Chapter 13

Implications of Changing the Delivery Mode on Reading Tests in Norway—A Gender Perspective

Ragnhild Engdal Jensen

Abstract What can be seen as a digital shift in society is also visible in the Norwegian educational system, as the use of digital devices has increased in both teaching and learning activities. Together with some practical and logistical reasons, the former has very much facilitated the change of delivery mode of the Norwegian National Assessment of Reading Literacy. At the same time, a concern arose regarding whether the test will continue to measure the same underlying concept of reading as before. Furthermore, from the equity perspective, it is important that the change of mode is not disfavoured to any particular group of students. As a solution to this, the format of the test is preserved using fixed, as opposed to dynamic, texts, assuming that fixed texts are consumed in the same way regardless of whether they are presented on paper or on screen. Building on this, this chapter reports on a field trial study for the 2016 Norwegian National Assessment in reading. Nine hundred seventy-three eighth graders from nine different schools participated in completing reading tests on either paper or screen. The main aim of the study is to explore to what extent delivery mode seems to influence students' outcomes. In particular, we investigate whether the change in delivery mode affects boys' and girls' results on reading comprehension tests in the same way. For the purpose of analysis, the Rasch model will be used as a measure of student ability and a multiple regression model will be used to investigate gender differences across the modes. Based on the research so far, we assume that the change in mode will not have a significant impact on student performance relative to gender. The results will be discussed in the light of the gender gap in reading achievement present in the Norwegian educational system.

Keywords Reading comprehension · National tests in reading · Norway · Delivery mode · Gender

R. Engdal Jensen (✉)

Department of Teacher Education and School Research, University of Oslo, Oslo, Norway

e-mail: r.e.jensen@ils.uio.no

© The Author(s) 2020

T. S. Frønes et al. (eds.), *Equity, Equality and Diversity in the Nordic Model of Education*, https://doi.org/10.1007/978-3-030-61648-9_13

337

13.1 Introduction

Over the past decades, digital technologies have changed how we read and manage information. This phenomenon is evident in many aspects of our lives, and in the field of education, digital technologies are transforming teaching and learning, as well as the ways in which schools assess students. The background for this study relates to recent trends in paper-based reading assessments and their replacement with on-screen assessments. In 2015, the Programme for International Student Assessment (PISA) was delivered on computers for the first time, and in 2016, the Norwegian National Assessment of Reading Literacy Skills was too. The change reflects how students and societies now commonly access, use and communicate information (OECD, 2019), and it is advantageous considering the logistical aspects and security issues of administering the assessments. At the same time, there is a concern regarding whether the tests continue to measure the same underlying concept of reading as before. The use of digital devices as reading tools calls into question how these potentially alter perceptions of what it means to read and the comprehension that results from the activity itself (Singer & Alexander, 2017b).

This chapter explores to what extent delivery mode affects students' outcomes in reading, using the 2016 field study data of the Norwegian national test in reading. While the idea of 'school for all' dominates the school system in Norway, the change in the delivery mode may have significant implications for educational justice. From an equity perspective, it is important that the change of mode is not disadvantageous to any particular group of students. As the trend of girls outperforming boys on reading assessments is well known (Jensen et al., 2019; Solheim & Gourvennec, 2017), we investigate if the change in delivery mode affects boys' and girls' results differently and whether this change has implications for boys and girls having equal opportunities in the test situation.

13.2 Theoretical Background

13.2.1 *Mode Effect*

Dillon's (1992) review of the literature, intended to examine differences that might exist between reading from a print compared to an electronic source, is referred to as a starting point by several researchers (Delgado, Vargas, Ackerman, & Salmerón, 2018; Singer & Alexander, 2017b). In recent years, a large body of research has emerged, and several updated reviews have been published (Clinton, 2019; Delgado et al., 2018; Kong, Seo, & Zhai, 2018; Singer & Alexander, 2017b). The reviews vary in content and scope; still, all reviews find that, overall, readers demonstrate better comprehension when reading on paper compared to when they read on screen or digitally.

Singer and Alexander's (2017b) narrative review includes 36 studies from the period 2001–2017. Their examination of the literature showed that studies were diverse both in how they define reading in the different media, as well as in how text comprehension is measured. One important finding was that there seems to be an association between the length of the text and the medium, and that readers demonstrate significantly better comprehension when reading on paper if the texts are longer than 500 words or one page. If the texts are shorter, there is no significant difference in the reading comprehension of texts presented in different media. This was evidenced by over 90% of the charted studies in which text length was specified. Further, they emphasize that print seems to be the favourable processing medium when individuals are reading for depth of understanding and not solely for gist.

The review of Delgado et al. (2018) includes 54 studies conducted between 2000 and 2017 that compare reading comprehension when reading printed and digital texts. Thirty-eight of these studies had a between-participants design – participants read either on paper or on the screen – whereas 16 were within-participants studies in which participants read texts in both modes. The results of their meta-analysis showed an advantage for printed texts regardless of the design, with effect sizes being significant (Hedge's $g = -.21$, $d_c = -.21$). The reviews of Kong et al. (2018) and Clinton (2019) include a smaller number of studies – 17 and 33, respectively. Still, the meta-analyses show similar effect sizes to those found by Delgado et al. (2018) (Hedge's $g = -.21$ and $-.25$, respectively). Reading texts from a screen had a small but significant adverse effect on comprehension scores compared to reading from paper.

The tree reviews also included analysis on possible moderators. Clinton (2019) found that readers had a significantly better-calibrated judgement of their performance when reading from paper compared to digitally. Both Delgado et al. (2018) and Clinton (2019) found that the advantage of reading in print increased when participants read expository texts as opposed to narrative texts. Further, Delgado et al. (2018) found that the advantage of reading in print was significantly higher in studies with time constraints compared to studies where participants were allowed to self-pace their reading. This finding was not confirmed by Kong et al. (2018) and Clinton (2019). Although the moderating effect of scrolling did not reach significance, Delgado et al. (2018) emphasize this variable, as their analysis showed a substantial advantage for paper-based reading when scrolling was necessary to read texts on the screen. Finally, it is worth mentioning that text length was not found to be a significant moderator by Delgado et al. (2018), as suggested by Singer and Alexander (2017b).

13.2.2 Mode and Text Processing

The fore-mentioned reviews comprise much of the literature on the field of reading comprehension in different modes. In this section, single studies of particular relevance to the focus of this chapter will be highlighted and discussed in more detail. In a study by Wästlund, Reinikka, Norlander, and Archer (2005), two experiments were performed to investigate the influence of video display terminals (VDT) and paper presentation of text on reading comprehension and the production of information. The results from the study showed that participants reading from computers reported higher levels of experienced stress and tiredness compared to those reading on paper. Furthermore, they found that in both experiments, performance in the VDT presentation condition was inferior to that of the paper presentation condition. Hence, they concluded that the dual-task effects of fulfilling the assignment and working with the computer resulted in a higher cognitive workload.

The supposition that working with a computer results in a higher cognitive workload compared to working on paper is supported by several studies. In a comparison of an identical comprehension task presented on paper and on a computer, Mayes, Sims, and Koonce (2001) showed a significant negative relationship between workload and comprehension scores. The comprehension task was to read a text and answer ten multiple-choice questions, and the workload was measured by the Task Load Index (NASA-TLX). The result showed that increased workload was associated with lower scores. This finding was replicated with thirty undergraduate students by Noyes, Garland, and Robbins (2004). The students read an article, presented in a closely matched form either on paper or on a computer, and then answered ten multiple-choice questions to measure comprehension. Finally, the NASA-TLX was administered. The results showed that there was a significant difference in the perceived effort needed for the computer-based test, and further, that those with lower comprehension scores experienced a higher workload. These findings indicate that lower-performing individuals might be disadvantaged when completing computer-based assessments as compared to similar tasks on paper.

