fatiguing than reading from paper, although this issue could change with improvements in technology (Dillon, 1992). Furthermore, Dillon (1992) noted a preference for reading paper books over electronic. Despite these issues, reading comprehension performance scores were not determined to be affected by medium (Dillon, 1992).

Noyes and Garland (2008) followed up on Dillon (1992) and conducted a critical review of research on the equivalence of performance between paper-and-pencil and computer-based testing. Based on this review, computer-based reading comprehension tasks were determined to be more difficult than paper versions. This is based on differences in the level of effort, fatigue and stress reported (Noyes & Garland, 2008). Noyes and Garland (2008) concluded that computer-based and paper-based tasks were inherently too different to be equivalent but that technological advances had made the two types of tasks more similar than they were when Dillon's review was published in 1992.

Wang, Jiao, Young, Brooks, and Olson (2008) conducted a meta-analysis to compare computer-based to paper-and-pencil testing on K-12 student reading assessments. Overall, Wang et al. (2008) found that there was no reliable difference between computer-based and paper-and-pencil testing on reading achievement scores for K-12 students. However, these analyses were limited to K-12 students in testing environments.

There have been meta-analyses conducted on the effects of technological enhancements on reading performance (Cheung & Slavin, 2012; Moran, Ferdig, Pearson, Wardrop, & Blomeyer, 2008; Slavin, Cheung, Groff, & Lake, 2008). These meta-analyses have looked at the influence of educational technology intended to support reading instruction compared to traditional reading instruction (Cheung & Slavin, 2012; Moran et al., 2008) or compared a variety of reading instructional techniques, including computer-assisted instruction (Slavin et al., 2008). These meta-analyses did not specifically address the differences in reading performance or processes between the same texts from paper and screens.

Moreover, the readers in these meta-analyses were developing their reading skills as opposed to reading independently.

Nichols (2016) provided an overview of research findings on reading from screens and paper. Nichols (2016) argued that, across studies, the theme was that reading comprehension performance from screens and paper were not reliably different. However, reading text from screens was determined to be more difficult than reading text from paper. This overview qualitatively summarised selected previous findings in a helpful way; however, a systematic approach is necessary to provide a thorough understanding of the topic in a manner with less potential for bias.

In a systematic review of the literature, Singer and Alexander (2017b) critically examined trends since Dillon's (1992) review of empirical studies examining how the medium (electronic or paper) of reading related to text comprehension. Singer and Alexander's (2017b) review raised important issues in terms of what areas had been well researched and what areas needed further work on this topic. However, there was not a meta-analysis on the findings to provide a quantitative overview of the findings on performance. Furthermore, Singer and Alexander (2017b) noted that the number of publications of empirical studies comparing reading from screens and paper has increased dramatically over the last 10 years. This uptick in dissemination indicates a need for a meta-analysis examining findings since the reviews of Noyes and Garland (2008) or Wang et al. (2008).

A meta-analysis by Kong et al. (2018) found a benefit for reading comprehension when reading from paper compared to screens. There were no reliable differences in reading times by medium, which indicates that reading from paper is more efficient than reading from screens considering that there is better performance with similar time investments. They did not however, identify any moderators, although they noted considerable variability in the performance findings. Furthermore, Kong et al. (2018) did not include literal reading performance measures in their meta-analysis. When making broad comparisons of reading performance, such as between reading from paper and screens, including both literal and inferential measures is preferable as it provides a more comprehensive examination of reading (Mislevy & Sabatini, 2012). Moreover, Kong et al. (2018) did not examine calibration measures, which have been shown to be a critical issue when reading from paper compared to screens (Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014).

Objectives

In light of the multitude of issues related to reading from screens and paper, numerous experiments have been conducted pertaining to performance and processes when reading from different media. As more reading is performed from screens and more research is conducted (Singer & Alexander, 2017b), there is a need to have a comprehensive understanding of the issues involved in reading from screens compared to paper. The purpose of this systematic review and meta-analysis is to consolidate the experimental findings on reading text from screens and paper on reading performance, time and/or metacognition, specifically calibration. The studies reviewed had participants with fundamental reading skills who were reading in their native language and used the same texts for both the screen and paper reading conditions.

In this meta-analysis, the work of Kong et al. (2018) is expanded upon by examining moderators they did not examine, specifically genre and age. In addition, this meta-analysis examines literal, inferential and general measures of reading comprehension whereas Kong et al. (2018) excluded work with only literal measures. Furthermore, this meta-analysis considers metacognition in terms of calibration, which is how accurate readers are in their judgements of reading performance. This was not examined by Kong et al. (2018). Three research questions were addressed in the current review and meta-analysis:

1 How does reading text from paper compared to screens influence performance (literal, inferential and general) on reading assessments?
2 How does reading text from paper compared to screens influence the process of reading in terms of reading time and metacognition?
3 How do the performance and processes findings vary by genre (narrative or expository) and age (child or adult readers)?

Method

Eligibility criteria

A systematic search was conducted for studies that examined (a) reading performance from paper compared to screens and (b) the process of reading from paper compared to screens (reading time and metacognition). There were three sets of inclusion criteria. The first set of inclusion criteria was based on the purpose of this review and meta-analysis. In order to be included, the study needed to examine reading performance, time and/or calibration between screen and paper reading (e.g., studies of different types of digital reading without a paper comparison were excluded). The study needed to be published after 2008 and not included in the Noyes and Garland (2008) critical review or the Wang et al. (2008) review and meta-analysis.

The second set of inclusion criteria for studies was used to minimise confounds. The participants in the studies needed to have fundamental reading skills because learning to read involves different processes than reading to learn (van den Broek & Kendeou, 2017). Also, the studies needed to have participants who did not report disabilities, including visual impairments, to avoid possible confounds related to disabilities (e.g., electronic text can be enhanced to ease reading for individuals with visual impairments; Mulloy, Gevarter, Hopkins, Sutherland, & Ramdoss, 2014). Any texts used in the studies to examine reading needed to be longer than one sentence to avoid confounds related to sentence-level and discourse-level reading (Berninger, Nagy, & Beers, 2011; Carpenter, Miyake, & Just, 1995). Moreover, participants must have read in their native language to avoid possible confounds related to reading in one's second language (Melby-Lervåg & Lervåg, 2014). The language of the study materials could be any language provided it was the participants' native language, and the findings were reported in English (because of the language background of the author of this systematic review and meta-analysis, reports in English were necessary).

Finally, there was a third set of inclusion criteria to screen the methodological quality of the studies. These criteria were to ensure a basic quality standard to allow for clarity of causal inferences (Cooper, 2015; Valentine & Cooper, 2008). The conditions for screen and paper reading needed to be comparable to allow direct comparisons of the media. The experimental procedures must have been carried out in a supervised environment (e.g., classroom or laboratory) so that measures would be accurate. Also, it was necessary to have random assignment to screen and paper reading conditions in which the same texts were used in both the screen and paper conditions or used a within-subjects design with counterbalancing of texts.

Resources and collection processes

The systematic search for studies comparing reading from paper and screens involved multiple steps. First, in October and November of 2016, searches for relevant literature were conducted in the following databases: Educational Resources Information Center (119 records), Science Direct (270 records), Taylor and Francis Online (474 records), Sage (232 records), Springerlink (179 records), PsychINFO (105 records) and Wiley Online (663 records). In these searches, 'paper', 'electronic', 'print' and 'screen' were used with 'read*' (with * as a joker) as search terms for the titles. These searches led to 2,042 hits across the databases. There were 24 duplicates that were removed. Then, 1,983 records were removed after an initial screen of the title and abstract indicated that they were not relevant for the review.

After this initial screening based on abstracts and titles, 51 relevant full texts were assessed. These 51 relevant full texts were further screened and 38 were removed (see Figure 1 for reasons for removal). The corresponding authors of relevant reports were contacted and asked if they would share additional studies in this area either published or unpublished. An additional four reports were obtained through author recommendations for a total of 17 relevant reports.

Following Wohlin (2014), the 17 reports were used as a start set for 'snowballing' in which citations within any of these reports would be considered. A backward search of the citations in the 17 reports led to the identification of two additional reports for a total of 19 reports. Finally, a forward search examining work that had cited these 19 reports in Google Scholar was conducted in May 2018. Google Scholar was chosen to avoid bias of a particular publisher (Wohlin, 2014). This led to the identification of 10 additional articles for a total of 29 reports with 33 studies each with independent effect sizes. Following Moher, Liberati, Tetzlaff, Altman, and Prisma Group (2009), a flow diagram outlines this process in Figure 1.

Data items

The descriptive information of the studies in the selected reports were coded and double coded by the author, and then, an independent research assistant coded 25% of the reviewed articles (see Follmer, 2018, for similar approach; k = .93; see Data S1). The coding involved major features of each study: bibliographic information, participant ages, number of participants, measures of reading performance, if there were measures of reading time or calibration processes, statistics for meta-analyses (means and standard deviations, t tests or F statistics) and average text length. The potential moderators for meta-analyses were coded: genre of texts (narrative or expository) and age group (child or adult). Child was defined as under the age of 18 and/or high school/secondary school and younger. Adult was over the age of 18 and/or college and older. Performance measures were further coded as literal or inferential. Literal measures were those that involved factual recall of information whereas inferential measures involved making inferences by connecting ideas (Cain et al., 2004; Hosp & Suchey, 2014). If there were insufficient information to categorise measures as literal or inferential, they were coded as general. In addition, if there were both inferential and literal measures, a general effect size was calculated based on the means of the reading performance outcomes. For all data items, if the information was not reported in the study, the author was contacted with requests for information (e.g.,

