Computers & Education 151 (2020) 103861

Contents lists available at ScienceDirect

Computers & Education

journal homepage: http://www.elsevier.com/locate/compedu

Assessing children's reading comprehension on paper and screen: A mode-effect study

Hildegunn Støle [a], Anne Mangen [a,*], Knut Schwippert [b]

[a] Norwegian Reading Centre, University of Stavanger, Norway
[b] Universität Hamburg, Germany, and University of Stavanger, Norway

ARTICLE INFO

Keywords:
Reading comprehension
Mode-effect
Children
Reading assessment
Screen reading

ABSTRACT

Recent meta-analyses (Delgado et al., 2018; Kong et al., 2018; Clinton, 2019) show that reading comprehension on paper is better than on screen among (young) adults. Children's screen reading comprehension, however, is underexplored. This article presents an experiment measuring the effect of reading medium on younger (10-year old) readers' comprehension, carried out in Norway in 2015. In a within-subjects design, students (n = 1139) took two comparable versions of a reading comprehension test – one on paper, and another digitally, with test version and order of medium counterbalanced. Probabilistic test theory models (two-parameter logistic (2 PL) and partial credit models) were employed for both versions of the test, allowing direct comparisons of student achievement across media. Results showed that the students in average achieved lower scores on the digital test than on the paper version. Almost a third of the students performed better on the paper test than they did on the computer test, and the negative effect of screen reading was most pronounced among high-performing girls. Scrolling and/or misplaced digital reading habits may be salient factors behind this difference, which sheds further light on children's reading performance and how this may be affected by screen technologies. Implications of these findings for education and for reading assessment are discussed.

1. Introduction and background

Computers, laptops, tablets and smartphones are ubiquitous in the lives of today's children and youth in large parts of the world (Mullis, Martin, Foy, & Hooper, 2017). Screen technologies have in many instances, replaced paper-based materials for reading, both in classrooms as well as in leisure contexts. Whether the current transition from paper-based reading to reading on screens affects cognitive learning outcomes, e.g. reading comprehension, has been the topic of an increasing number of empirical studies over the past couple of decades. Until recently, research findings on paper versus digital reading were inconsistent. Some studies (Aydemir, Öztürk, & Horzum, 2013) found better reading comprehension on screens than on paper, whereas others found no difference between the media (Hermena et al., 2017; Margolin, Driscoll, Toland, & Kegler, 2013; Porion, Aparicio, Megalakaki, Robert, & Baccino, 2016; Rockinson-Szapkiw, Courduff, Carter, & Bennett, 2013). Yet others found an advantage of paper reading (Golan, Barzillai, & Katzir, 2018; Halamish & Elbaz, 2019; Lenhard, Schroeders, & Lenhard, 2017; Mangen, Walgermo, & Brønnick, 2013; Singer & Alexander, 2017).

However, the emergence of meta-analyses has provided more clarity on this issue. The most comprehensive of these (Delgado,

* Corresponding author. Olav Hanssens vei 10, NO-4021, Stavanger, Norway.
E-mail address: anne.mangen@uis.no (A. Mangen).

https://doi.org/10.1016/j.compedu.2020.103861
Received 20 June 2019; Received in revised form 18 February 2020; Accepted 23 February 2020
Available online 28 February 2020

Vargas, Ackerman, & Salmerón, 2018), comprising 54 studies (n = 171,055 students) published between 2000 and 2017, found an advantage of paper over digital reading (Hedge's $g = -0.21$ for between-participants designs; $d_c = -0.21$ for within-participant designs) for the reading of informational, but not for narrative, texts. A comparable effect of medium on reading comprehension is reported in two other meta-analyses (Clinton, 2019; Kong, Seo, & Zhai, 2018). Moreover, Delgado et al. (2018) found that the advantage of paper-based reading had in fact increased rather than decreased during the period 2000–2017, casting doubt on claims about so-called digital natives displaying superior performance on screen. The fact that Delgado et al. (2018) statistically control for a potential publication bias, adds further weight to their results.

With a few exceptions (most recently, Halamish & Elbaz, 2019), studies measuring the effect of digitisation on reading comprehension are most typically performed among university students (e.g. Singer & Alexander, 2017; Chen, Cheng, Chang, Zheng, & Huang, 2014 [1]). Given the fact that also young children increasingly read from digital devices (Barzillai & Thomson, 2018; Duncan, McGeown, Griffiths, Stothard, & Dobai, 2016; Merga & Roni, 2016; Livingstone, Haddon, Vincent, Mascheroni, & Olafsson, 2014), and the increasing use of digital technologies in schools, it is important to find out whether children's reading comprehension differs when they read texts on screens compared to when they read in print. The fact that international (PIRLS,[2] and PISA[3]) as well as national (e.g. the US NAEP[4]) reading assessments are also increasingly digitised (Backes & Cowan, 2018; Eyre, Berg, Mazengarb, & Lawes, 2017), adds further weight to this topic and is a major motivation behind the present study. So far only a few studies have assessed the effect of reading medium on children's reading comprehension (e.g., Golan et al., 2018; Halamish & Elbaz, 2019; Kerr & Symons, 2006).

1.1. Digitised reading assessment

Norway has participated in close to all of the computer-based assessment (henceforth, CBA) additions in the international assessments, except PISA 2006 (CBA science), and it has had its own annual reading assessments for 10- and 14-year olds (5th and 8th grade) since 2004. These tests are modelled on the international PIRLS and PISA reading assessments and resemble them in many ways: the reading construct is similar, as is the operationalisation of this construct, and the item types employed to measure reading comprehension. The Norwegian National Reading Tests (henceforth: NRT) were paper-based until 2016, when the reading tests were digitised. [5]

Digital tests have some advantages over paper-based testing. One is cost reduction through automaticity of scoring. Automatic scoring eliminates scorer mistakes for some item types (multiple choice items), but not for other types (constructed response items; item types are explicated below). The widespread adoption of computers in education could potentially have allowed for further cost reduction in the actual data collection. Where school computer labs can be used for the tests, investigators no longer have to bring digital equipment like they did for the 2006 PISA Science CBA (Haldane, 2009). However, this has not come true, and for e.g. eTIMSS[6] 2019 in Germany, investigators still had to carry computers to many schools. Many countries prefer tablets (e.g., iPads, Chromebook) in instruction, and these cannot always be used for large-scale assessment, national or international. Cross-national assessments are valid only insofar as test conditions are identical. Thus, electronic assessment is sometimes harrowed by the same problem as research on effects of digital technologies: technology varies across borders and time.

On the positive side, CBA yields greater flexibility in text presentation than paper-based assessment (henceforth, PBA), for instance by using hyperlinks and dynamic illustrations, more like the online texts children are increasingly meeting on various electronic platforms. The IEA PIRLS reading assessment in 2016 took both traditional and digital reading features into account by providing two tests, one PBA and one CBA (ePIRLS) measuring specifically online informational reading.

However, most often digital assessments have simply replaced paper based reading assessments, because mode equivalence has been assumed. Jerrim (2016) found that in spite of bridging PBA and CBA PISA in order to ensure valid trend measurement and to check for mode equivalence internationally, there still were significant mode differences in some countries and demographic groups. Specifically, girls in many countries were disfavoured in the digital mathematics test condition compared to the paper version.