In their research, Noyes and Garland (2003) have also paid attention to the potential impact of presentation mode on cognitive processing and, in turn, learning performance. In a study that examined directly comparable text presented on screen and paper, they included a measure of memory awareness in addition to looking at reading time and comprehension. Such memory awareness measures have been widely used in psychology as a means of gauging recall and, hence, learning. They are based on the work of Tulving (1985), who developed the Remember-Know paradigm. The paradigm describes two types of retrieval response, 'Remember' and 'Know'. 'Remembered' knowledge is typically being recalled in association with related information about the learning episode, whereas 'known' knowledge is recalled without being tied to contextual details or associations. Tulving argued that with time, memory of specific events fades or reduces in contextual details. This implies that 'remembered' knowledge gets less accessible with time. Findings by

Conway, Gardiner, Perfect, Anderson, and Cohen (1997) suggest that knowledge that is 'known' is more readily applied and, as such, indicative of better learning.

The results from the study of Noyes and Garland (2003) indicate that when the material is matched adequately across media, reading time and number of correct answers do not differ. However, a significant effect of awareness frequencies was found in the study. The rate of 'remember' responses was approximately twice that of 'know' responses when reading on screen. In contrast, levels of 'remember' and 'know' responses were similar when reading on paper. The results indicate that cognitive processing associated with memory assimilation differs across mode conditions. Noyes and Garland (2003) suggest that characteristics of the computer screen, such as refresh rate and fluctuating luminance, might interfere with cognitive processing for long-term memory. These findings were confirmed in a later study (Garland & Noyes, 2004) that showed that the manner in which the knowledge was retrieved varied between presentation formats – screen and paper. The study was longitudinal, and the results suggest that repeated exposure to and rehearsal of computer-based information is needed to equate knowledge retrieval with that achievable from paper. The knowledge transition when reading from the screen was much more rapid compared to paper, which indicates that knowledge seems to be better adapted and, in turn, more easily applied when presented in paper format. Garland and Noyes conclude that "there still appears to be a benefit attached to learning from paper-based rather than computer-based material" (2004, p. 51).

Another possible explanation of the apparent comprehension differences across modes might be related to metacognitive skills (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014). When comparing the reading performance of undergraduate students who read identical texts on screen and paper, Ackerman and Goldsmith (2011) found that, under a fixed study time, test performance did not differ between the two media. However, when the study time was self-regulated, students performed poorer on screen than on paper. Further, the results showed that the students were less able to give an accurate prediction of their performance, tending to overestimate their comprehension when reading on screen. This was also accompanied by poorer allocation of study time. Hence, Ackerman and Goldsmith (2011) conclude that the primary difference between the two media is not cognitive but rather metacognitive. The authors conclude that metacognitive processes might be less effective on screen due to higher-order metacognitive beliefs. Previous research shows that people seem to perceive printed paper as the medium best suited for effortful learning, whereas the electronic medium is suited to fast and shallow reading of short texts, such as news and e-mails (Shaikh, 2004; Spencer, 2006). Such a perception might reduce the mobilization of cognitive resources that are needed for effective self-regulation (Ackerman & Goldsmith, 2011).

Also, research shows that people's use of digital media makes them less likely to engage in reflective thought (Annisette & Lafreniere, 2017). This is consistent with what has come to be known as the 'shallowing hypothesis' (Carr, 2010). The hypothesis proposes that the frequent use of ultra-brief social media, such as texts and tweets, characterized by quick, social interactions, promotes rapid, shallow and

non-reflective thought. Further, people typically process digital texts in a shallow or superficial way, and such digital activities might, in turn, prevent success when performing more complex activities that require sustained attention – for instance, processing longer texts.

The assumptions associated with the shallowing hypothesis are in line with findings showing that readers spend less time processing digital texts compared to paper-based texts. A study by Singer Trakhman, Alexander, and Berkowitz (2017) explored the effects of print and digital texts on readers' comprehension and processing time. They predicted that there would be differences in the time spent reading digital compared to printed texts, and further that processing time would serve as a mediator between medium and comprehension performance. This is in keeping with the speed-accuracy trade-off hypothesis (Wickelgren, 1977), which suggests a trade-off between the speed at which a certain task is performed and the quality of the product. The results showed that participants read significantly faster when texts were displayed on a computer than when texts were on paper, and that there was a significant direct effect of the medium on overall comprehension. Further, medium predicted processing time, which in turn predicted comprehension scores. Processing time significantly mediated the effects of the medium on readers' comprehension (Singer Trakhman et al., 2017).

Another topic that has been of interest to Singer Trakhman and colleagues (Singer & Alexander, 2017a; Singer Trakhman et al., 2017) is students' calibration when they read in print or digitally. Calibration can be defined as the distance between perceived performance and demonstrated levels of understanding or competence (Alexander, 2013). Singer and Alexander (2017a) examined whether students' judgments of their reading comprehension abilities under print and digital conditions would match their actual comprehension performance. The results showed that when asked to judge the medium in which they performed best, the majority of the participants indicated the digital medium. However, more students demonstrated stronger comprehension when they were reading on paper. This indicates that participants were generally poorly calibrated. The number of participants that presumed they would be better at performing in the digital medium but, in reality, comprehended better on paper, was significant. This is in line with the research of Ackerman and Goldsmith (2011) and was also confirmed by Singer Trakhman et al. (2017), who found that the participants' calibration was significantly worse when reading on screen compared to paper. They suggest that this may be explained by the potential influence of processing speed, and that calibration, as well as comprehension, might be subject to the speed-accuracy trade-off. They also suggest that there might be an association between the level of effort exerted in a task and the judgement of comprehension, referring to research by Koriati, Ma'ayan, and Nussinson (2006) that showed that the less effort exerted in task performance, the higher the judgement of learning.

Another recent study that confirms poorer calibration when reading on screen was published by Halamish and Elbaz (2019). This study makes an important contribution as it discusses the mode effect on children's comprehension and meta-comprehension judgements. In their meta-analysis, Delgado et al. (2018) did not

find age to be a moderator for the effect of medium on reading comprehension. Halamish and Elbaz (2019) suggest that this finding should be considered with caution as the number of studies on children included in the meta-analysis was small. However, it implies that children also tend to comprehend texts better on paper than on screen.

The study by Halamish and Elbaz (2019) gathered 38 fifth-grade children who read short texts on paper and screen. The students estimated their comprehension of each text and answered a reading comprehension test. The results showed that the children's comprehension was better when reading on paper compared to on the screen. Nevertheless, most children judged their comprehension to be the same on paper and screen, which suggests that they were metacognitively unaware of the effect of medium on their comprehension. Another study with 82 children of 11–12 years of age (Dahan Golan, Barzillai, & Katzir, 2018) also found that performance was better when reading on paper and that the children were more confident and better calibrated than when reading on screen. However, the majority of the children stated that they preferred to read on screens. This preference underpins the suggestion that children are unaware of the effect of medium on their comprehension. A recent study by Støle, Mangen, and Schwippert (2020) also found paper to be advantageous for children's reading comprehension. In this study, 1139 fifth-grade students participated, taking two comparable versions of a reading comprehension test, one on paper, and one digitally. Their results further showed that the negative effect of screen reading was evident for both boys and girls, but most profound among high-performing girls.