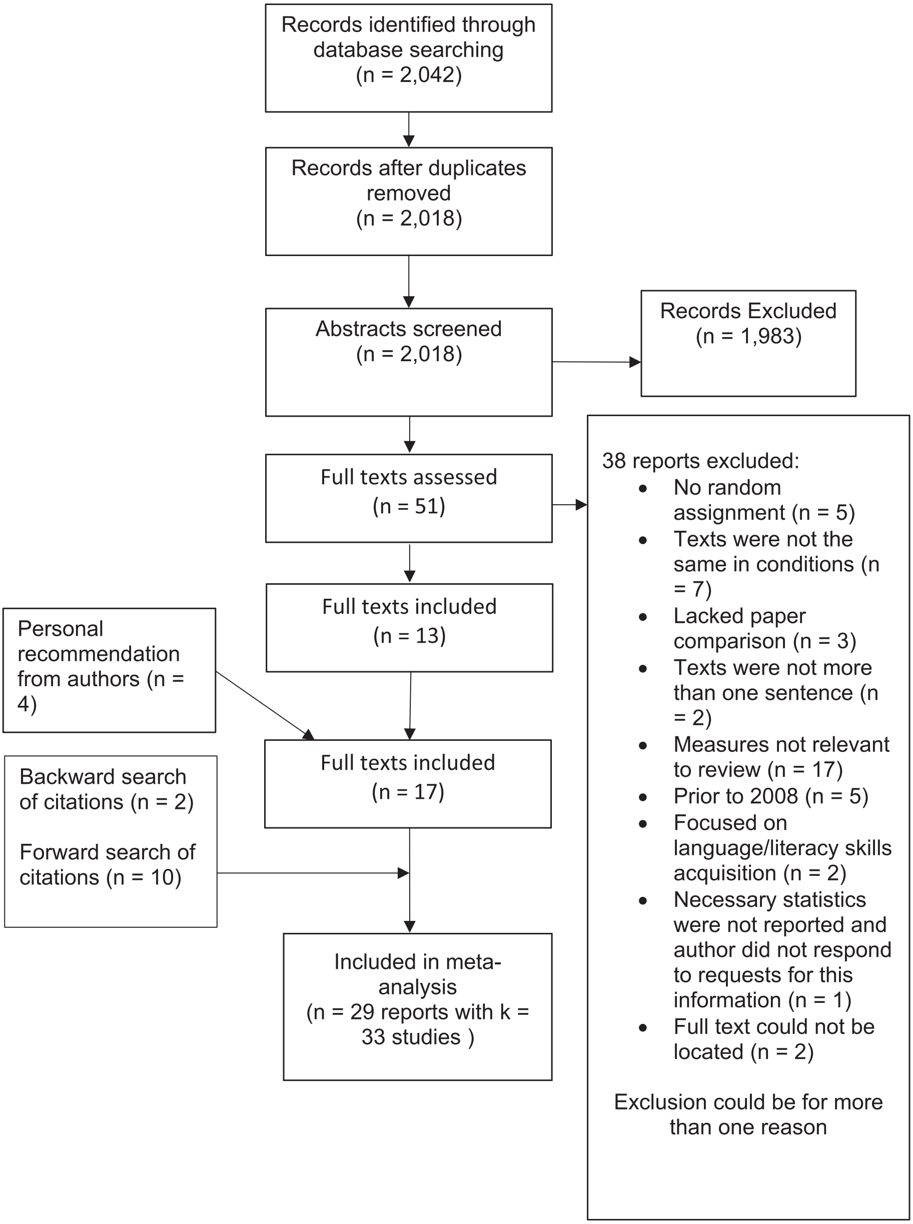
Records identified through database searching
(n = 2,042)
Records after duplicates removed
(n = 2,018)
Abstracts screened
(n = 2,018)
Records Excluded
(n = 1,983)
Full texts assessed
(n = 51)
38 reports excluded:
• No random assignment (n = 5)
• Texts were not the same in conditions (n = 7)
• Lacked paper comparison (n = 3)
• Texts were not more than one sentence (n = 2)
• Measures not relevant to review (n = 17)
• Prior to 2008 (n = 5)
• Focused on language/literacy skills acquisition (n = 2)
• Necessary statistics were not reported and author did not respond to requests for this information (n = 1)
• Full text could not be located (n = 2)
Exclusion could be for more than one reason
Full texts included
(n = 13)
Personal recommendation from authors (n = 4)
Full texts included
(n = 17)
Backward search of citations (n = 2)
Forward search of citations (n = 10)
Included in meta-analysis
(n = 29 reports with k = 33 studies)

Figure 1. Flow diagram of the systematic review process.

performance measures separated by literal and inferential items for Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014).

The descriptive information for the studies is shown in Table 1. There were 33 studies (all experiments with random assignment per the inclusion criteria) and a total of 2,799 participants. Note that some studies had conditions that were not relevant to the current research questions, such as do-nothing controls (Daniel & Woody, 2013; Green et al., 2010) and screen conditions that were not comparable to the paper conditions (Hou, Rashid, &

Table 1. Description of studies.

Author/s (year)	Participants	Texts	Experiment design (for medium)	Reading performance measure(s)	Type of reading performance measure(s)	Reading time and/or calibration measures	Limitations
Ackerman and Goldsmith (2011) Exp 1	70 college students	Six expository (1,000–1,200 words each)	Between	Multiple-choice questions	Literal and inferential	Calibration	Reliability not reported; small sample size
Ackerman and Goldsmith (2011) Exp 2	74 college students	Six expository (1,000–1,200 words each)	Between	Multiple-choice questions	Literal and inferential	Reading time and calibration	Reliability not reported; small sample size
Ackerman and Lauterman (2012) Exp 1	80 college students	Five expository (1,000–1,200 words each)	Between	Multiple-choice questions	Literal and inferential	Reading time and calibration	Reliability not reported; small sample size
Ackerman and Lauterman (2012) Exp 2	76 college students		Between	Multiple-choice questions	Literal and inferential	Reading time and calibration	Reliability not reported; small sample size
Baker (2010)	104 college students	Narrative and expository (details not available)	Between	Multiple-choice questions	General	Calibration	Items were not described; findings by genre were not available; reliability for performance measures not reported
Ben-Yehudah and Eshet-Alkalai (2018)	102 college students	One expository (858 words)	Between	Multiple-choice questions; essay question	Literal and inferential	Reading time	Low reading performance measure reliability

(*Continues*)

Table 1. (Continued)

Author/s (year)	Participants	Texts	Experiment design (for medium)	Reading performance measure(s)	Type of reading performance measure(s)	Reading time and/or calibration measures	Limitations
Chen and Catrambone (2015)	92 college students	Three expository (1,000 words each)	Between	Multiple-choice and short-answer questions	Literal and inferential	Reading time and calibration	Small sample size; Reliability for multiple-choice questions not reported
Chen, Cheng, Chang, Zheng, and Huang (2014)	90 college students	Four expository (1,050–1,099 words each)	Between	Multiple-choice questions; summaries	Literal	No	Reliability not reported; only literal measures; no process measures
Connell, Bayless, and Farmer (2012)	201 college students	One expository (length not stated)	Between	Multiple-choice questions	Literal	Reading time	Prior knowledge was measured, but not part of analyses; only literal measures
Daniel and Woody (2013)	141 college students	One expository (textbook chapter)	Between	Multiple-choice questions	General	Reading time	Reliability not reported; minimal description of performance items
Dundar and Akcayir (2017)	20 elementary school students	Narrative and expository (details not provided)	Between	Question format not stated	General	Reading time	Small sample size; Reliability not reported

(*Continues*)

Table 1. (Continued)

Author/s (year)	Participants	Texts	Experiment design (for medium)	Reading performance measure(s)	Type of reading performance measure(s)	Reading time and/or calibration measures	Limitations
Grace (2011)	18 elementary school students (Grade 3)	Two narrative book chapters	Between	Short-answer questions	Literal and inferential	No	Separate analyses for literal and inferential items were not reported
Green, Perera, Dance, and Myers (2010)	55 college students	One expository (two pages long)	Between	Multiple-choice questions	Literal	No	Reliability not reported; only literal measures
Heij and van der Meij (2014)	16 college students	Two expository (2,147 and 6,475 words)	Between	Essay questions	Literal and inferential	Reading time	Small sample size; Separate analyses for literal and inferential items were not reported
Hermena et al. (2017)	24 college students	Two narrative (each 604 words)	Within	Multiple-choice questions	General	Reading time	Small sample size
Hou, Rashid, and Lee (2017)	30 college students	One narrative (30 pages long)	Between	One open-ended question and multiple-choice questions	General	Reading time	Reliability of performance measure not reported; minimal description of performance items

(*Continues*)

Table 1. (Continued)

Author/s (year)	Participants	Texts	Experiment design (for medium)	Reading performance measure(s)	Type of reading performance measure(s)	Reading time and/or calibration measures	Limitations
Hou, Wu, and Harrell (2017)	81 adults (over the age of 50)	One narrative (3,469 words) and one expository (3,150 words)	Between	Sequence-of-events questions and multiple-choice questions	General	Reading time	Reliability of performance measure not reported; minimal description of performance items
Kim and Kim (2013)	108 high school students	Two expository (each two pages long)	Within	Multiple-choice questions	General	Reading time	Reliability of performance measure not reported; minimal description of performance items; answering performance items electronically may have been more difficult than on paper
Kretzschmar et al. (2013)	36 younger adults (mean age 25.7 years) and 21 older adults (mean age 66.8 years)	Three expository and three narrative (average length 222 words)	Within	Yes/no questions	Literal	No	Reliability of performance measure not reported

(*Continues*)

Table 1. (Continued)

Author/s (year)	Participants	Texts	Experiment design (for medium)	Reading performance measure(s)	Type of reading performance measure(s)	Reading time and/or calibration measures	Limitations
Lauterman and Ackerman (2014) Exp 1	87 college students	Six expository (1,000–1,200 words each)	Between	Multiple-choice questions	Literal and inferential	Calibration	Reliability of performance measure not reported
Lauterman and Ackerman (2014) Exp 2	76 college students	Six expository (1,000–1,200 words each)	Between	Multiple-choice questions	Literal and inferential	Calibration	Reliability of performance measure not reported
Mangen, Walgermo, and Brønnick (2013)	72 tenth-grade students	One expository and one narrative (1,400–1,600 words each)	Between	Multiple-choice and short-answer questions	General	No	Minimal description of performance items; answering performance items electronically may have been more difficult than on paper
Margolin, Driscoll, Toland, and Kegler (2013)	90 college students	Five expository (542 words each) and five narrative (average 541.8 words each)	Between	Multiple-choice questions	Inferential	No	Small sample size for paper condition; only inferential items

(*Continues*)

Table 1. (Continued)

Author/s (year)	Participants	Texts	Experiment design (for medium)	Reading performance measure(s)	Type of reading performance measure(s)	Reading time and/or calibration measures	Limitations
Neijens and Voorveld (2018)	90 college students	Expository (24 pages)	Between	Recall	Literal	No	Real newspaper articles were used so participants may have been familiar; reliability for performance measure not reported
Norman and Furnes (2016) Exp 1	100 college students	Four expository (1,000 words each)	Between	Multiple-choice questions	Literal	Calibration	Small sample size for paper condition; only literal items; reliability for performance measure not reported
Norman and Furnes (2016) Exp 2	50 college students	Four expository (1,000 words each)	Between	Multiple-choice questions		Calibration	Small sample size; only literal items; reliability for performance measure not reported

(*Continues*)

Table 1. (Continued)