1.2. Mode equivalence in reading – research on children's reading comprehension

A common way to test the effect of reading medium (i.e., paper and screen) on reading comprehension, is to compare the informants' achievement when given two comparable versions of a reading test, one administered on screen and another administered

[1] Chen et al. (2014) used Chinese characters and texts in their study. For future research, it would be of interest to compare effects of reading device on reading across languages that vary along different parameters (e.g. reading direction; type of script).

[2] PIRLS: the *Progress in International Reading Literacy Study* takes place every fifth year and is organised by the International Association for the Evaluation of Education Assessment; IEA. The 2021 digital PIRLS is to continue the trend assessment from PBA PIRLS, measuring the traditional reading literacy concept which includes literary reading. In contrast, ePIRLS does not measure literary reading, but is described as an assessment of "online informational reading".

[3] PISA: the *Programme of International Skills Assessment* is a triennial international study comprising more than 70 countries and economies in 2015. The OECD is responsible for the PISA study. PISA measures not only reading, but also maths and science skills.

[4] NAEP: *National Assessment of Educational Progress*.

[5] Please note that the present study bears no relation to the National Reading Tests used in Norway today. The digital reading test has undergone radical changes since the present experiment took place.

[6] TIMSS: the *Trends in Mathematics and Science Study* takes place every fourth year since 1995. Like PIRLS, it is an IEA study. The eTIMSS is the digital version of TIMSS. It was first conducted in 2019.

on paper. While there is a number of studies comparing digital and paper-based reading comprehension among adults (for recent meta-analyses, see Delgado et al., 2018; Kong et al., 2018; Clinton, 2019), only a handful of mode-effect studies have been performed among children and adolescents.

Kerr and Symons (2006) conducted a mode-effect reading experiment among 5th grade students, i.e. children of about 10 years of age, finding that comprehension was more efficient from paper than from computer. Golan et al. (2018) found that 90 children of 11–12 years of age, preferred reading on screens, but performed better on a reading comprehension test when they had read from paper. Another study with children of the same age (Halamish & Elbaz, 2019) also found that the 5th graders (n = 38) had better reading comprehension on paper than on screens. In both studies (Golan et al., 2018; Halamish & Elbaz, 2019), most children had no metacognitive awareness that they read better in one mode (paper) than the other.

Eyre (2017) reports that in spite of careful preparation of an online version of the New Zealand reading comprehension assessment (grades 4 to 10), results were better for the paper version of the test. Mangen et al. (2013) compared 15-year olds' comprehension of text content from paper and digital formats, and Rasmusson (2014) conducted a similar study among 14-year olds taking a reading test. All these studies indicate that children's and adolescents' reading comprehension is more efficient when texts are presented in a paper-format compared to screens. Those studies asking about medium preference and checking self-efficacy in relation to medium (Golan et al., 2018; Halamish & Elbaz, 2019) reveal that many children had little awareness that their reading comprehension was better in PBA than in CBA. This parallels findings among university students in studies by Ackerman and Goldsmith (2011) and by Singer and Alexander (2017).

In contrast, a mode-effect study among 14-year olds conducted by Porion et al. (2016), showed no significant difference on comprehension and recall between the two modalities of paper and computer. As with several of the studies finding mode differences in favour of paper, the need to scroll was eliminated in Porion et al. (2016) study. Scrolling can be avoided by including only short texts. Whereas this ensures more identical conditions than experiments in which digital reading requires scrolling (or paging), this solution comes at the cost of ecological validity. Online reading, for instance of a Wikipedia article, typically requires scrolling. Chen and Lin (2016) found that scrolling negatively affected cognitive load when university students (n = 20) read (Chinese texts) from small mobile screens.

Another aspect of the present study adding to its ecological validity, is the fact that both test conditions contained multimodal texts. Children's repertoire of reading material, whether textbooks for educational purposes or texts for leisure reading, is very often multimodal. Hence, the types of material in the present study are more ecologically valid than texts consisting exclusively of linear text. Moreover, online reading often involves multimodal texts that may consist of non-continuous and non-verbal elements in addition to verbal text. Consequently, some reading assessments – whether national or international (e.g., PIRLS and ePIRLS) – have turned multimodal, containing images, tables, and other graphic elements in addition to written text.

Mode effect can potentially be affected by time limits in testing. Reading is self-paced in some studies, whereas it is limited in others. The latter ensures equal test conditions and assesses reading comprehension in terms of efficacy, but it is rarely how real-life reading occurs. Ackerman and Lauterman (2012) compared self-paced to time-limited reading in two conditions (paper and screen) finding that the test scores were lower on screen compared to paper under time pressure only, not when regulating reading time themselves (n = 80 university students; mean age 25.5 years). They attribute the difference to learners' inferior self-regulation in digital contexts compared to reading from paper. Our mode-effect study had a time limit (90 min per condition/test), since the final digital reading test.

1.3. Definition of reading comprehension

The latest PIRLS framework (2015) provides an overview of acknowledged theories of reading comprehension forming the basis for an encompassing definition of reading which includes non-verbal text elements and reading in different media. The PIRLS definition describes *reading literacy* (PIRLS' term denoting comprehension) as "a constructive and interactive process" (Mullis, Martin, & Sainsbury, 2015, chap. 1, p. 12).[7] The current definition informed both of the PIRLS 2016 reading literacy tests; the PBA PIRLS as well as the new ePIRLS assessment of online informational reading.

> Reading literacy is the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment. (Mullis et al., 2015, chap. 1, p. 12).

The PIRLS definition views the young reader as actively constructing meaning from texts, and that reading ability covers diverse text types, including multimodal texts. Reading occurs for learning and for enjoyment. The PIRLS definitions have served as models for NRT in Norway since 2004. It thus informs the mode effect experiment presented in this study.

1.4. Measuring reading comprehension in PIRLS and Norwegian Reading tests

The PIRLS reading construct is operationalised by deconstructing the complex ability *reading literacy* into four comprehension processes, used for two reading purposes: literary and informational. Both PIRLS and the NRT (and, consequently, our study) therefore include literary and informational texts. The four comprehension processes described for the PIRLS assessment (Mullis & Martin, 2015,

[7] For a full explication of the theoretical background to the reading framework employed in PIRLS, see Mullis et al. (2015), pp. 11–13.

p. 6) are reduced to three in NRT, but they are still very similar. Italics highlight identical concepts in PIRLS and NRT in the following description of reading comprehension processes: In order to understand texts, both literary and informational, the reader must be able to *find/retrieve explicitly stated information* (low level processing); s/he must be able to *make straightforward inferences* and *interpret and integrate* information (high level processing); and, at the highest level, the reader demonstrates the ability to *evaluate and critique content and text* form. Both tests aim at covering reading competence across the whole range of student ability. In other words, there is no ceiling effect, as both test have to yield information about advanced as well as poor readers.

The NRT and PIRLS employ identical item types to measure reading comprehension; multiple choice (MC) items require the reader to select one out of (usually) four options, of which only one is correct, and constructed response (CR) items require of the student that s/he writes or types a response. Such items are common also elsewhere in the assessment literature. While MC items are automatically scored in digital tests, the CR items are scored manually in accordance with guidelines for acceptable vs. non-acceptable answers (correct spelling is not a criteria for acceptable response).