The same tendency is visible when looking at studies that concern adolescents. In 2017, Eyre et al. published a report on the digitization of the PAT: Reading Comprehension, a low-stakes, standardized assessment developed for use in New Zealand Schools, grades 4–10. Close to 200,000 assessment records were collected, and results showed that comprehension was lower when texts and items were presented on screen compared to when they were presented on paper. Mangen, Walgermo, and Brønnick (2013) also found that students who read texts in print showed significantly better comprehension than students who read on screen when exploring mode effect on 15-year-olds' reading of linear texts. In line with this are the findings from a study by Rasmussen (2015), who investigated differences in performance when 14-year-olds did the same reading test on paper and screen. The results showed a difference in favour of reading in print.

13.2.3 Gender Differences in Reading

The overarching values within the Norwegian education system include social justice, equity, equal opportunities to learn, inclusion and democratic participation for all students, regardless of their social and cultural background and abilities. All these ideas are interwoven within what is known as the Nordic model (Imsen, Blossing, & Moos, 2017). Results from PISA 2018 indicate that the Norwegian

educational system can be seen as equitable with respect to the socioeconomic status (SES) of the students. The influence of SES on achievement is significantly lower in Norway than the average across the OECD countries. Furthermore, only small differences related to students' performance are observed between schools (Jensen et al., 2019). However, the trend of girls outperforming boys on reading assessments is well known, and this trend applies to almost all countries that participate in large-scale assessments, such as PISA and Progress in International Reading Literacy Study (PIRLS) (Roe, 2013). In Norway, the phenomenon has been paid a great deal of attention, as the gender gap is significantly bigger than the OECD average (Jensen et al., 2019). Furthermore, this gap has been stable since the first PISA cycle in 2000, indicating that Norwegian schools fall behind on gender equity in reading.

Looking at the distributions of boys and girls across the levels of reading proficiency gives a more detailed picture of the gender gap. The proportion of Norwegian boys performing below level 2¹ was 26% in PISA 2018. For girls, this was 12%. Correspondingly, more girls were performing at the highest levels compared to boys (Jensen et al., 2019). The results from PIRLS show the same tendency. On the two lowest proficiency levels, the proportion of boys is almost 70%. Correspondingly, the proportion of girls is larger on the high comprehension levels. On the most advanced level, 64% are girls and 36% are boys (Solheim & Gourvennec, 2017).

Girls also outperform boys on the Norwegian National Assessment of Reading Literacy. Roe and Vagle (2012) found that open constructed-responses show a larger difference for boys and girls than multiple-choice items. This is partly due to boys skipping the open constructed-responses and partly due to short or wrong answers. This result was also confirmed when observing the performance of Norwegian students in PISA (Roe & Vagle, 2010). In their review of the national assessments, Roe and Vagle (2012) also found that the gender gap was larger for fictional texts than for factual texts. In particular, if the main character was female, the gender gap was twice as big as when the main character was a boy. This is in line with research showing that boys perform better on texts they like and find interesting than on texts they do not like (Oakhill & Petrides, 2007). Girls' performance, on the other hand, does not seem to be affected by motivational factors to the same extent, as their achievement is largely the same across all text types. Frønes (2016) suggests that this could be related to leisure time reading habits, with girls reading more diverse texts compared to boys.

Concerning reading habits, girls report that they read fictional literature more often than boys, who prefer reading newspapers both on paper and online. Results show that students who often read fictional literature demonstrate better reading comprehension than those who do not (Roe, 2020). Further, boys and girls express different engagement in reading. Girls spend significantly more time reading for pleasure than boys. In Norway, the results from PISA show that boys view reading

¹Level 2 is set by OECD as the baseline where students begin to demonstrate the competencies that will enable them to participate effectively and productively in life as continuing students, workers and citizens.

as a mere necessity and that they, to a larger degree than girls, only read if they have to. The same picture is portrayed across all participating OECD countries. However, Norwegian boys are among the least positive, a tendency that has persisted since the first PISA administration in 2000 (Roe, 2020).

In PISA 2009 and 2018, student's metacognitive reading strategies were also measured. The students were asked to rate the usefulness of different strategies proposed for different reading situations, and their answers were compared to the judgements of expert raters. The results show that Norwegian students score close to average when judging the strategies. However, the score difference between boys and girls was similar to that of the reading comprehension test; that is, favouring girls. The reason for boys' lower scores is that they do not distinguish between good and poorer strategies. Instead, they tend to rate all strategies as fairly good (Hopfenbeck & Roe, 2010; Jensen et al., 2019).

In terms of motivation, it is a common expectation that doing tests in a digital environment will benefit the performance of boys, because computers may motivate them more than paper and pencil tests do (Martin & Binkley, 2009). However, such an assumption should be tied to research on students' digital habits. In PISA 2009, the students reported on their use of computers (Frønes & Narvhus, 2011). The results showed that more than 70% of both boys and girls used computers daily or almost every day for chatting and for surfing on the Internet. For most other activities, such as homework and reading email, gender differences were small as well. Also, in PISA 2018, boys and girls report on their digital habits quite similarly (Roe, 2020). In PISA 2009, close to 50% of Norwegian boys, compared to less than 10% of the girls, reported that they used computers for gaming daily or close to daily. Updated numbers from The Norwegian Media Authority (2020) show that more girls are now interested in gaming, but the gender differences remain large. Ninety-six percent of the boys and 76% of the girls play games. In all age groups, the proportion among boys is larger than among girls, and among the girls, gaming becomes less widespread the older they get.

13.2.4 Digitization of Reading Assessments

Since 2004, The Norwegian National Assessment has been administered annually to students in grades 5 and 8 (10- and 13-year-olds). Students' skills in reading, mathematics and English as a second language are assessed. The tests provide information concerning individual students, student groups and schools, and are used both as an indicator for school improvement at a political level as well as the basis for formative assessments of students learning by teachers. The reading tests are, to a large extent, modelled like the international large-scale assessments PISA and PIRLS and share many similarities in terms of how the reading construct is defined and operationalized. The purpose of the tests is to measure students' reading literacy skills in terms of text comprehension as a basic skill (The Norwegian Directorate for Education and Training, 2017a). Thus, reading literacy is broadly defined as

being able to understand, use, reflect on and engage with texts. The definition is consistent with the definition used for the reading assessment in PISA: “Reading literacy is understanding, using, evaluating, reflecting on and engaging with texts in order to achieve one’s goals, to develop one’s knowledge and potential and to participate in society” (OECD, 2019, p. 28).

Following the digitization of PISA in 2015, The Norwegian National Assessment was administered on screen for the first time in 2016. The digitization of assessments has some advantages. In many cases, costs can be reduced; data collection is automatized and does not necessarily need to be supervised by researchers. Further, for some item types, scoring can be done by computers, which also eliminates error from manual scoring. Another advantage, pointed out by Støle et al. (2020), is the greater flexibility in text presentation when tests are computer-based, for instance, by using hyperlinks and dynamic elements. This allows for displaying texts that resemble the online texts children and adolescents meet in different types of electronic platforms. However, this also sheds light on some of the challenges with digitizing reading assessments. As new opportunities arise; consideration must be paid to ensure continuity with previous paper-based reading assessments. In addition, it is important to ensure that the change in the test conditions does not hinder students’ opportunities to succeed, regardless of the possible constraints related to some underlying factors (Espinoza, 2007).