Author/s (year)	Participants	Texts	Experiment design (for medium)	Reading performance measure(s)	Type of reading performance measure(s)	Reading time and/or calibration measures	Limitations
Porion, Aparicio, Megalakaki, Robert, and Baccino (2016)	72 secondary school students	One expository (one page long)	Between	Yes/no, multiple-choice and true/false questions	Literal and inferential	No	Small sample size; reliability for performance measure not reported
Singer and Alexander (2017a)	90 college students	Four expository (each approximately 450 words)	Within	Explain main idea, list key points, free recall	Literal and inferential	Calibration of medium (not performance on items)	Interrater reliability reported as percentage agreement rather than kappa
Singer Trakhman, Alexander, and Berkowitz (in press)	86 college students	Two expository (approximately 550 words each)	Within	Explain main idea, list key points, free recall	Literal and inferential	Reading time and calibration	Interrater reliability reported as percentage agreement rather than kappa
Singer Trakhman, Alexander, and Silverman (2018)	57 college students	Two expository (1,800 words each)		Explain main idea, list key points, free recall	Literal and inferential	Reading time and calibration	Interrater reliability reported as percentage agreement rather than kappa

(*Continues*)

Table 1. (Continued)

Author/s (year)	Participants	Texts	Experiment design (for medium)	Reading performance measure(s)	Type of reading performance measure(s)	Reading time and/or calibration measures	Limitations
Stevens (2014)	187 middle school students	Narrative (details not provided)	Between	Multiple-choice questions	General	No	Minimal description of performance items; reliability based on study data not reported
Taylor (2011)	74 college students	One expository (textbook chapter)	Between	Multiple-choice questions	General	No	Small sample size; reliability for performance measure not reported
Wells (2013)	140 middle and high school students	Narrative and expository (details not provided)		Multiple-choice questions	General	No	Reliability based on study data not reported

Lee, 2017; Stevens, 2014). Participants in these sorts of conditions were not included in Table 1 nor in the meta-analyses. If there was more than one screen condition (e.g., Chen et al., 2014 had tablet and computer monitor conditions), the screen conditions were combined in the meta-analyses. There were 14 studies that reported reading times (n = 1,233) and 11 studies that reported calibration for judgement of reading performance (n = 698).

Study quality

There was an examination of study quality (also referred to as risk of bias; Liberati et al., 2009) based on the Study Design and Implementation Assessment Device (Study DIAD), which is a framework for assessing construct, internal, external and statistical conclusion validity (Valentine & Cooper, 2008). This was in addition to the third set of study quality criteria presented in the section. The Study DIAD provides a framework with over 30 specific questions based on four categories of study quality: fit between concepts and operations, clarity of causal inference, generality of findings and precision of outcome estimation. Researchers are to omit Study DIAD questions that are irrelevant to their question(s). For this review and meta-analysis, the items evaluating the quality of quasi-experiments and longitudinal designs were omitted as they were irrelevant. Redundant items were also omitted. The questions selected for the purposes of the review were then used to provide guidance as to possible threats to the validity of the reviewed studies (Cooper, 2015). Based on questions from the Study DIAD, the reliability of the measures, suitability of outcome measures, reporting of statistical tests and appropriateness of sample size were examined. The specific questions are listed in Data S1. Note that some of the questions were used as inclusion criteria, and studies that did not meet them were not included in this meta-analysis (see Figure 1 for the number of studies excluded with reasons related to quality). Noted issues are in the Limitations column for the tables in which studies are described (Table 1).

Statistical procedures

The principal summary measure was Hedges' g. Hedges' g is an unbiased approach for estimating standardised mean differences because it is corrected for sample size (Enzmann, 2015; Hedges, 1981). Relevant statistics (e.g., means and standard deviations, t tests and sample sizes) from each experiment were used to calculate Hedges' g using Comprehensive Meta-Analysis software (version 3; Biostat, Englewood, NJ). If the relevant statistics were not reported, then the corresponding author was contacted with requests for the necessary statistical information. If the relevant statistics were not reported and the author did not respond to requests, then the experiment could not be included (as shown in Figure 1, this was the case for one full text report examined). A positive Hedges' g indicates that the mean values for reading from screens were greater.

Heterogeneity of effect sizes were calculated based on the I^2 index which indicates the degree of heterogeneity from 0 to 100 with higher levels indicating greater heterogeneity (i.e., the amount of variability in effect sizes among the included studies is more than would be expected from sampling error; Liberati et al., 2009). The I^2 index is the percentage of variability across studies that is not due to chance or sampling error and is instead thought to be due to heterogeneity (Higgins & Green, 2011). If the I^2 index is more than

20%, a moderator analysis to examine potential sources of variability is warranted (Bloch, 2014). The proportion of variance explained by a moderator is reported with the R^2 index.

If more than one outcome measure of reading performance and/or more than one condition per medium was reported, the means of the outcomes were used. This was because using more than one outcome from a particular study would violate assumptions of independence thereby introducing bias, as studies with more outcomes would receive more weight in the meta-analysis (Scammacca, Roberts, & Stuebing, 2014).

RQ1: reading performance results and discussion

Overall performance

The overall difference in reading performance by medium across all measures was first examined (k = 33). The I^2 was 70.35, which indicates heterogeneity sufficient to warrant investigation into moderators (Bloch, 2014; Higgins, Thompson, Deeks, & Altman, 2003). Given this heterogeneity and that the samples in the studies were from different populations (Table 1), a random effects model was used because the assumptions for a fixed effects model were not met (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, 2015; Field & Gillett, 2010). Based on the random effects model, reading text from screens had a small, but significant negative effect on performance scores compared to reading from paper, $g = -.25$, $k = 33$, SE = .06, 95% CI = [−.37, −.12], $p < .001$ (see Table 2 for statistics for each study).

To test for potential outliers, 'one study removed' analyses were conducted (Borenstein et al., 2009). This approach calculates the effect size if each of the studies in the meta-analysis were individually removed. In this way, the influence of any given study can be noted if its removal leads to a substantial change in the overall effect size (i.e., the effect size with the study removed was not within the confidence intervals; see Healy, Nacario, Braithwaite, & Hopper, in press, for a similar approach). As shown in Data S2 (Table B1), the removal of any one of the studies would not substantially change the results of the meta-analysis on reading performance.

Publication bias (that statistically significant findings were more likely to be reported) was examined with the graphical technique of a funnel plot and the statistical technique of Egger's test of the intercept (Cooper, 2015; see Follmer, 2018, for a similar approach). In funnel plots, studies are graphed according to their size along the y-axis (larger studies towards the top of the graph) and their effect size along the x-axis with the mean effect represented by a line in the middle. Publication bias is indicated by an asymmetrical distribution on the sides of the mean effect size as well as smaller studies being scattered more widely at the bottom (Egger, Smith, Schneider, & Minder, 1997). As can be noted in Figure 2, the studies were distributed in a symmetrical manner across the funnel plot. In Egger's test of the intercept, the intercept did not significantly differ from zero, $\beta = 1.29$, $p = .20$, 95% CI [−.71, 3.29]. Based on these analyses, publication bias was unlikely.

Moderator analyses. Moderators considered were the genre of the experimental texts (narrative or expository) and the age group of the participants (child or adult). Moderator analyses were assessed using $Q_{between}$ statistics (see Takacs, Swart, & Bus, 2015, for similar approach). Based on recommendations from Borenstein et al. (2009), there needed to

Table 2. Reading performance statistics for each study and model statistics (positive Hedges' *g* indicates better performance with screens compared to paper).

Study name	Age group	Genre	Hedges' *g*	Standard error	95% CI lower limit	95% CI upper limit	*p* value	Sample size E	P	T
Ackerman and Goldsmith (2011) Exp 1	Adult	Expository	.02	.24	−.44	.49	.92	35	35	70
Ackerman and Goldsmith (2011) Exp 2			−.69	.24	−1.15	−.22	.004	37	37	74
Ackerman and Lauterman (2012) Exp 1			−.29	.22	−.73	.15	.20	40	40	80
Ackerman and Lauterman (2012) Exp 2			−.30	.23	−.74	.15	.20	38	38	76
Baker (2010)		Combined	−.20	.21	−.60	.21	.34	69	35	104
Ben-Yehudah and Eshet-Alkalai (2018)		Expository	−.48	.20	−.87	−.09	.02	52	50	102
Chen and Catrambone (2015)			−.07	.21	−.47	.34	.75	46	46	92
Chen et al. (2014)			−.58	.23	−1.02	−.13	.01	60	30	90
Connell et al. (2012)			.01	.15	−.29	.32	.93	138	60	198
Daniel and Woody (2013)			.08	.14	−.19	.35	.56	120	89	209
Dundar and Akcaryir (2012)	Child	Combined	.29	.43	−.56	1.13	.51	10	10	20
Grace (2011)		Narrative	−.09	.44	−.95	.77	.84	9	10	19
Green et al. (2010)	Adult	Expository	−.31	.22	−.74	.12	.16	41	41	82
Heij and van der Meij (2014)			−.31	.44	−1.17	.55	.48	8	8	16
Hermena et al. (2017)		Narrative	.00	.20	−.39	.39	.99			24
Hou, Rashid, and Lee (2017)			−.16	.36	−.86	.54	.66	15	15	30
Hou, Wu, and Harrell (2017)		Combined	−.13	.22	−.56	.31	.57	41	40	81
Kim and Kim (2013)	Child	Expository	−1.18	.13	−1.42	−.94	<.001			108
Kretzschmar et al. (2013)	Adult	Combined	−.18	.13	−.44	.08	.18			57
		Narrative	−.14	.13	−.40	.12	.28			

(*Continues*)

Table 2. (Continued)

Study name	Age group	Genre	Hedges' *g*	Standard error	95% CI lower limit	95% CI upper limit	*p* value	Sample size		
								E	P	T
		Expository	−.21	.13	−.47	.05	.11			
Lauterman and Ackerman (2014) Exp 1	Adult	Expository	−.12	.22	−.54	.31	.59	38	49	87
Lauterman and Ackerman (2014) Exp 2			−.17	.23	−.62	.27	.45	37	39	76
Mangen et al. (2013)	Child	Combined	−.40	.25	−.88	−.09	.11	47	25	72
		Narrative	−.34	.13	−.82	.14	.17			
		Expository	−.45	.25	−.94	.03	.07			
Margolin et al. (2013)	Adult	Combined	−.09	.18	−.45	.26	.61	60	30	90
		Narrative	.03	.22	−.41	.46	.90			
		Expository	−.22	.22	−.65	.22	.33			
Neijens and Voorveld (2018)	Adult	Expository	−.33	.21	−.74	.08	.11	45	45	90
Norman and Furnes (2016) Exp 1			−.19	.23	−.64	.27	.42	75	25	100
Norman and Furnes (2016) Exp 2			−.53	.29	−1.10	.03	.07	25	25	50
Porion, Aparicio, Megalakaki, Robert, and Baccino (2016)	Child		.06	.23	−.40	.52	.79	36	36	72
Singer and Alexander (2017a)	Adult		−.44	.11	−.66	−.22	<.001			90
Stevens (2014)	Child	Narrative	.19	.15	−.10	.47	.21	94	93	187
Taylor (2011)	Adult	Expository	.01	.23	−.46	.47	.98	36	36	72
Trakhman et al. (in press-a)			−.34	.11	−.56	−.12	.002			86
Trakhman et al. (in press-b)			−.77	.16	−1.08	−.47	<.001			57
Wells (2013)	Child	Combined	.02	.17	−.31	.35	.92	70	68	138

(*Continues*)

Table 2. (Continued)

Study name	Age group	Genre	Hedges' *g*	Standard error	95% CI lower limit	95% CI upper limit	*p* value	Sample size		
								E	P	T
Random model (all, *k* = 33)	Combined		−.26	.04	−.33	−.19	<.001			
Random model (child, *k* = 7)	Child		−.18	.26	−.70	.33	.49			
Random model (adult, *k* = 26)	Adult		−.21	.06	−.33	−.09	<.001			
Random model (expository, *k* = 22)	Combined	Expository	−.32	.08	−.48	−.16	<.001			
Random model (narrative, *k* = 7)		Narrative	−.04	.07	−.18	.11	.64			

Notes: Sample size E = electronic text condition(s), P = paper text condition(s) and T = Total sum of participants in electronic and paper text conditions. For within subject experiments, only the total provided. The performance statistics available for Baker (2010), Dundar and Akarcir (2017), Hou, Wu, and Harrell (2017) and Wells (2013) were a composite of expository and narrative texts, so these studies were not included in genre analyses.