NRTs are developed over the course of 1 ½ years, as the texts and items are assessed in an early field test, and then improved twice, before the final test is ready for launching. To ensure that the 5th grade reading test would be valid and reliable after the transition to a digital test mode, the piloting of texts and items was particularly careful. First, five sets of texts (25 texts in total) and items (about 200 different items) were pre-piloted in a field test among 5th grade students employing computers only. No paper condition was offered, since the goal of the field test was to check that all items and distractors functioned according to guidelines of discrimination and item and test difficulty in the finalised digital test. Items that did not meet the criteria of item fit (see below), were discarded or amended prior to the main pilot, i.e. the mode-effect experiment. We discarded five of the texts from the pre-pilot from further piloting. For the main pilot, i.e. our mode experiment, ten texts with high quality items were selected to measure reading performance in a CBA vs. a PBA, but otherwise identical, reading test.

1.5. The present study

This large-scale study (n = 1139) adds knowledge about mode-effect in reading achievement among 10-year old children. We present results from the double mode assessment of reading comprehension conducted in autumn 2015 in preparation of the first digital NRT among ten-year olds in Norway. The purpose of the experiment was thus two-fold: Its main goal was to provide a valid and reliable digital reading test for the 5th grade without changing the reading construct that for a decade informed the paper-based reading test. The study was carefully prepared to yield as valid results in the digital mode as it had yielded previously in the paper mode (much like that of the New Zealand PAT, cf. Eyre, 2017). The second purpose was to establish empirical evidence refuting or supporting a null hypothesis (e.g. Noyes & Garland, 2008) that test mode does not matter for reading comprehension among 10-year old children.

In testing, mode equivalence has sometimes been assumed (Noyes & Garland, 2008), but not always found (e.g. Backes & Cowan, 2018; Drabowicz, 2014; Jerrim, 2016). Based on the literature, we *hypothesise a mode effect* (H1) when comparing student results from PBA and CBA in our experiment. To get a more precise picture of the mode effect, we explore two additional hypotheses: mode effects for different reading skills levels (H2) and for gender (H3).

Reading ability varies among 10-year olds. The second hypothesis takes into account that children are developing readers who may still have to consolidate their reading competence and accumulate world knowledge as well as reading experiences. It is conceivable that good readers are relatively better at understanding texts presented in diverse media than poor readers are. Therefore, the second hypothesis (H2) is that *a mode effect on reading comprehension is less pronounced among high-performing readers compared to low-performing readers*.

Girls outperforming boys is a well-known phenomenon in reading assessment, occurring almost consistently in the countries participating in large-scale surveys such as PIRLS and PISA (Solheim & Lundetræ, 2018; Mullis et al., 2017). Since boys perform better on texts they like than on texts they do not like (Oakhill & Petrides, 2007), the present experiment included texts supposed to be of interest to boys. In addition, the tests employed many more multiple choice than constructed response items, since the latter may discourage boys (Solheim & Lundetræ, 2018). Further, it is a common expectation that boys perform better in digital modes, because computers may motivate them more than paper and pencil tests do (Martin & Binkley, 2009). Therefore, our third hypothesis (H3) was that *a mode effect on reading comprehension is less pronounced among boys than among girls.*

The three hypotheses are:

H1. There is an advantage of reading in a PBA condition vs. a CBA condition documented by the mean achievement of reading comprehension in both modes

H2. There is a mode-effect correlating with reading skill: low achievers show a larger mode effect (favouring PBA) than high-achieving students.

H3. There is a mode-effect correlating with gender: Boys show a smaller mode effect than girls.

2. Methods

2.1. Student sampling, test design and administration

The school and student samples were drawn by the Directorate of Education, ensuring representation of schools in Norway by

stratifying the sample by school-sizes, rural and urban districts, and regions of Norway. Around 1500 students were sampled; cf. the total number of students in 5th grade was c. 61,000 in 2015 (Skoleporten, Utdanningsdirektoratet.no). The digital tests were conducted on computers only, i.e. no tablets (e.g. iPad) have been used. The automatized data collection included a school code, gender and names of the students, in addition to the student's responses. The PBA test booklets were distributed to the schools with information about the pilot and instructions for teachers. All schools were also contacted by email with the same information and instructions enclosed.

The students responded to two sets, A and B, of comprehension-probing items connected to text reading, one set on paper and one set on computer, with set (A vs B) and medium (paper vs digital) counterbalanced. The students in each class in the A-sample were randomly assigned by their teacher to respond to test A either digitally (half of the class) or on paper (the other half) (cluster-randomized assignment of condition). The following week, those children having responded to test A on paper would now take test B digitally, while those having done A digitally were now assigned to test B on paper. The B-schools were instructed to follow the same design, only starting with test B in the first of the two test weeks, and then do A the week after.

Each test comprised five texts (10 texts in total, see 1.5 for selection criteria) with 36 and 35 items respectively in the A and B test. Both short and long texts were included, ranging in length from 204 to 683 words. The text types and formats varied from linear narrative fiction to multi-modal informational texts with non-linear elements. Most were informational texts written for children learning about curriculum-based subjects, such as history, natural science, social science, and literature.[8] We prepared a test that would engage boys and girls by including texts believed to appeal to them, and by including more MR items than CR items. These choices make our test somewhat different from e.g. PIRLS, as they may reduce gender effects (Solheim & Lundetræ, 2018).

The paper tests had a booklet format of A4 size. An instruction to the students occupied the first two pages, leaving the test itself to start from page 3. Many multimodal texts required scrolling in the computer mode, whereas they were presented in a double page layout on paper. Regarding scrolling, a text's word count did not matter as much as the amount of multimodal features, i.e. the more tables and illustrations the more scrolling. Students often had to turn a page to see the items connected to the text in PBA, while in the CBA solution, items appeared at the left hand side when the student had scrolled/read through the text. Items also required scrolling if there were many. In both test conditions, the student could move back and forth between texts and items. Both gave the opportunity to change responses, and the digital test even notified the student of unsolved items before submission.

Each student had to finish each test during 90 min at school, with maximum 5 min longer for the CBA as the instruction to students took slightly longer for the teacher to explicate. The two tests were administered by the teachers in two different, but consecutive, weeks. In total, the design yielded 2999 data sets. Matching data from each student who took both tests (and discarding poor quality data, described below), gave an $n = 1139$ students with two test conditions.

The teachers acted as instructors in line with information and guidelines we provided. We presumed that the national reading tests were well known among 5th grade teachers, even though the test mode was new. Many teachers have had previous experience with the PBA reading tests, while some may also have had experience conducting the CBA tests in mathematics or in English, since these were introduced a year before the reading mode study (in 2014). We expected CBA experience among the teachers to be of marginal, if any, effect, since computers have been common in schools for years.

Even though Norwegian 5th graders are familiar with computers,[9] we aimed at a low-threshold digital solution that all students would likely handle without difficulty. We did not conduct a usability test, because the digital platform had already been employed for assessments of mathematics and English. The students read from standard computer screens (c. 20 inches), while using a mouse to navigate and click selected multiple choice responses. They used keyboards to type their answers to constructed response items. The digital tests took place in the school computer lab or in classrooms equipped with computers (one per student). All students would likely have worked on these pc's previously.

After the two test weeks, the schools returned all paper booklets with student names, gender and information about first language. Each booklet was given a numeric code to allow matching with the automatically generated student-IDs from the digital data collection. A trained team of scorers conducted the punching of paper data and scoring of responses. Two persons scored each response, so that inter-scorer reliability could be ensured by checking scores for a randomly selected 10 per cent of the responses.