In many cases, paper-based assessments have simply been replaced by digital assessments because mode equivalence has been assumed (Noyes & Garland, 2008). This could, however, be considered a break with traditional ways of categorizing reading activity and texts, as a distinction has often been made between paper-based and digital texts. The framework for PISA 2009 uses the terminology ‘print-medium texts’ and ‘electronic-medium texts’ (OECD, 2009). Print-medium texts have a static existence – the amount of text is immediately visible and the physical status of the text encourages the reader to approach the content in a certain order. Electronic-medium texts, on the other hand, are hypertext featuring navigation tools that make non-sequential reading possible, and often necessary. The reader chooses his or her reading path, and since the text is undefined and dynamic, it can be customized during the reading, often by the reader himself. On screen, only a fraction of the available text can be seen at any one time, and the extent of the text is unknown.

The distinction between print-medium and electronic-medium texts used in the framework for PISA 2009 (OECD, 2009) underlines the importance of medium for the categorization of texts. However, in the PISA 2015 framework, this distinction is no longer made due to the digitization of the test, meaning that the texts that were previously presented on paper were now delivered on screen. Although it is emphasized that the change of mode implies a break with previous assessments, it is argued that “both ‘print-medium’ and ‘electronic-medium’ texts can be consumed onscreen” (OECD, 2013, p. 15). Hence a new distinction is made between fixed and dynamic texts, moderating the link between text and medium (OECD, 2017). It is, however, a concern whether fixed texts are processed in the same way regardless of presentation mode. This concern pertains to the assumption that the medium itself

might set the premises for how texts are read. When processing dynamic texts, the use of navigation tools is essential for constructing meaning through the non-sequential reading path. Such navigation tools might be scrollbars, tabs, and various displays of hyperlinks. Although they have not been paid as much attention, navigation tools are also available when reading fixed texts, for instance, tables of contents, chapters, headlines and page numbers. Transferring fixed texts to the screen implies that the navigation tools of the print condition need to be accompanied by tools that are unique to the electronic medium. On this basis, one might question whether delivery mode can be disregarded when categorizing texts. Mangen and Kristiansen (2013) argue that texts read on screen, in essence, are volatile, dynamic and changeable, even if they are not multimodal or hypertext, but linear and in most ways look as if they are printed on paper. Even if the text is the same, the different affordances of the print and electronic media might affect the reading processing in different ways (Mangen, 2010).

In the framework for The Norwegian National Assessment, it is emphasized that the texts included in the test are meant to reflect the diversity of texts that the students typically encounter in the different subjects – not only verbal text but also illustrations, graphic representations, symbols and other possible ways of expression. Knowledge about different types of texts and text functions is therefore considered a crucial part of students' reading literacy skills (The Norwegian Directorate for Education and Training, 2017a). Regarding the digitization of the test in 2016, it is essential to point out that it resembles the way it was carried out for PISA in 2015. Despite the quite broad definition of text in the framework of The Norwegian National Assessment, computers are used for assessing fixed texts and not dynamic texts.

In the framework for The Norwegian National Assessment, it is further stated that the description of the tests might be revised as more results are obtained on how the digitized version of the test is working (The Norwegian Directorate for Education and Training, 2017a). On this notion, it is recognized that mode equivalence cannot easily be assumed. Moreover, it is crucial to obtain knowledge on how or if delivery mode affects students' reading comprehension and whether it affects everybody in the same way. From an equity perspective, one would want to assure that the change of mode is not disadvantageous to any particular group of students. Seen in the context of 'the equality-equity model' of Espinoza (2007), this can be linked to the output stage of the educational process and the importance of securing equity for equal achievement. When implementing new test conditions, it is essential for fairness that students who have achieved the same in the past continue to achieve similarly irrespective of the mode change. More specifically, if mode change is beneficial to some students and not to others, knowledge needs to be obtained so that fairness can be assured. This could, for instance, have implications for teacher practice, requiring change and customization of the reading instruction.

13.2.5 The Present Study

The present study presents results from the double mode assessment of reading comprehension, which was part of the preparations for changing the delivery mode of the Norwegian National Assessment in reading. The study uniquely contributes to the understanding of how delivery mode may affect the reading of Norwegian adolescents. An essential purpose of the study was to establish empirical evidence refuting or supporting the assumption of mode equivalence. Against the background of the research that has been shown so far, we can see that, overall, readers demonstrate better comprehension when reading on paper compared to when they read on screen or digitally (Clinton, 2019; Delgado et al., 2018; Kong et al., 2018; Singer & Alexander, 2017b). The first research question aims at further investigating this, based on the Norwegian context:

1. To what extent does overall comprehension performance differ when students process texts and solve items on paper and screen?

Further, in terms of equity, the change of mode should not be disadvantageous to any particular group of students (Espinoza, 2007). Research shows that girls outperform boys on reading tests (Jensen et al., 2019; Solheim & Gourvennec, 2017). However, little attention has been paid to see how delivery mode may affect gender differences, and the findings so far are inconclusive (Støle et al., 2020). The second research question, therefore, aims at exploring if the gender gap seen on paper-based reading assessments will translate to digitally delivered assessments:

2. Does change in delivery mode affect boys' and girls' results on reading comprehension tests in the same way?

13.3 Methods

13.3.1 Participants, Test Design and Administration

The study was administered in February 2016. Nine hundred seventy-three students from eighth grade (age 13–14) participated (48.7% female). The students came from nine different lower secondary schools. The schools were randomly picked from a list provided by the Norwegian Directorate for Education and Training and were spread geographically across the country. Both urban and rural schools were represented. The number of students from each school was distributed evenly. In conclusion, the number of participating schools makes the sample non-representative. However, the process has resulted in a sample of schools covering a relevant variation of contextual factors in Norway.

The study entailed two reading comprehension tests, Test 1 and Test 2. Each test consisted of seven texts that were similar concerning text length, text types and

formats. Both short and long texts were included, ranging in length from 228 to 1022 words. As the purpose of the assessment was to measure students' reading literacy skills in terms of text comprehension as a basic skill, the tests were designed from a wide selection of texts within different subjects (The Norwegian Directorate for Education and Training, 2017b). Both tests included expository texts, continuous and non-continuous, representing diverse subjects, such as history, natural science, social science and language arts. In each test, there was also one narrative text. To avoid gender effects as a result of text topic, the texts included in the tests were assumed to appeal to both boys and girls. The National Assessment typically contains 40 reading items. In all, 92 items were piloted. Test 1 contained 47 items (35 multiple-choice and 12 short-answer constructed-response items). Test 2 contained 45 items (only multiple-choice items). All multiple-choice items had four alternatives – one correct answer and three distractors. All items were scored dichotomously.

Both tests were administered digitally and on paper, and for each test, the screen version and paper version were made as close to identical as possible. The digital tests were completed on computers, and the students read from standard computer screens, typically 20 inches or a little smaller if using a laptop. Mouse and keyboard were used for navigation, selecting multiple-choice responses, and to type answers to constructed-response items. No training was provided in advance as most students were likely to previously have used the computers in classrooms or computer labs at the schools. Furthermore, no training was provided in using the digital platform, as it was familiar to students from the national assessments of mathematics and English that were digitized in 2014. The paper versions of the tests were formatted in A4 size and were handed out as booklets. Most texts filled about two pages, including tables, illustrations and graphics. Some of the texts were presented double-paged, while in other cases students had to turn a page. The comprehension items were displayed after each text, and students had to turn up to two pages to see the items connected to the text. Due to the length of most texts, students had to scroll when taking the digital version of the tests. After reading or scrolling through a text, items would appear on the left side in the platform window, while the text continued to be visible to the reader. In both conditions, students had the opportunity to move back and forth between texts and items, and they could revise their responses. The time limit of the test was 90 min.