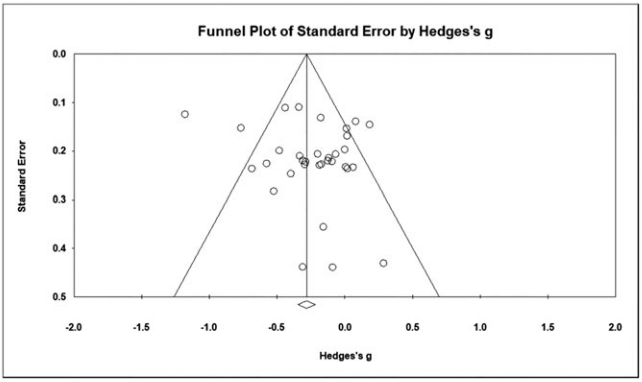

Figure 2. Funnel plot for reading performance findings.

be at least six effect sizes in a potential moderator category in order to conduct analyses (see Elleman, 2017, for a similar approach).

Genre. In studies in which both expository and narrative texts were used, only the narrative texts were considered in the moderator analyses. This is following guidelines from Higgins and Green (2011) and allows for a more even distribution given that there were more studies using expository texts than narrative texts. If a study used both narrative and expository texts but did not separate findings by genre, the findings by genre were requested from the author. Four studies were excluded from the genre moderator analyses because findings by genre were not available (Baker, 2010; Dündar & Akçayır, 2017; Hou, Wu, & Harrell, 2017; Wells, 2013). The genre of the experimental texts was significant as a moderator, $Q_{between}(1) = 6.86$, $p = .01$, $R^2 = .13$. When tested separately, reading performance for expository texts was worse for screens compared to paper, $g = -.32$, $k = 22$, SE = .08, 95% CI = [−.48, −.16], $p < .001$. In contrast, there was no difference in reading performance by medium for narrative texts, $g = -.04$, $k = 7$, SE = .07, 95% CI = [−.18, .11], $p = .64$. However, this finding should be interpreted with caution given that the genres were unevenly distributed (there were more expository findings than narrative).

Based on the overall meta-analysis findings, there is a benefit for performance when reading from paper compared to screens. Based on moderator analyses, this benefit appears to be primarily for expository texts with no reliable differences in reading performance by medium noted for narrative texts. This finding is consistent with views that reading from screens is more appropriate for light pleasure reading, which is more likely to be from narratives, than for challenging reading, which is more likely to be from expository texts (Baron, 2015; Narvaez, van den Broek, & Ruiz, 1999; Nell, 1988). It should be noted that there were considerably more experiments with expository texts than narrative, which diminishes the robustness of this moderator analysis. However, in three studies in which both narrative and expository texts were used (and findings by genre were available), there appeared to be more of a negative effect of screens on performance for expository compared to narrative texts (Kretzschmar et al., 2013; Mangen et al., 2013; Margolin et al., 2013).

Age. The age group (child or adult) of the participants was examined, but it was not a significant moderator, $Q_{between}(1) = .07$, $p = .80$, $R^2 = .04$, indicating that the negative effect on performance for reading text from screens rather than paper did not vary for readers who were adults or children. However, this finding should be interpreted with caution because there were more studies with adult participants ($k = 26$) than child ($k = 7$).

The reading performance findings did not vary by the age of the reader with both children and adults having similar benefits in reading performance from paper compared to screens. One reason for this could be that the adult participants in the reviewed studies were almost entirely college students often considered to be 'digital natives' who are comfortable with technology (Akçayır, Dündar, & Akçayır, 2016) so both age groups (i.e., children and adults) were similarly accustomed to reading from screens. One study, by Kretzschmar et al. (2013) compared a group of older adults (average age of 66.8 years) with younger adults (average age of 25.7 years) and found similar patterns by medium for each age group. In other words, the view that older adults have more difficulty than younger adults when reading from screens was not supported in their analyses (Kretzschmar et al., 2013).

Literal and inferential reading. To examine if the effect of medium on reading performance fluctuates depending on the type of performance assessment, separate analyses were conducted for literal and inferential reading performance (see Elleman, 2017, for a similar approach). For literal reading performance, the I^2 was 67.13, and a random effects model was used. As can be seen in Table 3, literal reading performance from screens was worse than from paper, $g = -.33$, $k = 19$, SE = .08, 95% CI = [−.48, −.18], $p < .001$. Moderator analyses were not conducted given that none of the proposed moderators had a minimum of six effect sizes. For inferential reading performance, the I^2 was 0.00. Similar to literal reading performance, inferential reading performance from screens was worse than from paper, $g = -.26$, $k = 13$, SE = .05, 95% CI = [−.36, −.17], $p < .001$ (Table 4). Based on the one-study removed analyses, there were no outliers in either the literal (Table B2) or inferential (Table B3) analyses.

The benefit of paper for reading performance was noted for both literal and inferential measures. Given that literal reading tasks are typically considered easier than inferential reading tasks (Basaraba et al., 2013), this finding was contrary to expectations that screens would be more detrimental for challenging tasks than easier tasks. In three studies with similar materials and populations (college students), there were no differences by medium for determining the main idea of expository texts, but readers were better able to recall details from texts read from paper than screens (Singer & Alexander, 2017a; Singer Trakhman et al., in-press; Singer Trakhman et al., 2018). Determining the main idea of a text requires making connections (inferences) throughout the text in order to get an overall understanding, whereas recall does not require making connections (Leopold & Leutner, 2012; Schiefele & Krapp, 1996). Singer Trakhman et al. (in press), Singer and Alexander (2017a) and Singer Trakhman et al. (2018) concluded that readers may be able to make connections to understand the overall main idea similarly with different media, but reading from screens may interfere with encoding specific details relative to paper. Literal measures are based on memory of the text, and subsequently, it is logical that encoding difficulty would affect performance on literal measures. Difficulty encoding details from text read from screens compared to paper could cause issues with

Table 3. Literal reading performance statistics for each study and model statistics (positive Hedges' *g* indicates better performance with screens compared to paper).

Study name	Hedges' *g*	Standard error	95% CI lower limit	95% CI upper limit	*p* value	Sample size		
						E	P	T
Ackerman and Goldsmith (2011) Exp 1	.05	.24	−.41	.52	.82	35	35	70
Ackerman and Goldsmith (2011) Exp 2	−.77	.24	−.124	−.30	.001	37	37	74
Ackerman and Lauterman (2012) Exp 1	−.24	.22	−.68	.19	.27	40	40	80
Ackerman and Lauterman (2012) Exp 2	−.30	.23	−.74	.15	.20	38	38	76
Ben-Yehudah and Eshet-Alkalai (2018)	−.45	.20	−.84	−.06	.02	52	50	102
Chen and Catrambone (2015)	−.07	.21	−.47	.34	.75	46	46	92
Chen et al. (2014)	−.58	.23	−1.02	−.13	.01	60	30	90
Connell et al. (2012)	.01	.15	−.29	.32	.93	138	60	198
Green et al. (2010)	−.31	.22	−.74	.12	.16	41	41	82
Kretzschmar et al. (2013)	−.18	.13	−.44	.08	.18			57
Lauterman and Ackerman (2014) Exp 1	−.20	.22	−.62	.22	.34	38	49	87
Lauterman and Ackerman (2014) Exp 2	−.21	.23	−.66	.24	.35	37	39	76
Neijens and Voorveld (2016)	−.33	.21	−.74	.08	.11	45	45	90
Norman and Furnes (2016) Exp 1	−.19	.23	−.64	.27	.42	75	25	100
Norman and Furnes (2016) Exp 2	−.53	.29	−1.10	.03	.07	25	25	50
Porion et al. (2016)	.23	.24	−.23	.69	.33	36	36	72
Singer and Alexander (2017)	−.67	.12	−.93	−.47	<.001			90
Trakhman et al. (in press-a)	−.32	.11	−.53	−.10	.004			86
Trakhman, Alexander, and Silverman (in press)	−1.12	.17	−1.45	−.79	<.001			57
Random model (all, *k* = 19)	−.33	.08	−.48	−.18	<.001			

Note: Sample size E = electronic text condition(s), P = paper text condition(s) and T = Total sum of participants in electronic and paper text conditions.

inferential items if answering the inferential measures required memory of specific details in the text.

RQ2: reading processes results and discussion

The meta-analytic procedures used to examine reading performance were used for reading times and calibration of performance (metacognition).

Table 4. Inferential reading performance statistics for each study and model statistics (positive Hedges' *g* indicates better performance with screens compared to paper).

Study name	Hedges' *g*	Standard error	95% CI lower limit	95% CI upper limit	*p* value	Sample size E	P	T
Ackerman and Goldsmith (2011) Exp 1	.01	.24	−.48	.45	.95	35	35	70
Ackerman and Goldsmith (2011) Exp 2	−.60	.24	−1.06	−.14	.01	37	37	74
Ackerman and Lauterman (2012) Exp 1	−.33	.22	−.77	.10	.14	40	40	80
Ackerman and Lauterman (2012) Exp 2	−.30	.23	−.74	.16	.20	38	38	76
Ben-Yehudah and Eshet-Alkalai (2018)	−.51	.20	−.90	−.12	.01	52	50	102
Chen and Catrambone (2015)	−.07	.21	−.47	.34	.74	46	46	92
Lauterman and Ackerman (2014) Exp 1	−.03	.21	−.45	.39	.89	38	49	87
Lauterman and Ackerman (2014) Exp 2	−.13	.23	−.58	.31	.56	37	39	76
Porion et al. (2016)	−.10	.23	−.56	.36	.66	36	36	72
Singer and Alexander (2017)	−.18	.11	−.39	.03	.09			90
Singer Trakhman et al. (in press-a)	−.36	.11	−.58	−.14	.001			86
Singer Trakhman et al. (in press-b)	−.41	.14	−.68	−.14	.003			57
Random model (all, *k* = 13)	−.26	.05	−.36	−.17	<.001			

Note: Sample size E = electronic text condition(s), P = paper text condition(s) and T = Total sum of participants in electronic and paper text conditions.