2.2. Scaling procedure

Since reading ability is not directly observable, tests are designed using a set of items that probe the comprehension of texts that students have read. Technically, reading tests assess the performance of students in relation to the actual testing materials. Test performance, in the form of achievement score, thus serves as a proxy for the latent student reading ability. The scores can be used for calculations of correlations with variables such as test mode and gender.

For each student, two achievement scores (one for PBA and CBA) were calculated by the employment of probabilistic test theory models in general, and 2 PL and partial credit models specifically, making it possible to compare directly student reading achievement in both test conditions. The logit scale was transformed to a linear scale with a mean of 50 and a standard deviation of 10. For the calculation, the scaling software Xcalibre 4.2 was used. Further analyses were performed by use of the standard software SPSS Vers. 23,

[8] Examples of text topics: the human skin, Harry Potter – an interview with the translator, the Viking warrior Harald Fairhair, the lost Inka civilisation.

[9] In 2016, 99 per cent of Norwegian 5th graders reported access to a computer or tablet at home and 98 per cent reported to have Internet access (PIRLS data, Støle & Schwippert, 2017).

if nothing else is mentioned.

A first descriptive Item Response Theory (IRT) analysis of the student responses to test items, revealed that some items occurred biased against the digital mode, meaning that the students got significantly ($p < 0.05$) poorer results in the CBA than in the PBA condition on identical items. Feedback from teachers and students had indeed indicated occasional problems with the technical solutions of the CBA: illustrations in.jpg formats did not download successfully in all classrooms. However infrequent the problem, to be on the safe (unbiased) side, we excluded from the mode effect analyses all items related to illustrations that may have been affected by problems in downloading. An unrelated problem was that some students had taken the same test twice, even though technical preparations were supposed to have eliminated this possibility by assigning individual student user names for the digital versions. We discarded all this data, from both paper and digital mode.

Our decision to eliminate potentially problematic data was due to a conservative strategy of seeking to avoid bias from factors bearing no relation to the comparison of PBA- and CBA-student responses. Finally, in agreement with this strategy, only data from students taking part in both test conditions was included in the calculations. We matched samples for the computer A and B test and the paper A and B test. Initially, the number of students was 1461 responses to the paper versions (A and B), and 1538 responses to the digital versions (A and B). After data cleaning and matching, the final number of students used for the analyses reported below is $n = 1139$.

We also discarded data from items that did not meet our criteria of item fit. We thus discarded items with poor discrimination (below $r_{bis} < 0.300$), while we kept items which were easy (item difficulties easier than −3.00 on the logit-scale) as long as they functioned according to the test theoretical principles.

The sample size ($n = 1139$) is sufficient to allow, in a matched pairs sample, for calculations of mean differences with a small effect size of $d = 0.2$ with a power of $1\text{-}\beta = 0.999$, and an α error of 5% (two-sided). For the calculation of the power, the program G-power (Version 3.1.9.2)[10] was used.

3. Results

3.1. Hypothesis 1: effect of test mode

Mean comparisons revealed that average test performance (i.e. achievement scores) was lower in CBA than in PBA. The digital reading test was significantly ($p < 0.05$) more difficult than the paper-and-pencil test. Hence, mode seems to matter for children's reading comprehension: in average, they comprehend texts better when reading them from paper. In order to shed further light on the effect of test mode on student performance, we calculated students' difference in performance between the two modes (see Fig. 1).

Fig. 1 shows the differences (rounded to full numbers) between PBA and CBA on the logit scale, where positive values indicate higher achievement in PBA, while negative values indicate an advantage in CBA. Fig. 1 illustrates that 599 students performed equally in both mode conditions (middle column): These students had the smallest performance deviation (−0.500 to 0.499 SD) between the CBA and the PBA condition. As many as 373 students performed better on PBA than on CBA (right-hand columns), whereas only 167 students gained from the CBA condition compared to PBA (left-hand columns). Fig. 1 also shows that 38 students vary considerably ($SD > |2$ points$|$) in performance in the two test modes.

Expressed as percentages, more than half of the students, 53 per cent achieved identical (or close to) results in the two test modes, while 13.6 per cent of the students scored about one scale point higher in CBA, and more than twice that number, 30.5 per cent, scored about one point higher in PBA. In other words, almost a third of the students performed better on paper. Next, we explore whether one test mode favours certain student group, for example high-performing students.

3.2. Hypothesis 2: skill level differences

Since the same students responded in both modes, PBA and CBA, the matched dataset can be used to calculate a *combined student ability theta value* by combining the individual theta values from both modes. This over-all student achievement was divided into three skills levels: Quartiles (25 + 50 + 25%) were used to differentiate between low (Q1), medium (Q2 & Q3) and high reading comprehension achievement (Q4). This allows for testing our second hypothesis that mode effect may be related to skills levels in such a manner that *low achievers show a larger mode effect (favouring PBA) than high-achieving students.*

Fig. 2 illustrates the differences between CBA and PBA in student performance at three skills levels. The difference in favour of print is significant at the $p < 0.05$ level for all three skills levels (Q1; Q2 & Q3; Q4).

In order to determine the value of the mode effects, we calculated effect sizes for each skills levels group. A common measure of the relevance of significance in means comparisons is effect size in terms of Cohen's *d* (e.g. Cohen, 1992; 1994). Means and effect sizes for each student skills group are reported in Table 1.

Table 1 reveals that the effect size is most pronounced for students in the Q4 group, i.e. the highest reading achievement group, at Cohen's $d = 0.44$. The effect size reaches $d = 0.41$ for the middle range students, while it is smallest for students who only reach Q1, the lowest achievers. Contrary to our hypothesis that weak readers would be more negatively affected by CBA than strong readers, it is the top-performing readers (Q4) who lose most in CBA. This may affect girls in particular.

[10] Available at: http://www.gpower.hhu.de/[13.3.2018].

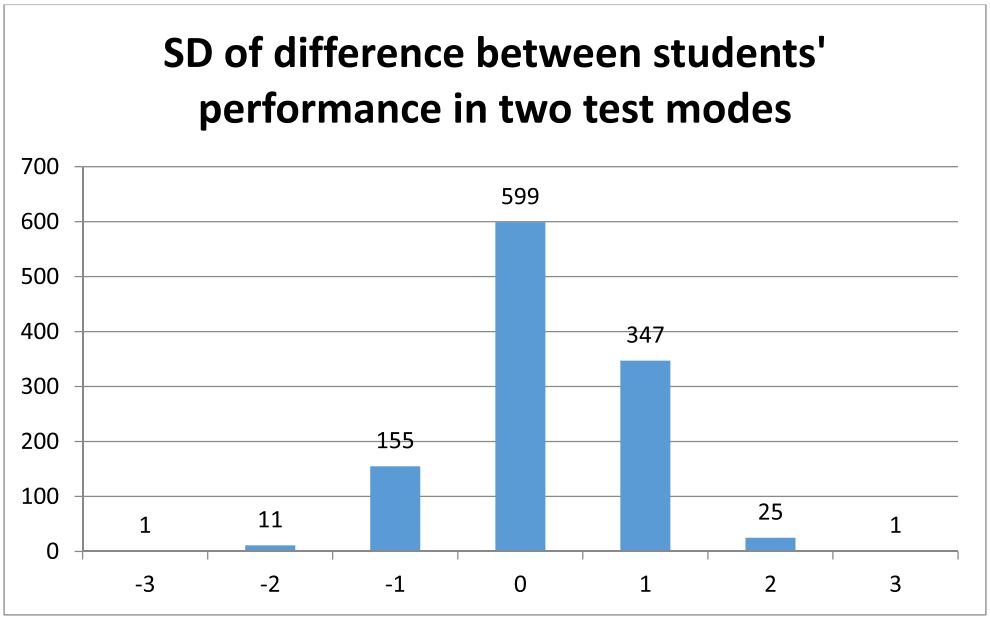

Fig. 1. Standard deviation of student differences in performance between the two test modes, CBA (negative values) vs PBA (positive values).