All students conducted the tests at school administered by their teachers. The teachers had been told to give instructions according to the guidelines provided by the researchers. As the national assessments are administered annually for the full cohort of fifth-, eighth- and ninth-grade students, it is reasonable to assume that many of the teachers would be experienced test administrators. However, the digitized version was new for the reading assessment. The study had a between-participant design, and each student was assigned to one of the two tests, taking it on either paper or screen. The students were assigned randomly to the two tests; 470 students completed Test 1, and 503 students completed Test 2. However, for delivery mode, the students were assigned class-wise, as we wanted to avoid a design too sophisticated for the teachers to handle. In support of this decision is the argument that randomness was ensured by randomly assigning students to the two different

tests. Further, we know that Norwegian classrooms tend to be quite heterogeneous considering that students with different home background and abilities are mixed. After completing the tests, data from the digital version was generated automatically, and the paper booklets were returned to the researchers, who scored the responses. For all short-answer constructed responses, both digital and on paper, at least two experienced raters scored each response to secure inter-rater reliability.

13.3.2 Data Analysis

The data collected from the study provided the following information for each student: school, mode, gender and score (item format and frequency on multiple-choice items). Probabilistic test theory was employed to give a measure for student achievement. To be more specific, the Rasch model was applied, which allows for characterization of students' proficiency and difficulty of items as locations on the same continuous scale. The origin of the scale is identified by the mean item difficulty. Students' proficiency corresponds to the point on the scale where they will have a 50% probability of responding correctly to an item. Given that the two tests were unique and non-linked test forms, the scaling was done separately for each of the two test forms. However, the two versions of each test (paper and on-screen) were calibrated together. The software package RUMM2030 was used for the scaling, while statistical analyses were conducted in SPSS 26.

Gender differences across the modes were investigated through multiple regression analysis, using dummy-coded variables for mode (1 = screen) and gender (1 = girl) as predictors. Also, an interaction term for mode and gender was included in the model as the product of the variables for mode and screen. As exemplified by Aiken and West (1991), this allows for the exploration of conditions under which causal relationships are moderated or strengthened. In the case of this study, the interaction term makes it possible to see if the effect of the mode change is the same for boys and girls.

Several assumptions (i.e., homoscedasticity, normality and independence of residuals, multicollinearity, as well as variables tolerance) were checked to ensure that the regression model fit the data (Cohen, Cohen, West, & Aiken, 2003). All assumptions were met. Furthermore, as extreme scores potentially have a great impact on regression models (Osborne & Overbay, 2004), the dataset was screened to detect outliers. As a result, 3 and 9 students were removed from the dataset in Test 1 and 2, respectively.

13.4 Results

The first research question guiding this study focused on the role of the medium in students’ overall reading comprehension when processing texts and solving items in print and on screen. Descriptive data on the two comprehension tests overall and by medium are displayed in Table 13.1. The data show that students scored a little lower on both tests when they were administered digitally. However, the differences between the mean scores of the two modes are low. On average, students score 0.05 higher on Test 1 and 0.07 higher on Test 2 when reading on paper compared to screen. This difference by medium on overall comprehension is not significant, and the prediction of students having higher comprehension scores when reading on paper is not confirmed.

As a first insight into the second research question for this study – whether the change in delivery mode affects boys’ and girls’ results on reading comprehension tests in the same way – descriptive data for reading comprehension on both tests, split on medium and gender, are provided in Table 13.2. On both tests, girls performed better when the test was administered on screen as compared to on paper. For Test 1 the mean score for girls’ comprehension was .53 (SD = 1.06) on screen and .42 (SD = .97) on paper, a mean difference of .11 between modes. On Test 2, the mean difference for girls’ comprehension was .03, favouring the digital condition. Boys, on the other hand, perform better on paper than screen, the mean difference being .25 on Test 1 and .17 on Test 2. The difference in mean scores between modes is larger for boys than it is for girls.

Table 13.1 Means and standard deviations for reading comprehension by medium on Test 1 and Test 2

	Test 1			Test 2		
	<i>N</i>	<i>M</i>	SD	<i>N</i>	<i>M</i>	SD
Paper	247	.31	1.02	250	.28	.95
Screen	223	.26	1.10	253	.21	.93
Total	470	.28	1.06	503	.25	.94

Table 13.2 Means and standard deviations for reading comprehension by medium and gender

Test	Mode	Gender	<i>N</i>	<i>M</i>	SD	<i>p</i> -Value
Test 1	Paper	Boy	131	.22	1.06	.139
		Girl	110	.42	.97	
	Screen	Boy	109	-.03	1.07	.000
		Girl	114	.53	1.06	
Test 2	Paper	Boy	126	.24	.98	.421
		Girl	115	.34	.94	
	Screen	Boy	133	.07	.91	.010
		Girl	120	.37	.94	

Note: *p*-Value from t-test for individual samples, comparing scores of boys and girls by mode on the two tests. *p*-value is considered statistically significant at *p* < 0.05

Table 13.3 Regression model with reading comprehension scores on Test 1 as an outcome (Outliers removed, $N = 3$)

Model	Variable	R^2	R^2 change	F change	b	p -Value
1		.032	.032	7.569		.001
	Mode (0 = paper, 1 = screen)				−.091	.330
	Gender (0 = boy, 1 = girl)				.357	.000
2		.041	.009	4.375		.037
	(Constant)				.247	.005
	Mode (0 = paper, 1 = screen)				−.281	.032
	Gender (0 = boy, 1 = girl)				.169	.193
	Interaction mode-gender				.391	.037

Note: p -Value is considered statistically significant at $p < 0.05$

Table 13.4 Regression model with reading comprehension scores on Test 2 as an outcome (Outliers removed, $N = 9$)

Model	Variable	R^2	R^2 change	F change	b	p -Value
1		.014	.014	3.359		.036
	Mode (0 = paper, 1 = screen)				−.053	.499
	Gender (0 = boy, 1 = girl)				.194	.013
2		.023	.009	4.464		.035
	(Constant)				.260	.001
	Mode (0 = paper, 1 = screen)				−.208	.052
	Gender (0 = boy, 1 = girl)				.026	.815
	Interaction mode-gender				.328	.035

Note: p -Value is considered statistically significant at $p < 0.05$

Even if the girls performed better than the boys overall, when breaking down the scores by test form, medium and gender, as shown in Table 13.2, it is evident that girls outperformed boys in the on-screen condition, with a difference of .56 and .44 for the two tests, respectively. As shown by the column to the far right in Table 13.2, listing the p -values from a t-test for individual samples, comparing scores of boys and girls by mode on the two tests, both differences are statistically significant ($p = .000$ and $p = .010$, respectively). However, the gender differences for the paper-based tests are trivial and non-significant. Before turning to the regression analysis, it is worth noting that these results indicate the existence of an interaction effect.

Tables 13.3 and 13.4 show the results of the regression analysis for Test 1 and Test 2, respectively, and overall, the results show the same tendencies for the two tests. The values of R^2 for the steps in both analyses show that very little variance in the criterion variable (the score) is explained by the models. This amounts to about 3–4% for Test 1 and about 1.5–2.5% for Test 2. It is, however, significant at the .05 level (Test 1, $p = .001$, Test 2, $p = .036$). Explanatory power was improved in the second model, as can be seen from the positive change in R^2 , the change being significant in both cases (Test 1, $p = .037$, Test 2, $p = .035$).