Reading time

For reading times, the heterogeneity of effect sizes was substantial, $I^2 = 92.47$, indicating a great deal of variability in the findings. Based on the random effects model, reading text from screens had no reliable effect on reading times compared to reading from paper, $g = .08$, $k = 14$, SE = .20, 95% CI = [−.32, .48], $p = .45$ (see Table 5 for statistics for each study).

One study removed analyses were conducted to test for outliers. As can be seen in Data S2 (Table B4), removal of any of the studies would not have changed the overall effect size outside of the confidence intervals nor change the (lack of) significance of the results.

The distribution on the funnel plot was not symmetrical (Figure 3). However, there were no differences in how smaller and larger studies were scattered. Differences in how the studies were scattered by size would indicate a bias in reporting statistically significant findings (Egger et al., 1997). In addition, the Egger's test of the intercept did not

Table 5. Reading time statistics for each study and model statistics (positive Hedges' *g* indicates longer reading times text with screens compared to paper).

						Sample size		
Study name	Hedges' *g*	Standard error	95% CI lower limit	95% CI upper limit	*p* value	E	P	T
Ackerman and Goldsmith (2011) Exp 2	−.58	.32	−1.21	.06	.08	37	37	74
Ackerman and Lauterman (2012) Exp 1	−.06	.31	−.67	.55	.85	40	40	80
Ackerman and Lauterman (2012) Exp 2	.41	.23	−.04	.86	.08	38	38	76
Ben-Yehudah and Eshet-Alkalai (2018)	.09	.28	−.45	.63	.75	52	50	102
Chen and Catrambone (2015)	−.54	.21	−.95	−.13	.01	46	46	92
Connell et al. (2012)	.42	.18	.07	.78	.02	138	60	198
Daniel and Woody (2013)	.43	.18	.09	.78	.01	120	89	209
Heij and van der Meij (2014)	−.27	.48	−1.21	.67	.58	8	8	16
Hermena et al. (2017)	−.28	.20	−.67	.12	.17			24
Hou, Rashid, and Lee (2017)	.16	.36	−.54	.85	.66	15	15	30
Hou, Wu, and Harrell (2017)	.25	.31	−.36	.86	.42	41	40	81
Kim and Kim (2013)	1.72	.15	1.43	2.02	<.001			108
Singer Trakham et al. (in press)	−.30	.11	−.51	−.08	.01			86
Singer Trakman et al. (2018)	−.49	.14	−.76	−.22	<.001			57
Random model	.08	.20	−.31	.47	.69			

Note: Sample size E = electronic text condition(s), P = paper text condition(s) and T = Total sum of participants in electronic and paper text conditions.

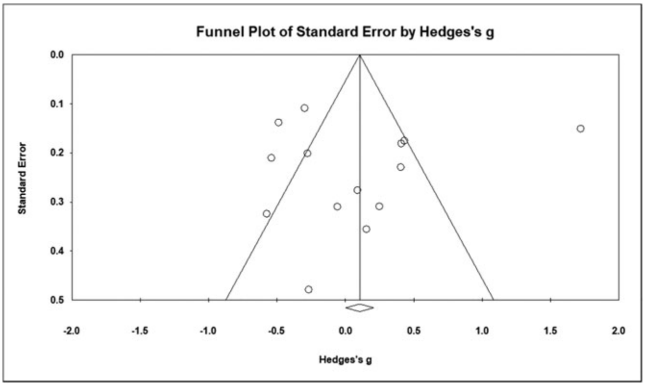

Figure 3. Funnel plot for reading time findings.

significantly differ from zero, β = −.14, p = .96, 95% CI [−6.01, 5.73]. Therefore, the lack of symmetry in the funnel plot was more likely due to heterogeneity than publication bias (Sterne et al., 2011).

Because of the small number of studies and the lack of a significant effect, a moderator analysis would not be appropriate (Field & Gillett, 2010). However, the substantial variability in findings prompted the question as to why results would be so different across studies. One notable outlier is Kim and Kim's (2013) experiment in which high school students answered multiple-choice questions about texts as they were reading. In the screen condition, students were instructed to circle their responses using their computer mice, whereas students in the paper condition circled their responses using a pencil. Despite the conclusion of Wang et al. (2008) that paper and computer-based testing are similar, it is possible that the different psychomotor demands of assessment in Kim and Kim (2013) were responsible for the substantially longer reading times in the paper than the screen condition.

The reading time findings varied across three experiments using the same materials and measures (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012). One reason for these different findings could be study population – the authors noted that the participants in the second experiment in Ackerman and Goldsmith (2011) had a much stronger preference for paper than those in Ackerman and Lauterman (2012). It is possible that one would opt to spend more time reading from a preferred medium than a non-preferred medium, which could change the direction of the findings.

Overall, the reading time findings that were significant at the study level varied from longer reading times for paper in some studies (Chen & Catrambone, 2015; Singer Trakhman et al., in press; Singer Trakhman et al., 2018) and longer reading times for screens in others (Connell et al., 2012; Daniel & Woody, 2013). These studies each involved college students and expository texts; therefore, age and genre are not potential explanations. Furthermore, text length is unlikely a factor given that Daniel and Woody (2013) and Singer Trakhman et al. (2018) had texts of similar length, and their findings conflicted with each other. However, one area in which they differ is in the presence of visual representations. In studies in which there was only text, reading times were longer from paper (Chen & Catrambone, 2015; Singer Trakhman et al., 2018; Singer Trakhman et al., in press). In contrast, the studies in which there were visual representations (e.g., graphs and illustrations), reading times were longer from screens (Connell et al., 2012; Daniel & Woody, 2013). The process of reading text with visual representation is different than that of text alone because text with visual representations requires splitting attention between the verbal and visual information as well as integrating the two modalities (Hillesund, 2010; Mason, Pluchino, Tornatora, & Ariasi, 2013; Mayer, 2009). This process could function differently reading from paper or screens if the layout of the page varies by medium. There are not clear empirical findings to support this possible explanation; therefore, this would be a potential direction for future research.

Calibration

There were several empirical studies of calibration, in which the accuracy of predictions of performance relative to actual performance, were calculated. Generally speaking, readers are overconfident in the accuracy of predictions, and calibration is calculated by subtracting the actual performance from the predicted performance (Ackerman & Goldsmith, 2011). The heterogeneity of effect sizes was low, indicating consistent findings, $I^2 = 19.65$. However, the studies involved different methodologies; therefore, a random effects model was used (Borenstein et al., 2009). Based on the random effects model, reading text from screens caused less calibrated and more overconfident predictions of performance than reading from paper, g = .20, k = 11, SE = .07, 95% CI = [.07, .33], p = .002 (see Table 6 for statistics for each study).

One study removed analyses were conducted to test for outliers. As can be seen in Data S2 (Table B5), removal of any of the studies would not have changed the overall effect size outside of the confidence intervals nor change the significance of the results.

As can be seen in the funnel plot, the distribution of studies is fairly symmetrical (Figure 4). In Egger's test of the intercept, the intercept did not significantly differ from zero, β = 1.61, p = .16, 95% CI [−.75, 3.97], which does not indicate bias. Based on these analyses, there was little evidence of publication bias.

In these analyses, the calibration bias effect sizes for each of the studies indicates better calibration for paper compared to screens (i.e., positive effect sizes) except for Singer Trakhman et al. (2018). Despite not being identified as an outlier in the one study removed analyses, the difference in the direction of the effect size raises the question as to what the cause of the difference could be. One possibility could be that participants in Singer Trakhman et al. (2018) experiment were instructed to track their reading with a pencil (paper condition) or an enlarged cursor (screen condition). Previous research findings have indicated that readers are more likely to keep track of their reading through using their fingers

Table 6. Reading calibration statistics for each study and model statistics (positive Hedges' g indicates better calibration with screens compared to paper).

Study name	Hedges' g	Standard error	95% CI lower limit	95% CI upper limit	p value	Sample size E	P	T
Ackerman and Goldsmith (2011) Exp 1	.38	.24	−.09	.85	.12	35	35	35
Ackerman and Goldsmith (2011) Exp 2	.61	.24	.14	1.07	.01	37	37	37
Ackerman and Lauterman (2012) Exp 1	.45	.23	.00	.89	.05	40	40	40
Ackerman and Lauterman (2012) Exp 2	.27	.24	−.20	.74	.26	38	38	38
Chen and Catrambone (2015)	.06	.21	−.35	.46	.79	46	46	92
Lauterman and Ackerman (2014) Exp 1	.22	.22	−.21	.65	.32	38	49	87
Lauterman and Ackerman (2014) Exp 2	.14	.23	−.32	.59	.56	37	39	76
Norman and Furnes (2016) Exp 1	.16	.15	−.16	.49	.32	75	25	100
Norman and Furnes (2016) Exp 2	.33	.29	−.23	.89	.25	25	25	50
Trakhman et al. (in press-a)	.24	.11	.03	.45	.03			86
Trakhman et al. (in press-b)	−.15	.133	−.41	.11	.27			57
Random model (all, k = 11)	.20	.07	.07	.33	.002			

Note: Sample size E = electronic text condition(s), P = paper text condition(s) and T = Total sum of participants in electronic and paper text conditions.

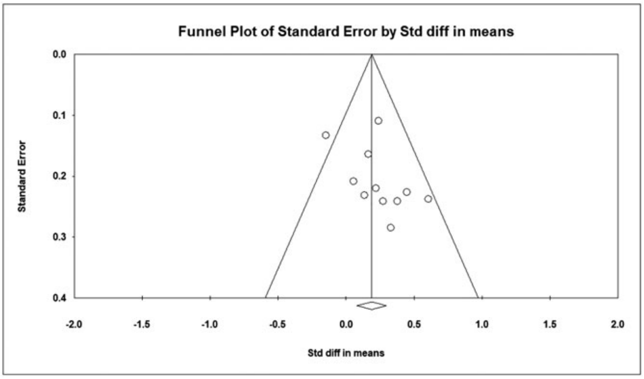

Figure 4. Funnel plot for calibration findings.

or a pencil when reading from paper rather than screens (Zaphiris & Kurniawan, 2001). Being encouraged to track while reading from screens could have possibly helped with focus that would have assisted with comprehension monitoring. However, this possibility is only conjecture without empirical findings to support it.