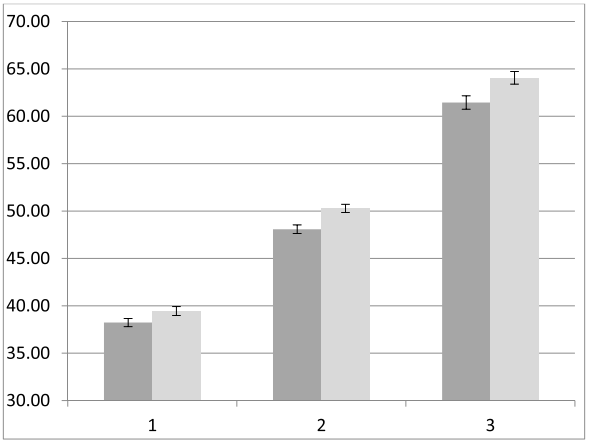

Fig. 2. Three levels of reading comprehension skills: from the lowest 25%, middle 50% to the highest 25%. Differences in performance is significant in favour of paper (light grey) across all three ability levels, cf. CI bars not overlapping.

In sum, we did not find support for the second hypothesis in our data. It emerged that students at all skills levels performed significantly better on PBA than on CBA (see Fig. 2 and Table 1) with an increasing mode effect (see the adjusted mean differences as Cohen's *d* in Table 1) for higher achieving students.

3.3. Hypothesis 3: gender differences

We used the combined theta value to explore reading achievement of boys and girls in the two test modes. Since we had six students whose gender could not be identified, the n = 1139 is reduced to n = 1133 in the following analyses. Table 2 shows the gender distribution on all three levels of reading achievement. T-values larger than 1.96 (or smaller than −1.96) indicate significant differences.

Table 2 shows the expected pattern for reading assessment, insofar as more boys (n = 162) than girls (n = 120) perform on the lowest reading achievement level (Q1), and more girls (n = 154) than boys (n = 130) perform at the highest level (Q 4). As means are calculated from a lower number of boys (130) than girls (154) at skills level 3/Q4, Table 2 shows numerically higher means for the top-performing boys than for girls at the same level of ability, but only the difference in CBA is significant between the genders with a t-value of 2.16.

Table 1
Three levels of reading achievement, means, standard deviation, standard error, confidence intervals (CI) and Cohen's d expressing the effect size of the difference between test modes; CBA to the left, PBA to the right. The effect sizes occur as negative values due to CBA being the point of comparison.

Reading Achievement	CBA Mean	SD	s.e.		CI		PBA Mean	SD	s.e.		CI		Cohen's *d*
Q 1	38.23	3.73	0.22	37.80	–	38.67	39.45	3.98	0.24	38.98	–	39.91	−0.31
Q 2 & 3	48.08	5.41	0.23	47.64	–	48.53	50.28	5.29	0.22	49.85	–	50.72	−0.41
Q 4	61.45	6.03	0.36	60.75	–	62.15	64.06	5.79	0.34	63.39	–	64.73	−0.44

Table 2
Boys and girls at three levels of mean reading achievement based on combined theta values with differences between both gender and test mode.

Gender	CBA			PBA			Diff. CBA		Diff. PBA	
Low to high achievement	Mean	s.e.	n	Mean	s.e.	n	boy-girl	t-value	boy-girl	t-value
boy	38.15	0.31	162	39.18	0.32	162	−0.15	−0.34	−0.58	−1.22
girl	38.30	0.32	120	39.76	0.35	120				
boy	48.07	0.33	285	50.33	0.33	285	0.02	0.04	0.08	0.18
girl	48.05	0.31	282	50.25	0.30	282				
boy	62.29	0.52	130	64.23	0.51	130	1.54	2.16	0.30	0.43
girl	60.75	0.49	154	63.93	0.47	154				

Table 2 also shows that the means are consistently higher in PBA than in CBA for both boys and girls. In other words, we find no support for hypothesis 3 that boys perform better on CBA. However, Table 2 reveals another interesting result: at the highest skills level (Q4), we find a larger mode-effect in favour of paper for girls than for boys. Hence, the digitisation of reading tests seem to come with a particular disadvantage for girls that are at the highest level of reading performance.

We explored mode effects within reading skills levels and gender groups further. Tables 3 and 4 show the differences between test modes and their effect on reading achievement among boys (Table 3) and girls (Table 4) at three levels of reading comprehension.

Means with confidence intervals (CI) in Tables 3 and 4 document consistent and significantly higher scores for both boys and girls in the PBA mode than in the CBA mode. Negative values in effect sizes indicate better achievement in PBA. The strongest effects appear in the following groups: boys at medium achievement level (level 2, i.e. Q2 & Q3), and girls at medium and high achievement levels (levels 2 and 3). Top-performing girls especially, experience a significant mode effect, with an effect size of $d = 0.53$. In other words, we found that top-performing girls suffer more negative effects from CBA than girls performing less well, and also than boys at all levels of reading achievement.

4. Discussion

In light of the emerging knowledge we have that young adults tend to comprehend what they read from paper better than when they read from screens (see e.g. the meta-study by Delgado et al., 2018 for an overview), we seized the opportunity to perform a mode effect study among children when preparing for a digital national reading test in Norway. Our experiment revealed significant differences in reading performance among 10-year olds as an effect of assessment mode. The mode effect reached significance not only on average achievement, but for students at all three levels of reading comprehension, and among both boys and girls. Top-performing girls were particularly disfavoured by the digital test mode.

Several factors may have contributed to the result. One of them could be children's (lack of) digital skills and experience. However, two aspects of the experiment make this unlikely: First, we aimed at a low-threshold digital solution not requiring more technical experience than a 10-year old can be expected to master. Secondly, children's digital skills have been a prioritised area in Norwegian educational policy. Computers, and increasingly tablets, are common learning tools in primary schools. There is also evidence that Norwegian 10-year olds have plentiful experience with digital devices at home. In PIRLS 2016, Norway was on top in children's access to digital devices: 99% had access to a computer or other online digital device; 98% had access to the Internet in the home (Mullis et al., 2017).

Table 3
Boys' reading achievement with means, SD, s.e. and CI in test modes CBA and PBA, with effect sizes: Cohen's *d*.

Boys' Reading	CBA						PBA						Cohen's *d*
Achievement	Mean	SD	S.E.		CI		Mean	SD	S.E.		CI		
Q 1	38.15	3.90	0.31	37.55	–	38.75	39.18	4.08	0.32	38.55	–	39.81	−0.26
Q 2 & 3	48.07	5.54	0.33	47.43	–	48.71	50.33	5.52	0.33	49.69	–	50.97	−0.41
Q 4	62.29	5.88	0.52	61.28	–	63.30	64.23	5.82	0.51	63.23	–	65.24	−0.33

Table 4
Girls' reading achievement with means, SD, s.e. and CI in test modes CBA and PBA with effect sizes: Cohen's *d*.