In line with the results from the descriptive analysis, the first model of the regression analysis shows no significant difference for mode, neither for Test 1 nor Test 2. However, there is a strong and significant difference for gender (Test 1, $b = .357$, p

= .000, Test 2, $b = .194, p = .013$), most prominent in Test 1. The fact that R^2 is low indicates that the variance in reading comprehension is much larger among each gender, respectively, compared to the difference between boys and girls. Turning to the second model of the regression analysis, with the interaction term included, this also confirms the descriptive analysis, as the interaction effect is significant on both tests (Test 1, $b = .391, p = .037$, Test 2, $b = .328, p = .035$). Further, with the interaction term included, the gender difference is no longer significant on any of the tests. However, mode turns out to have a significant effect on reading comprehension in the second model for Test 1($b = -.281, p = .032$), and it is close to significant for Test 2 ($b = -.208, p = .052$).

As a further illustration of the interaction effect documented by both the descriptive analyses and the regression models, Figs. 13.1 and 13.2 show predicted values for boys and girls across modes for Test 1 and Test 2, respectively. The predicted values are calculated from the regression coefficients by using the equation:

$$\hat{Y} = b_1X_1 + b_2X_2 + b_{12}X_1X_2 + b_0$$

In the equation, X_1 pertains to ‘Mode’ and X_2 to ‘Gender’. Given that we, for example, want to know the predicted scores on Test 1 for boys reading on screen, the calculation will be: $(1*(-.281)) + (0*.169) + (0*.391) + (.247) = -.034$. The same is done for the other conditions; the results are given in the plots. The fact that the lines for the two groups go in separate directions illustrates very well the interaction effect with a widening of the gender gap from paper to screen.

Fig. 13.1 Predicted values for Test 1, model 2

Fig. 13.2 Predicted values for Test 2, model 2

13.5 Discussion

This study was motivated by recent trends in the field of large-scale assessments, as paper-based reading tests are being replaced with digitally delivered, on-screen assessments. The preparation of digitizing the Norwegian National Assessment in reading in 2016 offered a unique opportunity to perform a mode effect study among adolescents. In particular, two questions were addressed: first, to what extent overall comprehension performance differs when students process texts and solve items on paper and screen, and second, if change in delivery mode affects boys' and girls' results on reading comprehension tests in the same way. Investigating these questions is relevant for understanding how delivery mode may affect students' reading and, in turn, how changes in test conditions may have implications for fairness in student assessment. From an equity perspective, it is important that students have the same opportunity to succeed as they have had in the past (Espinoza, 2007).

The results of this study did not reveal significant differences in overall reading performance among 13–14-year-olds as an effect of delivery mode. This is contrary to what could be expected, reviewing the literature in the field (Clinton, 2019; Delgado et al., 2018; Kong et al., 2018; Singer & Alexander, 2017b). At the same time, 13–14-year-olds are often labelled as digital natives (Prensky, 2001), and most of them are likely to possess extensive digital skills and experience. Within the Norwegian educational policy, children's digital skills have been prioritized (The Norwegian Directorate for Education and Training, 2017b), and computers and tablets are widely used as learning tools in primary and secondary schools. Norwegian children and adolescents also have experience with digital devices at home. Ninety-nine percent of 17–18-year-olds have their own mobile phone, and more than 98% have their own computer (The Norwegian Media Authority, 2020). Also, younger students have wide access to digital devices. In PIRLS 2016, Norway ranked highest concerning children's access to digital devices; high access was reported for 58% of the children, whereas low access was not reported (Mullis, Martin, Foy, & Hooper, 2017).

Furthermore, when preparing the test for each mode, efforts were made to secure a low-threshold digital solution. In order to keep the digitized test in line with the previous paper-based version, it was important to make the screen and print version as similar as possible. Equal attention was paid to ensuring that the technical requirements of using computers were not higher than what could be expected for the age group. More specifically, the students had to use the mouse and keyboard for responding to items and for navigation, the navigation tools mainly being scrollbars for displaying longer texts and tabs for moving back and forth between texts. Considering the navigation skills that were anticipated among these students, the requirements were not expected to be too challenging.

Turning to the second question addressed by this study, the reviewed literature did not propose a clear hypothesized answer, as little research has been done on mode effect and gender differences (Clinton, 2019; Delgado et al., 2018; Kong et al., 2018). The results showed a widening of the gender gap, with boys clearly not benefitting from completing the tests on screen. This is a matter of concern for educational justice, as possible constraints related to underlying factors should not hinder students' opportunity to succeed (Espinoza, 2007). Knowing that research indicates that Norwegian schools fall behind on gender equity in reading (Jensen et al., 2019) makes it particularly important to further understand what might affect the gender gap to increase when changing the test conditions. This could be of guidance to policy makers and teachers.

Computers have been assumed to motivate boys, and it is a common expectation that boys will benefit from tests being digitized (Martin & Binkley, 2009). The results of this study show that this assumption might not hold. The gender gap increased in the on-screen version of both tests, boys' comprehension scores being negatively affected by the screen condition. Several factors may have contributed to the results. First, it is uncertain to what degree boys' motivation for using computers is transferrable to completing reading comprehension tests on screen. As can be seen from the report on children and media (The Norwegian Media Authority, 2020), for instance, boys are motivated to use computers for gaming. Whether this would translate into motivation for digitized reading assessments is not clear. However, as children and adolescents today are considered to be digital natives, the use of computers as such has likely been de-mystified.

The use of screens for reading, both in and out of school, steadily increases. The activities children and adolescents most frequently use computers and digital devices for at home are watching video clips and listening to music, visiting social network profiles, socializing and communication, playing games, and searching for information to satisfy curiosity (Mascheroni & Cuman, 2014). Both boys and girls also report that they use computers for activities which, to a greater extent, are related to reading in particular, such as chatting, surfing on the Internet, searching for information and doing homework (Frønes & Narvhus, 2011; Roe, 2020). Most of these texts that are encountered on screen share the features of being dynamic, undefined and interactive. Considering the 'shallowing hypothesis' (Annisette & Lafreniere, 2017; Carr, 2010), digital texts are often processed in a shallow or superficial way, as digital texts may promote a way of reading that typically involves

skimming and scanning. In turn, this could contribute to some children and adolescents developing a screen reading behaviour that is not beneficial for deep reading and processing of longer texts. If the student's screen reading is modelled on strategies efficient for quick and superficial reading, this might explain why the scores of students taking the tests on screen were poorer than the scores of those who took the tests on paper. However, as boys and girls do not report very differently about their digital habits, further explanations are needed to understand why boys' comprehension scores on the reading tests are more negatively affected than those of girls.

One possible explanation may relate to metacognitive comprehension. Several studies (e.g. Ackerman & Goldsmith, 2011; Dahan Golan et al., 2018; Halamish & Elbaz, 2019; Singer & Alexander, 2017a; Singer Trakhman et al., 2017) show that readers are poorly calibrated when asked to judge the medium in which they perform best. This may imply that they are unaware of whether they comprehend better when reading from paper or screen. Many presume that they are better at reading in the digital medium, but, in reality, they comprehend better when reading on paper. The miscalibration is likely to be underpinned by the reading activity itself. As digital reading is perceived to be easy and fast, the reader's sense of achievement is likely to rise, even if this is not the case. For this reason, awareness of expedient reading strategies seems even more important when reading on screen. Considering that boys generally demonstrate lower metacognitive skills in reading, as shown by the Norwegian PISA results (Hopfenbeck & Roe, 2010; Jensen et al., 2019), this may have had a negative effect on their scores on the screen version of the tests. However, as collection of data on students' calibration was not within the scope of this study, such an explanation may not be ascertained. A future study should address more specifically the metacognitive comprehension of boys and girls reading across different modes.