Although there were not enough studies or heterogeneity to warrant a moderator analysis, it is possible that reader preferences were a key factor in calibration bias. Ackerman and colleagues found that calibration was equally or more accurate in a series of studies and conditions when reading from paper compared to screens (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014). However, calibration accuracy appeared to be improved when reading from one's preferred medium (Lauterman & Ackerman, 2014). This could be interpreted that calibration is not as dependent on medium per se, but from which medium one would prefer to read.

Discussion

The purpose of this systematic literature review and meta-analyses was to examine differences in performance and processes between reading from screens and paper. Reading from paper appeared to yield better performance on assessments than reading from screens. There were no reliable differences in reading times by medium, indicating that readers performed slightly better with paper, even though similar amounts of processing and effort appeared to be involved with reading from paper and screens. In other words, reading from paper appears to be more efficient in terms of performance outcomes than reading from screens. Given the meta-analytic findings that calibration accuracy (metacognitive awareness of performance) is better when reading text from paper compared to screens, the difference in performance could be due to better calibration accuracy. Readers may be processing text from screens less efficiently based on poor calibration accuracy, as they think they are understanding the text better than they actually are (Ackerman & Goldsmith, 2011; Sidi, Ophir, & Ackerman, 2016), which could lead to detriments in performance when reading from screens (Sidi, Shpigelman, Zalmanov, & Ackerman, 2017).

One possible reason for the issues with calibration and subsequent performance with screens could be mind wandering (i.e., engaging in task unrelated thoughts; Randall, Oswald, & Beier, 2014). Readers report that it is more difficult to focus when reading from screens compared to paper (Mizrachi, 2015) and that reading from a screen can be distracting (Muir & Hawes, 2013). Given these findings, it is likely that readers may be more likely to think of topics unrelated to the task of reading when reading text from screens compared to paper. Mind wandering while reading has been found to be negatively associated with reading performance (Unsworth & McMillan, 2013). Moreover, less focus on the text would presumably lead to poorer judgement of performance, as readers would be less aware of how well they are reading. Indeed, an intervention to reduce mind wandering during lectures also improved calibration (Szpunar, Jing, & Schacter, 2014). An interesting avenue for future work would be to examine mind wandering differences by medium while reading and how mind wandering relates to differences in reading performance and calibration for paper and screens.

The contextual cue provided by the medium on which information is presented may explain the differences in performance and calibration accuracy when reading text from paper and screens. In a comparison of solving word problems displayed on paper or screens, framing a task as challenging lead to similar calibration for task performance by medium (Sidi et al., 2017). In other words, the participants may have perceived the medium as an indication of the difficulty of the problems, with the paper medium being indicative of more challenge than the electronic medium, unless they were prompted to consider the problems as challenging in both media. Although not explicitly tested with text comprehension, these findings may carry over to the findings presented in this meta-analysis. Readers are more likely to use an electronic medium for leisure reading whereas reading for study tends to be performed from paper (Aharony & Bar-Ilan, 2018; Foasberg, 2011). In this way, an electronic medium may be a contextual cue to process the text as if for leisure and paper may be a contextual cue to process the text as if for study (Sidi et al., 2016). Given previous findings that processes and performance differ depending on whether one is reading for leisure or study (van den Broek, Lorch, Linderholm, & Gustafson, 2001), it is possible contextual cue are provided by the medium, which prime the reader to read for leisure from screens or study from paper.

One area in need of better understanding is predictors of medium preference. Logically, one's previous experience with reading from screens could lead to more comfort with the medium and a preference for screens over paper, but this has not necessarily been shown in previous findings (Kurata, Ishita, Miyata, & Minami, 2016; Woody, Daniel, & Baker, 2010). Given the interactions with medium preference and reading processes, as well as performance differences (Ackerman & Lauterman, 2012; Kim & Kim, 2013; Lauterman & Ackerman, 2014), there needs to be a better understanding of what drives medium preference.

Limitations

There were limitations of this review that should be noted. The search itself was limited to dissemination in English; therefore, findings reported in other languages were not incorporated. The scope of the review was restricted to native language reading and did not include readers with learning disabilities or visual impairments. Thus, the findings reported in this review may not apply to these populations. Future studies and reviews could build on this work by examining these populations. In addition, the scope of the review did not examine medium differences in learning to read, so the findings from this review do not apply to all readers (see Takacs et al., 2015, for a meta-analysis on technology and learning to read).

There were limitations in the quality of the studies covered in this review. Across many of the studies, a common limitation is a lack of reliability statistics reported for measures (e.g., Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Chen et al., 2014; Daniel & Woody, 2013; Green et al., 2010; Kim & Kim, 2013; Lauterman & Ackerman, 2014; Porion et al., 2016; Taylor, 2011). Reading comprehension is a multidimensional construct, which makes developing measures with good internal consistency challenging (Clinton, 2014; Kamalski, 2004). However, as also suggested in Singer and Alexander (2017b), details about reliability should be included in future work.

Future studies comparing reading from screens and paper would likely need larger sample sizes than were used by many of the studies of this review. The reading performance meta-analytic results indicated that the benefit of paper over screens is rather small; therefore, the null results found in several studies could likely be due to a sample size too small to detect an effect rather than no differences between media (e.g., Chen et al., 2014; Green et al., 2010; Margolin et al., 2013; Taylor, 2011).

Conclusion

Reading from screens, such as tablets, smartphones and computers, has become ubiquitous for leisure, academic and work-related reading. This review examined the literature on performance and two processes in reading text from screens and paper. There is legitimate concern that reading on paper may be better in terms of performance and efficiency. Future examination of key issues related to mind wandering, medium preference and contextual cues provided by medium will inform the practical implications of reading text from paper compared to screens.

Acknowledgements

S. Jill Burholder, R. Alex Karie, and Michael Burd are thanked for this assistance with the project.

References

*References included in the systematic review and meta-analysis

*Ackerman, R. & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied*, 17(1), 18–32 https://doi.org/10.1037/a0022086.

*Ackerman, R. & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior*, 28(5), 1816–1828 https://doi.org/10.1016/j.chb.2012.04.023.

Aharony, N. & Bar-Ilan, J. (2018). Students' academic reading preferences: An exploratory study. *Journal of Librarianship and Information Science*, 50(1), 3–13 https://doi.org/10.1177/0961000616656044.

Akçayır, M., Dündar, H. & Akçayır, G. (2016). What makes you a digital native? Is it enough to be born after 1980? *Computers in Human Behavior*, 60, 435–440 https://doi.org/10.1016/j.chb.2016.02.089.

Albrecht, J. & O'Brien, E. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(5), 1061–1070 https://doi.org/10.1037//0278-7393.19.5.1061.

*Baker, R. (2010). Comparing the readability of text displays on paper, e-book readers, and small screen devices. (Unpublished doctoral dissertation). University of North Texas, Denton Texas.

Bando, R., Gallego, F., Gertler, P., & Romero, D. (2016). Books or laptops? The cost-effectiveness of shifting from printed to digital delivery of educational content, No 22928, NBER Working Papers, National Bureau of Economic Research, Inc. Retrieved from http://EconPapers.repec.org/RePEc:nbr:nberwo:2292

Baron, N. (2015, January 12). The case against e-readers: Why reading paper books is better for your mind. *Washington Post.*, Retrieved from https://www.washingtonpost.com/posteverything/wp/2015/01/12/the-case-against-kindle-why-reading-paper-books-is-better-for-your-mind-and-body/?utm_term=.69f23f984bab.

Basaraba, D., Yovanoff, P., Alonzo, J. & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing*, 26(3), 349–379 https://doi.org/10.1007/s11145-012-9372-9.

*Ben-Yehudah, G. & Eshet-Alkalai, Y. (2018). The contribution of text-highlighting to comprehension: A comparison of print and digital reading. *Journal of Educational Multimedia and Hypermedia.*, Retrieved from https://www.learntechlib.org/primary/p/174353/.

Berninger, V., Nagy, W. & Beers, S. (2011). Child writers' construction and reconstruction of single sentences and construction of multi-sentence texts: Contributions of syntax and transcription to translation. *Reading and Writing*, 24(2), 151–182 https://doi.org/10.1007/s11145-010-9262-y.

Best, R., Floyd, R. & McNamara, D. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, 29(2), 137–164 https://doi.org/10.1080/02702710801963951.

Bloch, M. (2014). Meta-analysis and moderator analysis: Can the field develop further? *Journal of the American Academy of Child & Adolescent Psychiatry*, 53(2), 135–137 https://doi.org/10.1016/j.jaac.2013.12.001.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. & Rothstein, H.R. (2009). *Introduction to meta-analysis*, Vol 53, (pp. 135–137). New York, NY. https://doi.org/10.1016/j.jaac.2013.12.001: Wiley.

Bowyer-Crane, C. & Snowling, M.J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *British Journal of Educational Psychology*, 75(2), 189–201 https://doi.org/10.1348/000709904X22674.

Cain, K., Oakhill, J. & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, 96(1), 31–42 https://doi.org/10.1037/0022-0663.96.1.31.

Carpenter, P., Miyake, A. & Just, M. (1995). Language comprehension: Sentence and discourse processing. *Annual Review of Psychology*, 46(1), 91–120 https://doi.org/10.1146/annurev.ps.46.020195.000515.

*Chen, D. & Catrambone, R. (2015). Paper vs. screen effects on reading comprehension, metacognition, and reader behavior. *Proceedings of The Human Factors and Ergonomics Society Annual Meeting*, 59(1), 332–336 https://doi.org/10.1177/1541931215591069.

*Chen, G., Cheng, W., Chang, T., Zheng, X. & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education*, 1(2–3), 213–225 https://doi.org/10.1007/s40692-014-0012-z.

Cheung, A.C. & Slavin, R.E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review*, 7(3), 198–215 https://doi.org/10.1016/j.edurev.2012.05.002.

Clinton, V. (2014). The relationship between students' preferred approaches to learning and behaviors during learning: An examination of the process stage of the 3P model. *Instructional Science*, 42(5), 817–837 https://doi.org/10.1007/s11251-013-9308-z.

Clinton, V. & van den Broek, P. (2012). Interest, inferences, and learning from texts. *Learning and Individual Differences*, 22(6), 650–663 https://doi.org/10.1016/j.lindif.2012.07.004.

*Connell, C., Bayliss, L. & Farmer, W. (2012). Effects of e-book readers and tablet computers on reading comprehension. *International Journal of Instructional Media*, 39(2), 131–140.

Cooper, H.M. (2015). *Research synthesis and meta-analysis: A step-by-step approach*. (5th edn). Thousand Oaks, CA: Sage Publications.