Girls' Reading	CBA						PBA						Cohen's *d*
Achievement	Mean	SD	S.E.	CI			Mean	SD	S.E.	CI			
Q 1	38.30	3.51	0.32	37.67	–	38.92	39.76	3.83	0.35	39.07	–	40.44	−0.40
Q 2 & 3	48.05	5.27	0.31	47.44	–	48.67	50.25	5.06	0.30	49.66	–	50.84	−0.43
Q 4	60.75	6.11	0.49	59.79	–	61.72	63.93	5.80	0.47	63.02	–	64.85	−0.53

Another factor which could have influenced results is the time limit. There were noticeably more missing responses towards the end of the CBA compared to the PBA test. In other words, reading from paper may be more efficient than reading from screens when reading is not self-paced (Ackerman & Lauterman, 2012). In our calculations, however, this was accounted for by IRT procedures, as missing were treated as null-responses when occurring as a string of the five last responses and not as faulty responses. This means that only given responses (and random missing responses) were scored for calculations of student achievement.

Kerr and Symons (2006) found that reading time increased for 10-year olds when reading a computer presented text. In their experiment, comprehension relative to reading rate was less efficient in CBA compared to paper. One factor which might have influenced reading time in our study is the amount of scrolling the students had to do in order to read the texts and respond to the items in CBA. Several studies report scrolling to have a negative impact on text recall and/or reading comprehension among adult readers (e.g. Sanchez & Wiley, 2009; Delgado et al., 2018). It has been suggested that scrolling disrupts the reader's (and writer's) sense of text structure and spatial location of information, and thus makes it more difficult to infer a global text perspective (Eklundh, 1992: adults writing in a scroll format). Scrolling is likely to draw on the limited working memory capacity that we need when reading for comprehension (Sanchez & Wiley, 2009). It is possible that young readers are even more affected than adults by text presentation requiring scrolling for full read-throughs and overviews of longer texts. Given that the present platform used a scrolling option, this may have added to the cognitive load in the digital version (Chen & Lin, 2016), although care was taken to ensure that text and graphic elements that provided the correct responses, appeared within the same screen page. Nevertheless, the fact that scrolling was needed to access the multimodal texts in their entirety, may have introduced an additional challenge in the digital version. However, our inclusion of multimodal texts adds to the ecological validity of the study. Such texts are common in digital textbooks in school (as well as in print textbooks), and multimodal, informational texts requiring scrolling are frequent in online reading.

Kerr and Symons (2006) further suggest that inferential comprehension, as a higher-order reading process, is more affected by computer reading than lower-order processes, such as retaining facts from the text. Our study supports this suggestion, since we found that top-performing students, i.e. those students most likely to succeed on items that assess higher-order reading, are the ones who were most disadvantaged in the CBA mode. This is an indication that it is more difficult for children to perform equally well on higher-order reading in digital modes than on paper. Schulz-Heidorf and Støle (2018), exploring PIRLS and ePIRLS data, indeed found that students were less successful responding to complex items (CRs requiring students to write/type a response) in CBA than in PBA.[11] CR items typically probe higher-order reading processes, and mode differences found for these items disfavoured girls in particular (Schulz-Heidorf; Støle, 2018).

It thus appears that digital assessment of cognitive skills tend to disfavour girls, like we have found for reading in the present study. Jerrim (2016), exploring PISA data from the 2012 mathematics tests, found a gender difference in PISA between paper and digital modes disfavouring girls in 20 out of the 32 countries. In our mode experiment, the effect size of the difference between PBA and CBA for high performing girls was $d = 0.53$ in favour of PBA. Even if this can be characterised as a medium rather than large effect size (Cohen, 1992), it is far from ignorable. Smaller effect sizes in testing have caused great concern among educationalists. For example, the gender differences in the 2011 Norwegian Reading test and in Norwegian PIRLS results from the same year (and cohort) were of considerably smaller effects sizes with $d = 0.18$ and $d = 0.24$ respectively (Solheim & Lundetræ, 2013, p. 64).

A qualitative user test conducted prior to the mode effect study (Mangen & Støle, n.p.) provided support for the hypothesis that children's concentrated reading takes longer on screen than on paper. In the user test, we observed a number of children tracing text lines with a finger on their pc screens. This reading behaviour was also found among college students of c. 19 years of age taking a reading comprehension test on computers (Margolin et al., 2013). In the study by Margolin et al. (2013), p. 43% of the students reported tracing lines with a finger on computer, while 53% did so on paper. For some children in our user test, this behaviour carried over to a medium (computer) for which it may not be very efficient, as it likely slows down reading. It might further indicate that some children find it difficult to perform concentrated reading on screens. We thus have two seemingly contradictory but complementary phenomena which could help explain our results: Careful readers apply a tactile paper reading strategy that slows their reading down, perhaps impacting fluency, whereas good readers apply digitally acquired reading strategies that make their reading and responding to items faster and more shallow in CBA compared to PBA. A couple of studies among university students (Ackerman & Goldsmith, 2011; Singer & Alexander, 2017) and among children (Halamish & Elbaz, 2019) find that readers in both age groups are unaware of their better reading from paper than from screens. Often they overestimate how well they understand texts they have read on screen (Ackerman & Goldsmith, 2011), and many prefer to read on screens or have no specific preference (Halamish & Elbaz, 2019) even though they perform better on a reading comprehension test when they have read on paper.

More and more reading in and out of school occurs on screens of many sorts. In the home, children mainly use computers and

[11] This occurred in spite of better performance by the average student on CBA (ePIRLS) than on PBA (PIRLS) in Norway.

mobile digital devices for news updates, socialising and communication, information searches, searching for entertainment, watching films and listening to music (Livingstone et al., 2014). As online reading typically involves skimming and scanning, rather than reading for pleasure or to learn, it is possible that some children develop a screen reading behaviour that is not beneficial for deep reading for comprehension. Wolf (2018) refers to this as a "bleeding over" effect. If children's screen reading is modelled on reading strategies efficient for quick and superficial reading, this may explain why many students in our experiment performed more poorly on CBA reading compared to PBA. In other words, the mode differences may be due to children's adapting a reading (and writing) behaviour formed by frequent use of digital media (Baron, 2015), rather than by lack of such experiences. It also explains why the reading mode appears to affect top-performing students (often girls) in particular, if their reading and responding in CBA is influenced by reading and writing strategies they employ when reading and writing on their digital devices.

Long form text reading, specifically book reading, is a strong predictor of reading ability (Cunningham & Stanovich, 1997), even in our digital age. Pfost, Dörfler, and Artelt (2013) found that the best readers in a German longitudinal study read books frequently, while the poorest readers (out of five categories) read little, but used digital devices extensively for socialising. The poorest readers did not read books, and Pfost et al. (2013) even found their digital media use to have a negative impact on reading performance. Duncan et al. (2016) similarly found that traditional (print) reading predicts reading comprehension ability whereas digital reading does not. Extending this line of research and using the PISA 2009 database with data for more than 250,000 teenagers from across 35 OECD countries, Jerrim and Moss (2019) found evidence that teenagers who spend more time reading, specifically, *fiction texts* (typically, novels and stories in books) have significantly stronger reading skills than peers who do not read, or read less, fiction. The authors call it the "fiction effect", since no associations were found between the frequency of reading non-fiction, news, magazines, or comics, and reading skill (Jerrim & Moss, 2019). As long form text reading is well-known to develop reading ability, long form text reading still has its place both in schools and in digital reading assessment. Future digitised reading tests might ameliorate the mode effect caused by scrolling by measuring only reading of short text passages. However, this may threaten the validity of reading comprehension tests as it makes it more difficult to test deep comprehension and reflection.