Another possible factor that could have contributed to the widening of the gender gap from paper to screen could be that girls' reading habits are also beneficial for on-screen reading. Several studies confirm that reading traditional extended texts, especially fictional books, is a strong predictor of reading comprehension, even if many reading activities are digitized (Duncan, McGeown, Griffiths, Stothard, & Dobai, 2016; Pfost, Dörfler, & Artelt, 2013). This is in line with the Norwegian PISA results as well, showing that students who report that they read for enjoyment comprehend significantly better than those who do not read. Furthermore, students who report that they prefer reading books on paper outperform students who read books more often on digital devices and students who read books equally often in paper format and on digital devices (Roe, 2020). This indicates that book reading and reading for enjoyment have a positive effect on reading comprehension regardless of presentation mode. As girls spend significantly more time reading for pleasure than boys, who for a large part report that reading is seen as a mere necessity, girls are more likely to develop reading skills that are beneficial to reading across all text presentation media.

Results of several large-scale assessments administered to Norwegian children and adolescents consistently show that the proportion of boys on the lowest comprehension levels is significantly larger than the proportion of girls (Jensen et al., 2019;

Solheim & Gourvennec, 2017). This finding should also be considered when trying to understand the widening of the gender gap in on-screen testing. Research shows that reading from a screen involves a higher cognitive workload and can be more tiring than reading from paper (Wästlund et al., 2005). Especially low-performing students experience a higher workload when reading from a screen (Noyes et al., 2004). Hence, they may be additionally disadvantaged when completing computer-based assessments as compared to similar tasks on paper.

The higher cognitive load associated with the screen condition may be especially true for assessments that involve sophisticated tasks that require sustained attention (Eyre, Berg, Mazengarb, & Lawes, 2017). Further, it may also be related to issues of navigation. Bridgeman, Lennon, and Jackenthal (2003) suggest that the resolution of the monitor and amount of scrolling required by the test-taker could affect performance. They found that students who could see the whole passage of a text without scrolling comprehended better on reading assessments than those who had to scroll to see the full passage. This is in line with the result of Delgado et al. (2018), showing that the advantage of paper-based reading is significant when scrolling is necessary to read texts on screen.

As pointed out by Sanchez and Wiley (2009), scrolling is likely to draw on the limited capacity of the working memory needed for reading. Furthermore, Kingston (2008) argues that reading while scrolling is cognitively different from reading a page. While reading a page, the reader can use spatial memory clues to remember the location (e.g. toward the upper right portion of a page) of information that is pertinent – for instance, when answering a particular question. Parallel clues are not available when scrolling is needed for reading texts on screen. Scrolling constantly changes the spatial frame of reference, which may have a negative effect on the readers' mental reconstruction of the text. By implication, this also has a negative effect on comprehension, as having a good spatial mental representation of the physical layout of a text supports comprehension. Cataldo and Oakhill (2000) found that good comprehenders were more efficient than poor comprehenders at remembering and relocating the order of information in texts. This suggests that there is a relationship between mental reconstruction of text structure and reading comprehension.

The present study did not control for factors that seem to increase the demands of reading on screen, such as the resolution of the monitor and amount of scrolling. Consequently, the extra cognitive load of creating mental representations of texts the spatial frames of which constantly changed may have contributed to the fact that low-performing students participating in this study comprehended worse in the screen condition compared to the paper condition. As the proportion of low-performing students is higher among the boys than among girls, this may have further contributed to the increase of the gender gap from paper to screen. However, a future study should explore this assumption more closely, as comprehension differences across modes for high- and low-performing students were not in the scope of this study.

13.6 Concluding Remarks

As the empirical evidence of children's and adolescent's reading comprehension on paper compared to on screen remains rather sparse, this study adds valuable information about the way delivery mode affects reading comprehension. The study particularly broadens the field by exploring how the gender gap seen in reading is affected. Both the large sample size and the fact that both reading tests showed the same results are strengths of the study. The results of the study have several pedagogical implications. Showing that the gender gap increases when reading on screen, the results confirm that equivalence cannot easily be assumed. This has implications for policy makers, as consideration should be paid to the increasing use of digital technologies in education. Furthermore, care must be taken to ensure the fairness of student assessment. Even though the transition from paper to screen is the same for all students, this study exemplifies how equality in some cases does not contribute to equity (Espinoza, 2007). However, awareness of this matter may promote measures that can be levelling in an educational system aiming to be a 'School for All'.

Although no differences in overall reading performance among 13–14-year-olds as an effect of the delivery mode were found, the results of the present study indicate that different media may affect the reading of students differently. Moreover, students are likely to exhibit different reading behaviour and apply diverse strategies for different reading purposes. However, attention must be paid to what reading behaviour is useful. It is evident that the skimming and scanning strategies readily applied for online information-seeking and entertainment do not benefit all reading situations on screen. On the contrary, in-depth reading strategies are more beneficial for completing digitized reading assessments.

Garland and Noyes (2004) point out that repeated exposure and rehearsal of computer-based information is needed to equate knowledge retrieval with that achievable from paper, and research by Lauterman and Ackerman (2014) shows that encouragement of in-depth processing on screen may reduce the inferiority of screen reading. This has implications for teachers and educators. Children and adolescents need to develop awareness of useful reading behaviour and should be taught effective and expedient strategies for reading on screens. This may contribute to the overall fairness in the assessment situation and eliminate some of the adverse effects of the screen for students susceptible to these. Moreover, both boys and girls, as well as both high- and low-performing students, would benefit from this.

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, 1816–1828.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA and London: SAGE Publications, Inc..
- Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction*, 24, 1–3.
- Annisette, L. E., & Lafreniere, K. D. (2017). Social media, texting, and personality: A test of the shallowing hypothesis. *Personality and Individual Differences*, 115, 154–158.
- Bridgeman, B., Lennon, M., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer- based test performance. *Applied Measurement in Education*, 26, 191–205.
- Carr, N. G. (2010). *The shallows: What the internet is doing to our brains*. New York: W. W. Norton.
- Cataldo, M. G., & Oakhill, J. (2000). Why are poor comprehenders inefficient searchers? An investigation into the effects of text. *Journal of Educational Psychology*, 92(4), 791–799.
- 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., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Conway, M. A., Gardiner, J. M., Perfect, T. J., Anderson, S. J., & Cohen, G. (1997). Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General*, 126(4), 393–413.
- Dahan Golan, 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.
- 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.
- Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics*, 35, 1297–1326.
- Duncan, L. G., McGeown, S. P., Griffiths, Y. M., Stothard, S. E., & Dobai, A. (2016). Adolescent reading skills and engagement with digital literacies as predictors of reading comprehension. *British Journal of Psychology*, 107, 209–238.
- Espinoza, O. (2007). Solving the equity-equality conceptual dilemma: A new model for analysis of the educational process. *Educational Research*, 49(4), 343–363.
- Eyre, J., Berg, M., Mazengarb, J., & Lawes, E. (2017). *Mode equivalency in PAT: Reading comprehension*. Wellington, NZ: New Zealand Council of Educational Research.
- Frønes, T. S. (2016). Resultater i lesing [Results in reading]. In M. Kjærnsli & F. Jensen (Eds.), *Stø kurs. Norske elevers kompetanse i naturfag, matematikk og lesing I PISA 2015* (pp. 136–171). Oslo, Norway: Universitetsforlaget.
- Frønes, T. S. & Narvhus, E. K. (2011). *Elever på nett. Digital lesing i PISA 2009* [Students online. Digital reading in PISA 2009]. Oslo, Norway: ILS, UiO.
- Garland, K. J., & Noyes, J. (2004). CRT monitors: Do they interfere with learning? *Behavior and Information Technology*, 23(1), 43–52.
- Halamish, V., & Elbaz, E. (2019). Children's reading comprehension and metacomprehension on screen versus paper. *Computers & Education*, 145, 103737.
- Hopfenbeck, T. N., & Roe, A. (2010). Lese- og læringsstrategier [Reading and learning strategies]. In M. Kjærnsli & A. Roe (Eds.), *På rett spor. Norske elevers kompetanse i lesing, matematikk og naturfag I PISA 2009* (pp. 118–137). Oslo, Norway: Universitetsforlaget.