Cross, D. & Paris, S. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology*, 80(2), 131–142 https://doi.org/10.1037//0022-0663.80.2.131.

*Daniel, D. & Woody, W. (2013). E-textbooks at what cost? Performance and use of electronic v. print texts. *Computers & Education*, 62, 18–23 https://doi.org/10.1016/j.compedu.2012.10.016.

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics*, 35(10), 1297–1326 https://doi.org/10.1080/00140139208967394.

*Dündar, H. & Akçayır, M. (2017). Tablet vs. paper: The effect on learners' reading performance. *International Electronic Journal of Elementary Education*, 4(3), 441–450 Retrieved from http://iejee.com/index.php/IEJEE/article/download/188/184.

Egger, M., Smith, G., Schneider, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634 https://doi.org/10.1136/bmj.315.7109.629.

Elleman, A.M. (2017). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology*, 109(6), 761–781 https://doi.org/10.1037/edu0000180.

Enzmann, D. (2015, January 12) Notes on effect size measures for the difference of means from two independent groups: The case of Cohen's d and Hedges' *g*. Technical Report. doi: https://doi.org/10.13140/2.1.1578.2725

Field, A. & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63(3), 665–694 https://doi.org/10.1348/000711010x502733.

Foasberg, N. (2011). Adoption of e-book readers among college students: A survey. *Information Technology and Libraries*, 30(3), 108–128 https://doi.org/10.6017/ital.v30i3.1769.

Follmer, D.J. (2018). Executive function and reading comprehension: A meta-analytic review. *Educational Psychologist*, 53(1), 42–60 https://doi.org/10.1080/00461520.2017.1309295.

*Grace, K., (2011). Comparing the iPad to paper: Increasing reading comprehension in the digital age. (Unpublished Master Dissertation). Bowling Green State University, Bowling Green, Ohio, USA.

Graesser, A.C. & McNamara, D.S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3(2), 371–398 https://doi.org/10.1111/j.1756-8765.2010.01081.x.

Graesser, A., McNamara, D. & Kulikowich, J. (2011). Coh-Metrix. *Educational Researcher*, 40(5), 223–234 https://doi.org/10.3102/0013189x11413260.

*Green, T.D., Perera, R.A., Dance, L.A. & Myers, E.A. (2010). Impact of presentation mode on recall of written text and numerical information: Hard copy versus electronic. *North American Journal of Psychology*, 12(2), 233–242 Retrieved from http://arturo.ozunaeducators.com/wp-content/uploads/2011/10/50614008.pdf.

Hancock, G., Schmidt-Daly, T., Fanfarelli, J., Wolfe, J. & Szalma, J. (2016). Is e-reader technology killing or kindling the reading experience? *Ergonomics in Design: The Quarterly of Human Factors Applications*, 24(1), 25–30 https://doi.org/10.1177/1064804615611269.

Healy, S., Nacario, A., Braithwaite, R.E. & Hopper, C. (in press). The effect of physical activity interventions on youth with autism spectrum disorder: A meta-analysis. *Autism Research*. https://doi.org/10.1002/aur.1955, 11(6), 818–833.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128. https://doi.org/10.3102/10769986006002107.

*Heij, M., & van der Meij, H. (2014). The (in)effectiveness of pdf reading. (Unpublished Bachelor Thesis), Enschede, University of Twente, The Netherlands

*Hermena, E.W., Sheen, M., Al Jassmi, M., Al Falasi, K., Al Matroushi, M. & Jordan, T.R. (2017). Reading rate and comprehension for text presented on tablet and paper: Evidence from Arabic. *Frontiers in Psychology*, 8 https://doi.org/10.3389/fpsyg.2017.00257.

Herold, B. (2014). Digital reading poses learning challenges for students. *The Education Digest*, 80(1), 44.

Higgins, J.P.T. & Green, S. (Eds.) (2011). *Cochrane handbook for systematic reviews of interventions version 5.1.0*. London, UK: The Cochrane Collaboration Retrieved from www.handbook.cochrane.org.

Higgins, J.P., Thompson, S.G., Deeks, J.J. & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, 327(7414), 557–560 https://doi.org/10.1136/bmj.327.7414.557.

Hillesund, T. (2010). Digital reading spaces: How expert readers handle books, the web and electronic paper. *First Monday*, 15(4–5). https://doi.org/10.5210/fm.v15i4.2762.

Hosp, J.L. & Suchey, N. (2014). Reading assessment: Reading fluency, reading fluently, and comprehension--commentary on the special topic. *School Psychology Review*, 43(1), 59–69 Retrieved from http://www.nasponline.org/index2.html.

*Hou, J., Rashid, J. & Lee, K.M. (2017). Cognitive map or medium materiality? Reading on paper and screen. *Computers in Human Behavior*, 67, 84–94 https://doi.org/10.1016/j.chb.2016.10.014.

*Hou, J., Wu, Y. & Harrell, E. (2017). Reading on paper and screen among senior adults: Cognitive map and technophobia. *Frontiers in Psychology*, 8 https://doi.org/10.3389/fpsyg.2017.02225.

Hyman, J., Moser, M. & Segala, L. (2014). Electronic reading and digital library technologies: Understanding learner expectation and usage intent for mobile learning. *Educational Technology Research and Development*, 62(1), 35–52 https://doi.org/10.1007/s11423-013-9330-5.

Kamalski, J. (2004). How to measure the situation model. In A. Kerkhoff, J. de Lange & O.S. Leicht (Eds.), *Yearbook Utrecht Institute of Linguistics*, (pp. 121–134) Retrieved from http://www-uilots.let.uu.nl/research/Publications/Yearbook2004.pdf#page=129.

Kazanci, Z. (2015). University students' preferences of reading from a printed paper or a digital screen⊠A longitudinal study. *International Journal of Culture and History (Ejournal)*, 1(1), 50–53 https://doi.org/10.18178/ijch.2015.1.1.009.

*Kim, H. & Kim, J. (2013). Reading from an LCD monitor versus paper: Teenagers' reading performance. *International Journal of Research Studies in Educational Technology*, 2(1), 1–10 https://doi.org/10.5861/ijrset.2012.170.

Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessment. In J. Sabatini, E. Albro & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability*, (pp. 21–38). New York: Rowman & Littlefield Publishers, Inc.

Kong, Y., Seo, Y.S. & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, 123, 138–149 https://doi.org/10.1016/j.compedu.2018.05.005.

*Kretzschmar, F., Pleimling, D., Hosemann, J., Füssel, S., Bornkessel-Schlesewsky, I. & Schlesewsky, M. (2013). Subjective impressions do not mirror online reading effort: Concurrent EEG-eyetracking evidence from the reading of books and digital media. *PLoS One*, 8(2), e56178 https://doi.org/10.1371/journal.pone.0056178.

Kurata, K., Ishita, E., Miyata, Y. & Minami, Y. (2016). Print or digital? Reading behavior and preferences in Japan. *Journal of the Association for Information Science and Technology*, 68(4), 884–894 https://doi.org/10.1002/asi.23712.

*Lauterman, T. & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior*, 35, 455–463 https://doi.org/10.1016/j.chb.2014.02.046.

Leopold, C. & Leutner, D. (2012). Science text comprehension: Drawing, main idea selection, and summarizing as learning strategies. *Learning and Instruction*, 22(1), 16–26 https://doi.org/10.1016/j.learninstruc.2011.05.005.

Liberati, A., Altman, D., Tetzlaff, J., Mulrow, C., Gøtzsche, P., Ioannidis, J. et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine*, 6(7), e1000100 https://doi.org/10.1371/journal.pmed.1000100.

Lin, L. & Zabrucky, K. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, 23(4), 345–391 https://doi.org/10.1006/ceps.1998.0972.

MacGinitie, W.H., MacGinitie, R.K., Cooter, R.B. & Curry, S. (1989). Assessment: Gates-Macginitie reading tests. *The Reading Teacher*, 43(3), 256–258 Retrieved from https://www.jstor.org/stable/20200351.

Magliano, J.P., Millis, K., Ozuru, Y. & McNamara, D. (2007). A multidimensional framework to evaluate reading assessment tools. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies*, (pp. 107–136). New York: Lawrence Erlbaum Associates.

Mangen, A. & Kuiken, D. (2014). Lost in an iPad: Narrative engagement on paper and tablet. *Scientific Study of Literature*, 4(2), 150–177 https://doi.org/10.1075/ssol.4.2.02man.

*Mangen, A., Walgermo, B. & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, 58, 61–68 https://doi.org/10.1016/j.ijer.2012.12.002.

*Margolin, S., Driscoll, C., Toland, M. & Kegler, J. (2013). E-readers, computer screens, or paper: Does reading comprehension change across media platforms? *Applied Cognitive Psychology*, 27(4), 512–519 https://doi.org/10.1002/acp.2930.

Mason, L., Pluchino, P., Tornatora, M.C. & Ariasi, N. (2013). An eye-tracking study of learning from science text with concrete and abstract illustrations. *The Journal of Experimental Education*, 81(3), 356–384 https://doi.org/10.1080/00220973.2012.727885.

Mayer, R.E. (2009). *Multimedia learning*. (2nd edn). New York: Cambridge University Press https://doi.org/10.1017/CBO9780511811678.

Melby-Lervåg, M. & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, 140(2), 409–433 https://doi.org/10.1037/a0033890.

Mislevy, R.J. & Sabatini, J.P. (2012). How research on reading and research on assessment are transforming reading assessment (or if they aren't, how they ought to). In J. Sabatini, E. Albro & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability*, (pp. 119–134). New York: Rowman & Littlefield Publishers, Inc.

Mizrachi, D. (2015). Undergraduates' academic reading format preferences and behaviors. *The Journal of Academic Librarianship*, 41(3), 301–311 https://doi.org/10.1016/j.acalib.2015.03.009.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. & Prisma Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7), e1000097 https://doi.org/10.1371/journal.pmed.1000097.

Moran, J., Ferdig, R.E., Pearson, P.D., Wardrop, J. & Blomeyer, R.L., Jr. (2008). Technology and reading performance in the middle-school grades: A meta-analysis with recommendations for policy and practice. *Journal of Literacy Research*, 40(1), 6–58 https://doi.org/10.1080/10862960802070483.

Muijselaar, M., Swart, N., Steenbeek-Planting, E., Droop, M., Verhoeven, L. & de Jong, P. (2017). Developmental relations between reading comprehension and reading strategies. *Scientific Studies of Reading*, 21(3), 194–209 https://doi.org/10.1080/10888438.2017.1278763.

Muir, L. & Hawes, G. (2013). The case for e-book literacy: Undergraduate students' experience with e-books for course work. *The Journal of Academic Librarianship*, 39(3), 260–274 https://doi.org/10.1016/j.acalib.2013.01.002.