Even though Suang, Chang, and Liu (2016) identify several studies documenting positive learning outcomes when digital technologies are used in teaching, no consistent correlation has been found between a country's investment in digital technologies for education and the results in skills assessment of reading, science and mathematics as measured in PISA (OECD, 2015). It seems likely that digital devices are useful for concrete, short term learning goals, whereas complex cognitive skills like reading comprehension are best developed through traditional print reading. Reading ability is long term learning, developed throughout life. Støle and Schwippert (2017), for example, found that the association between book reading and achievement on the digital ePIRLS was much stronger than that of digital media use (at home and in school) and reading achievement on the digital ePIRLS. We need a more nuanced picture of what various digital technologies are good for, and when long form print (book) is preferable for learning. We also need to know whether CBA or PBA gives fairer results in assessment of young readers' comprehension.

4.1. Implications for testing and education

There are still some uncertainties as to whether CBA is appropriate for testing reading among all age groups. Children may not have the metacognitive skills necessary to apply differentiated reading strategies for screen reading and for paper reading. The deep reading for comprehension typically measured in reading tests, may be met with the skimming and scanning strategies often applied when searching for online information and entertainment (Baron, 2015; Liu, 2005; Wolf, 2018). These issues are important for policymakers and educators, as the trend towards increasing use of digital technologies in reading assessment and reading education may be based on assumptions rather than on facts about whether digitally presented texts actually offer children good opportunities of learning to read and develop their reading comprehension.

Frequent high-stakes digital assessment among children (e.g. annual national testing) may lead teachers and educators to infer that reading instruction is best done digitally. Even though digital technologies have potential for learning (Suang et al., 2016), they seem not to offer the qualities print reading has for developing reading skill among children. Children spend much time online in their spare time (e.g. Livingstone et al., 2014; OECD, 2015). For many, this happens at the cost of leisure reading (Merga & Roni, 2016), even when children report that they enjoy book reading. Therefore, it is important that schools do not supplant book reading with digital reading. Instruction in reading strategies for online (deep and critical) reading obviously has its place in school, but in order to ensure comprehension development, children still need time to read enjoyable long-form texts to consolidate reading, develop vocabulary, automaticity and fluency, and thereby comprehension. If this does not happen in the home, it is even more urgent that schools encourage book reading.

It has yet to be documented that digital technologies are apt for all kinds of learning in education. It is possible that reading behaviour, for instance, takes different forms in different media (Baron, 2015; Liu, 2005; Wolf, 2018), and that the reading behaviour which is useful for instance for information seeking on the Internet, is applied for all reading purposes when children read on screen, whether this is a functional reading strategy or not. This may be particularly true of children growing up in technology-rich environments spending more time online than with paper-based reading material such as textbooks and children's books. Children are still in the process of developing reading comprehension as well as discovering purposeful reading strategies for different kinds of texts in different media. Therefore, it is important that skills testing is sensitive to developmental aspects, especially when it comes to reading. It is also important that teachers teach children strategies for concentrated reading on screens.

5. Conclusions

Our results show that 10-year old children across levels of reading competence, in average performed significantly better on a reading test presented on paper than on screen. We did not find that CBA affected poor readers more than advanced readers. On the contrary, our results revealed that the mode effect (in favour of paper) was largest for high-performing girls. This finding occurs in spite of Norwegian children's plentiful access to and experience with digital devices and the Internet (Mullis et al., 2017; OECD, 2015). Digital reading habits as well as scrolling may indeed have affected results negatively in the CBA.

The added value of our study lies specifically in its contribution to the still relatively sparse empirical evidence documenting children's reading comprehension on paper vs on screen. It broadens the field by employing multimodal text types that are typical of student textbooks and of online texts. Another particular strength is the large sample size. Our finding that students across all skill levels perform more poorly on a digital test than on paper, is an urgent call for a more nuanced perspective on implementation of digital technologies in elementary education, and a signal to policy makers, school administrators and educators that the medium matters, especially for reading comprehension. Moreover, the finding that the negative effect of screen reading is particularly pronounced for the highest-performing girls calls for a reconsideration of assumptions that digitisation may make a difference only, or mostly, for the poorer readers.

6. Limitations

Due to the well tested definition of reading literacy and the operationalisation of this construct (both modelled on PIRLS), the results are valid to describe the reading ability of students. The conservative and careful handling of data, as well as the large sample size, add to the reliability and generalisability of our findings. However, generalisability of the results may be limited in two respects. First, as with all digital tests, results may in part depend on the selected digital solution. Digital technologies and test delivery platforms change. It makes sense to replicate the study on for instance, tablets or smartphones, which are increasingly introduced in many schools. However, since we employed a low-threshold solution that Norwegian children should be able to handle, it seems likely that the significant mode differences relate to aspects that are difficult to escape in real-world online reading, namely scrolling and superficial reading strategies.

The second limitation concerns the fact that the Norwegian 5th grade reading test is developed with an eye to reducing gender differences by selecting texts and items that engage boys in reading. Similar mode-effect studies among children may therefore yield other results regarding gender differences in reading.

Author contribution statement

Hildegunn Støle: Conceptualisation, Methodology, Analysis, Writing – Original Draft, Writing – Review & Editing. **Anne Mangen**: Conceptualisation, Methodology, Analysis, Writing – Original Draft, Writing – Review & Editing. **Knut Schwippert**: Conceptualisation, Methodology, Analysis, Writing – Original Draft.

Funding

No external funding but the employer of all authors at the time, the National Centre of Reading Research and Education, Norway.

Declaration of competing interest

None.

Acknowledgements

We are grateful that research for this article was made possible by the Norwegian Directorate of Education. Careful data management is of the utmost importance: We therefore wish to thank our colleague Brenda Spierings for her conscientious handling of paper test booklets, her help in preparing the digitised test, communication with schools, and the meticulous tracking of student-ids to enable matching. We would also like to thank the many teachers and students who participated in the mode-effect experiment. We also wish to thank the reviewers who provided us with good advice which helped update and improve and this text/article.

References

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

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.

Aydemir, Z., Öztürk, E., & Horzum, M. B. (2013). The effect of reading from screen on the 5th grade elementary students' level of reading comprehension on informative and narrative type of texts. *Educational Sciences: Theory and Practice, 13*(4), 2272–2276.

Backes, B., & Cowan, J. (2018). *Is the pen mightier than the keyboard? The effect of online testing on measured student achievement*. National Center for Analysis of Longitudinal Data in Education Research. Working Paper 190. April 2018.

Baron, N. S. (2015). *Words onscreen: The fate of reading in a digital world*. USA: Oxford: Oxford University Press.

Barzillai, M., & Thomson, J. M. (2018). Children learning to read in a digital world. *First Monday, 23*(10), 1–10.