Mulloy, A.M., Gevarter, C., Hopkins, M., Sutherland, K.S. & Ramdoss, S.T. (2014). Assistive technology for students with visual impairments and blindness. In G. Lancioni & N. Singh (Eds.), *Assistive technologies for people with diverse abilities*, (pp. 113–156). New York: Springer https://doi.org/10.1007/978-1-4899-8029-8_5.

Narvaez, D., van den Broek, P. & Ruiz, A.B. (1999). The influence of reading purpose on inference generation and comprehension in reading. *Journal of Educational Psychology*, 91(3), 488–496 https://doi.org/10.1037/0022-0663.91.3.488.

*Neijens, P. & Voorveld, H. (2018). Digital replica editions versus printed newspapers: Different reading styles? Different recall? *New Media & Society*, 20(2), 760–776 https://doi.org/10.1177/1461444816670326.

Nell, V. (1988). The psychology of reading for pleasure: Needs and gratifications. *Reading Research Quarterly*, 23(1), 6–50 https://doi.org/10.2307/747903.

Nichols, M. (2016). Reading and studying on the screen: An overview of literature towards good learning design practice. *Journal of Open, Flexible and Distance Learning*, 20(1), 33 Retrieved from http://www.jofdl.nz/index.php/JOFDL/article/view/263/200.

*Norman, E. & Furnes, B. (2016). The relationship between metacognitive experiences and learning: Is there a difference between digital and non-digital study media? *Computers in Human Behavior*, 54, 301–309 https://doi.org/10.1016/j.chb.2015.07.043.

Noyes, J. & Garland, K. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, 51(9), 1352–1375 https://doi.org/10.1080/00140130802170387.

Perfetti, C. & Adlof, S.M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. In J. Sabatini, E. Albro & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability*, (pp. 3–20). New York: Rowman & Littlefield Publishers, Inc.

Picton, I. (2014). *The impact of e-books on the reading motivation and reading skills of children and young people: A rapid literature review*. London: National Literacy Trust Retrieved from https://eric.ed.gov/?id=ED560635.

*Porion, A., Aparicio, X., Megalakaki, O., Robert, A. & Baccino, T. (2016). The impact of paper-based versus computerized presentation on text comprehension and memorization. *Computers in Human Behavior*, 54, 569–576 https://doi.org/10.1016/j.chb.2015.08.002.

Randall, J.G., Oswald, F.L. & Beier, M.E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin*, 140(6), 1411–1431 https://doi.org/10.1037/a0037428.

Rapp, D. & van den Broek, P. (2005). Dynamic text comprehension. *Current Directions in Psychological Science*, 14(5), 276–279 https://doi.org/10.1111/j.0963-7214.2005.00380.x.

Reinking, D. (1988). Computer-mediated text and comprehension differences: The role of reading time, reader preference, and estimation of learning. *Reading Research Quarterly*, 23(4), 484–498. Retrieved from https://www.jstor.org/stable/747645.

Rockinson- Szapkiw, A., Courduff, J., Carter, K. & Bennett, D. (2013). Electronic versus traditional print textbooks: A comparison study on the influence of university students' learning. *Computers & Education*, 63, 259–266 https://doi.org/10.1016/j.compedu.2012.11.022.

Rosenfield, M., Jahan, S., Nunez, K. & Chan, K. (2015). Cognitive demand, digital screens and blink rate. *Computers in Human Behavior*, 51, 403–406 https://doi.org/10.1016/j.chb.2015.04.073.

Scammacca, N., Roberts, G. & Stuebing, K. (2014). Meta-analysis with complex research designs. *Review of Educational Research*, 84(3), 328–364 https://doi.org/10.3102/0034654313500826.

Schiefele, U. & Krapp, A. (1996). Topic interest and free recall of expository text. *Learning and Individual Differences*, 8(2), 141–160 https://doi.org/10.1016/s1041-6080(96)90030-8.

Schneider, W. & Lockl, K. (2002). The development of metacognitive knowledge in children and adolescents. *Applied Metacognition*, 224–258 https://doi.org/10.1017/cbo9780511489976.011.

Shenoy, P. & Aithal, P. (2016). A study on history of paper and possible paper-free world. *International Journal of Management, IT and Engineering*, Retrieved from https://ssrn.com/abstract=2779181.

Sidi, Y., Ophir, Y. & Ackerman, R. (2016). Generalizing screen inferiority—Does the medium, screen versus paper, affect performance even with brief tasks? *Metacognition and Learning*, 11(1), 15–33 https://doi.org/10.1007/s11409-015-9150-6.

Sidi, Y., Shpigelman, M., Zalmanov, H. & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction*, 51, 61–73 https://doi.org/10.1016/j.learninstruc.2017.01.002.

*Singer, L. & Alexander, P. (2017a). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education*, 85(1), 155–172 https://doi.org/10.1080/00220973.2016.1143794.

Singer, L. & Alexander, P. (2017b). Reading on paper and digitally: What the past decades of empirical research reveal. *Review of Educational Research*, 87(6), 1007–1041 https://doi.org/10.3102/0034654317722961.

*Singer Trakhman, L.M., Alexander, P.A. & Berkowitz, L.E. (in press). Effects of processing time on comprehension and calibration in print and digital mediums. *The Journal of Experimental Education.* https://doi.org/10.1080/00220973.2017.1411877, 1–15.

*Singer Trakhman, L.M., Alexander, P.A. & Silverman, A.B. (2018). Profiling reading in print and digital mediums. *Learning and Instruction*, 57, 5–17 https://doi.org/10.1016/j.learninstruc.2018.04.001.

Slavin, R.E., Cheung, A., Groff, C. & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, 43(3), 290–322 https://doi.org/10.1598/RRQ.43.3.4.

Sterne, J., Sutton, A., Ioannidis, J., Terrin, N., Jones, D., Lau, J. et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, 343(jul22 1), d4002–d4002 https://doi.org/10.1136/bmj.d4002.

*Stevens, B. (2014). E-reading comprehension versus conventional reading comprehension of junior high students. (Unpublished doctoral dissertation). Northcentral University, Prescott Valley, Arizona, USA.

Stevens, R.J., Lu, X., Baker, D.P., Ray, M.N., Eckert, S.A. & Gamson, D.A. (2015). Assessing the cognitive demands of a century of reading curricula: An analysis of reading text and comprehension tasks from 1910 to 2000. *American Educational Research Journal*, 52(3), 582–617 https://doi.org/10.3102/0002831215573531.

Szpunar, K.K., Jing, H.G. & Schacter, D.L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition*, 3(3), 161–164 https://doi.org/10.1016/j.jarmac.2014.02.001.

Takacs, Z., Swart, E. & Bus, A. (2015). Benefits and pitfalls of multimedia and interactive features in technology-enhanced storybooks. *Review of Educational Research*, 85(4), 698–739 https://doi.org/10.3102/0034654314566989.

Tarchi, C. (2017). Comprehending expository texts: The role of cognitive and motivational factors. *Reading Psychology*, 38(2), 154–181 https://doi.org/10.1080/02702711.2016.1245229.

*Taylor, A. (2011). Students learn equally well from digital as from paperbound texts. *Teaching of Psychology*, 38(4), 278–281 https://doi.org/10.1177/0098628311421330.

Thiede, K.W., Griffin, T.D., Wiley, J. & Redford, J.S. (2009). Metacognitive monitoring during and after reading. In D.J. Hacker, J. Dunlosky & A.C. Graesser (Eds.), *Handbook of metacognition in education*, (pp. 85–106). New York: Routledge.

Unsworth, N. & McMillan, B.D. (2013). Mind wandering and reading comprehension: Examining the roles of working memory capacity, interest, motivation, and topic experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3), 832–842 https://doi.org/10.1007/s12671-016-0615-8.

Valentine, J. & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods*, 13(2), 130–149 https://doi.org/10.1037/1082-989x.13.2.130.

van den Broek, P. & Kendeou, P. (2017). Development of reading comprehension: Change and continuity in the ability to construct coherent representations. In K. Cain, D.L. Compton & R.K. Parrila (Eds.), *Theories of reading development*, (pp. 283–307). John Benjamin's Publishing Company https://doi.org/10.1075/swll.15.

van den Broek, P., Lorch, R., Linderholm, T. & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition*, 29(8), 1081–1087 https://doi.org/10.3758/bf03206376.

van den Broek, P., White, M.J., Kendeou, P. & Carlson, S. (2009). Reading between the lines: Developmental and individual differences in cognitive processes in reading comprehension. In R.K. Wagner, C. Schatschneider & C. Phythian-Sence (Eds.), *Beyond decoding: The behavioral and biological foundations of reading comprehension*, (pp. 107–123). New York, NY: The Guildford Press.

Vidal-Abarca, E., Mañá, A. & Gil, L. (2010). Individual differences for self-regulating task-oriented reading activities. *Journal of Educational Psychology*, 102(4), 817–826 https://doi.org/10.1037/a0020062.

Wang, S., Jiao, H., Young, M., Brooks, T. & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments. *Educational and Psychological Measurement*, 68(1), 5–24 https://doi.org/10.1177/0013164407305592.

*Wells, C.L. (2013). Do students using electronic books display different reading comprehension and motivation levels than students using traditional print books? Unpublished doctoral dissertation, Liberty University, Lynchberg, Virginia, USA.

Wohlin, C. (2014) Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *8th International Conference on Evaluation and Assessment in Software Engineer*ing (EASE 2014), pp. 321–330. https://doi.org/10.1145/2601248.2601268

Woody, W., Daniel, D. & Baker, C. (2010). E-books or textbooks: Students prefer textbooks. *Computers & Education*, 55(3), 945–948 https://doi.org/10.1016/j.compedu.2010.04.005.

Zaphiris, P. & Kurniawan, H. (2001). Effects of information layout on reading speed: Differences between paper and monitor presentation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 45(15), 1210–1214 https://doi.org/10.1177/154193120104501512.

Ziefle, M. (1998). Effects of display resolution on visual performance. *Human Factors: The Journal of The Human Factors and Ergonomics Society*, 40(4), 554–568 https://doi.org/10.1518/001872098779649355.

Virginia Clinton, **PhD** is an Assistant Professor at the University of North Dakota and holds a masters' degree in Teaching English to Speakers of Other Languages from New York University. Her PhD degree is in Educational Psychology with a minor in Cognitive Science from the University of Minnesota. Her current research focuses on the relationship between affect, particularly mindfulness, and student cognition in learning situations.

Received 25 May 2017; revised version received 19 November 2018.

Address for correspondence: Virginia Clinton, Assistant Professor of Educational Foundations and Research, University of North Dakota, 231 Centennial Dr., Grand Forks, ND 58202, USA. E-mail: *virginia.clinton@und.edu*