Chen, G., Cheng, W., Chang, T.-W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screen, and tablets: Does tablet familiarity matter? *Journal of Computers and Education*. https://doi.org/10.1007/s40692-014-0012-z. Springer online 06 August 2014.
Chen, C.-M., & Lin, Y.-J. (2016). Effects of different text display types on reading comprehension, sustained attention and cognitive load in mobile reading contexts. *Interactive Learning Environments, 24*(3), 553–571.
Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading, 42*(2), 288–325.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155.
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003.
Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*(6), 934–945.
Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review, 25*, 23–38.
Drabowicz, T. (2014). Gender and digital usage inequality among adolescents: A comparative study of 39 countries. *Computers & Education, 74*, 98–111, 10 1016/j.compedu.2014.01.016.
Duncan, L. G., McGeown, S. P., Griffiths, Y. M., Stothard, S. E., & Dobai, A. (2016). Adolescent reading skill and engagement with digital literacies as predictors of reading comprehension. *British Journal of Psychology, 107*, 209–238. https://doi.org/10.1111/bjop.12134.
Eklundh, K. S. (1992). Problems in achieving a global perspective of the text in computer-based writing. *Instructional Science, 21*, 73–84.
Eyre, J. (2017). On or off screen Reading in a digital world. *Assessment News set, 1*, 53–58, 10.18296/set.0072.
Eyre, J., Berg, M., Mazengarb, J., & Lawes, E. (2017). *Mode equivalency in PAT: Reading comprehension*. New Zealand Council for Educational Research.
Golan, D. D., Barzillai, M., & Katzir, T. (2018). The effect of presentation mode on children's reading preferences, performance, and self-evaluations. *Computers & Education, 126*, 346–358.
Halamish, V., & Elbaz, E. (2019). Children's reading comprehension and metacomprehension on screen versus on paper. *Computers & Education, 145*.
Haldane, S. (2009). Delivery platforms for national and international computer-based surveys. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 63–67).
Hermena, E. W., Sheen, M., AlJassmi, M., AlFalasi, K., AlMatroushi, M., & Jordan, T. R. (2017). Reading rate and comprehension for text presented on tablet and paper: Evidence from Arabic. *Frontiers in Psychology, 8*, 257.
Jerrim, J. (2016). Pisa 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice, 23*(4), 495–518. https://doi.org/10.1080/0969594X.2016.1147420.
Jerrim, J., & Moss, G. (2019). The link between fiction and teenagers' reading skills: International evidence from the OECD PISA study. *British Educational Research Journal, 45*(1), 181–200.
Kerr, M. A., & Symons, S. E. (2006). Computerized presentation of text: Effects on children's reading of informational material. *Reading and Writing, 19*, 1–19. https://doi.org/10.1007/s11145-003-8128-y.
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.
Lenhard, W., Schroeders, U., & Lenhard, A. (2017). Equivalence of screen versus print reading comprehension depends on task complexity and proficiency. *Discourse Processes, 54*(5–6), 427–445.
Liu, Z. (2005). Reading behaviour in the digital environment: Changes in reading behaviour over the past ten years. *Journal of Documentation, 61*(6), 700–712. https://doi.org/10.1108/00220410510632040.
Livingstone, S., Haddon, L., Vincent, J., Mascheroni, G., & Olafsson, K. (2014). *Net children go mobile: The UK report*. London: London School of Economics and Political Science.
Mangen, A., & Støle, H. (2014). *User test for digital national reading assessment, conducted in four classrooms at different schools in spring*.
Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper vs. computer screens: Effects on reading comprehension. *International Journal of Educational Research, 58*, 61–68.
Margolin, S. J., Driscoll, C., Toland, M. J., & Kegler, J. L. (2013). E-readers, computer screens, or paper: Does reading comprehension change across media platforms? *Applied Cognitive Psychology, 27*(4), 512–519.
Martin, R., & Binkley, M. (2009). Gender differences in cognitive tests: A consequence of gender-dependent preferences for specific information presentation formats? In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 75–82).
Merga, K. M., & Roni, S. M. (2016). The influence of access to eReaders, computers and mobile phones on children's book reading frequency. *Computers & Education, 109*, 187–196.
Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *ePIRLS 2016 international results in online informational reading*. Retrieved from Boston College. TIMSS & PIRLS International Study Center website http://timssandpirls.bc.edu/pirls2016/international-results/.
Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2015). PIRLS 2016 reading framework, 2015. In I. V. S. Mullis, & M. O. Martin (Eds.), *PIRLS 2016 assessment framework* (2nd ed.). TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*(9), 1352–1375. https://doi.org/10.1080/00140130802170387.
Oakhill, J. V., & Petrides, A. (2007). Sex differences in the effects of interest on boys' and girls' reading comprehension. *British Journal of Psychology, 98*, 223–235. https://doi.org/10.1348/000712606X117649.
OECD. (2015). *Students, computers and learning: Making the connection*. PISA OECD Publishing. https://doi.org/10.1797/9789264239555-en.
Pfost, M., Dörfler, T., & Artelt, C. (2013). Students' extracurricular reading behaviour and the development of vocabulary and reading comprehension. *Learning and Individual Differences, 26*, 89–102. https://doi.org/10.1016/j.lindif.2013.04.008.
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*, 569–579. https://doi.org/10.1016/j.chb.2015.08.002.
Rasmusson, M. (2014). Reading paper - reading screen. *Nordic Studies in Education, 35*, 3–19.
Rockinson-Szapkiw, A. J., 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.
Sanchez, C. A., & Wiley, J. (2009). To scroll or not to scroll: Scrolling, working memory capacity, and comprehending complex texts. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 51*(5), 730–738. https://doi.org/10.1177/0018720809352788.
Schulz-Heidorf, K., & Støle, H. (2018). Gender differences in Norwegian PIRLS 2016 and ePIRLS results at test mode, text and item format level. *Journal of Nordic Literacy Research, 4*(1).
Singer, L. M., & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education, 85*(1), 155–172.
Skoleporten. Utdanningsdirektoratet: Fakta ["School gate" facts by the Norwegian directorate of education]. https://skoleporten.udir.no/rapportvisning/grunnskole/fakta-om-opplaeringa/elevar-laerarar-skolar/nasjonalt?orgaggr=a&kjonn=a&trinn=5&sammenstilling=1&fordeling=2. (Accessed 3 December 2019).
Solheim, O. J., & Lundetræ, K. (2013). Prøveutformingens betydning for rapporterte kjønnsforskjeller: En sammenligning av kjønnsforskjeller i PIRLS og nasjonale prøver i lesing på 5. Trinn [the impact of test design on reported gender differences: A comparison of gender differences in PIRLS and national reading tests in grade 5]. In E. Gabrielsen, & R. G. Solheim (Eds.), *Over kneiken? Leseferdighet på 4. Og 5. Trinn i et tiårsperspektiv [Have we turned the corner? Reading skill in grades 4 and 5 in a ten-year perspective] 61-76*. Oslo: Akademika forlag.

[publ. online 2016]) Solheim, O. J., & Lundetræ, K. (2018). Can test construction account for varying gender differences in international reading achievement tests of children, adolescents and young adults? – a study based on nordic results in PIRLS, PISA and PIAAC. *Assessment in education: Principles. Policy & Practice, 25*(1), 107–126.

Støle, H., & Schwippert, K. (2017). Norske resultater fra e-PIRLS – online informational reading [Norwegian results from e-PIRLS – online informational reading]. In E. Gabrielsen (Ed.), *Klar framgang! Leseferdighet på 4. og 5. trinn i et femtenårsperspektiv [Visible progress! Reading skill in Grades 4 and 5 in a fifteen-year perspective]* (pp. 50–74). Oslo: Universitetsforlaget.

Suang, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education, 94*, 252–275.

Wolf, M. (2018). *Reader, come home: The reading brain in a digital world.* New York.