- Imsen, G., Blossing, U., & Moos, L. (2017). Reshaping the Nordic education model in an era of efficiency. Changes in the comprehensive school project in Denmark, Norway, and Sweden since the millennium. *Scandinavian Journal of Educational Research*, 61(5), 568–583.
- Jensen, F., Pettersen, A., Frønes, T.S., Kjærnsli, M., Rohatgi, A., & Narvhus, E. K. (2019). *PISA 2018. Norske elevers kompetanse i lesing, matematikk og naturfag* [PISA 2018. What Norwegian students know in reading, mathematics and science]. Oslo, Norway: Universitetsforlaget.
- Kingston, N. M. (2008). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education*, 22(1), 22–37.
- Kong, Y., Seo, Y., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, 123, 138–149.
- Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationship between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, 135(1), 36–69.
- Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior*, 35, 455–463.
- Mangen, A. (2010). Lesing – på skjerm eller papir; er det så nøye, da? [Reading – on screen or paper; does it really matter?]. *Norsklæreren*, 3–10, 16–20.
- Mangen, A., & Kristiansen, M. (2013). Tekstlesing på skjerm: Noen implikasjoner av et digitalt grensesnitt for lesing og forståelse [Reading texts on screen: Implications of a digital interface for reading and comprehension]. *Norsk Pedagogisk Tidsskrift*, 97, 52–62.
- Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, 58, 61–68.
- 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 and implications for large-scale testing* (pp. 75–82). Luxembourg, Luxembourg: OPOCE.
- Mascheroni, G., & Cuman, A. (2014). *Net children go mobile: Final report. Deliverables D6.4 & D5.2*. Milano: Educatt.
- Mayes, D. K., Sims, V. K., & Koonce, J. M. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics*, 28, 367–378.
- 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: <https://timssandpirls.bc.edu/pirls2016/international-results/>
- Noyes, J., & Garland, K. J. (2003). VDT versus paper-based text: Reply to Mayes, Sims and Koonce. *International Journal of Industrial Ergonomics*, 31, 411–423.
- Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, 51(9), 1352–1375.
- Noyes, J., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: Is workload another test mode effect? *British Journal of Educational Technology*, 35, 111–113.
- 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.
- OECD. (2009). *PISA 2009 assessment framework. Key competencies in reading, mathematics and science*. Retrieved from <https://www.oecd.org/pisa/pisaproducts/44455820.pdf>
- OECD. (2013). *PISA 2015 draft reading literacy framework*. Retrieved from <https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Reading%20Framework%20.pdf>
- OECD. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematics, financial literacy and collaborative problem solving, revised edition*. Paris: PISA, OECD Publishing.
- OECD. (2019). *PISA 2018 reading framework*. In *PISA 2018 assessment and analytical framework*. Paris: PISA, OECD Publishing.
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(6), 1–12.
- Pfost, M., Dörfler, T., & Artelt, C. (2013). Students' extracurricular reading behavior and the development of vocabulary and reading comprehension. *Learning and Individual Differences*, 26, 89–102.

- Prensky, M. (2001). Digital natives, digital immigrants. Part 1. *On the Horizon*, 9(5), 1–6.
- Rasmussen, M. (2015). Reading paper – Reading screen. *Nordic Studies in Education*, 35, 3–19.
- Roe, A. (2013). Norske gutters resultater på nasjonale og internasjonale leseprøver [Norwegian boys' results on national and international reading assessments]. In K. Kverndokken (Ed.), *Gutter og lesing* (pp. 13–31). Bergen, Norway: Fagbokforlaget.
- Roe, A. (2020). Elevenes lesevaner og holdninger til lesing [Students' reading habits and attitudes towards reading]. In T. S. Frønes & F. Jensen (Eds.), *Like muligheter til god leseforståelse? 20 år med lesing i PISA* (pp. 107–134). Oslo, Norway: Universitetsforlaget.
- Roe, A., & Vagle, W. (2010). Resultater i lesing [Results in reading]. In M. Kjærnsli & A. Roe (Eds.), *På rett spor: Norske elevers kompetanse i lesing, matematikk og naturfag I PISA 2009* (pp. 59–93). Oslo, Norway: Universitetsforlaget.
- Roe, A., & Vagle, W. (2012). Kjønnforskjeller i lesing – et dybdeddykk i nasjonale prøver på åttende trinn fra 2007 til 2011 [Gender differences in reading – A closer look at the national assessments for grade eight from 2007 to 2011]. *Norsk Pedagogisk Tidsskrift*, 96, 425–440.
- 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.
- Shaikh, D. (2004). Paper or pixels: What are people reading online? *Usability News*, 6.
- Singer, L. M., & Alexander, P. A. (2017a). Reading across mediums: Effects of reading print and digital texts on comprehension and calibration. *Journal of Experimental Education*, 85, 155–172.
- Singer, L. M., & Alexander, P. A. (2017b). Reading on paper and digitally: What the past decades of empirical research reveal. *Review of Educational Research*, 87, 1007–1041.
- Singer Trakhman, L. M., Alexander, P. A., & Berkowitz, L. E. (2017). Effects of processing time on comprehension and calibration in print and digital mediums. *The Journal of Experimental Education*, 87, 1–15.
- Solheim, O. J., & Gourvennec, A. F. (2017). Gutte- og jentelesere i topp og bunn. En utforsking av leserprofiler [High and low performing readers among boys and girls. Examining reading profiles]. In E. Gabrielsen (Ed.), *Klar framgang! Leseferdighet på 4. og 5. trinn i et femtenårsperspektiv* (pp. 186–203). Oslo, Norway: Universitetsforlaget.
- Spencer, C. (2006). Research on learners' preferences for reading from a printed text or from a computer screen. *Journal of Distance Education*, 21, 33–50.
- Støle, H., Mangen, A., & Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: A mode-effect study. *Computers & Education*, 151, 1–13.
- The Norwegian Directorate for Education and Training. (2017a). *Rammeverk for nasjonale prøver* [Framework for the national assessments]. Retrieved from <https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover/>
- The Norwegian Directorate for Education and Training. (2017b). *Rammeverk for grunnleggende ferdigheter* [Framework for basic skills]. Retrieved from <https://www.udir.no/laring-og-trivsel/lareplanverket/grunnleggende-ferdigheter/rammeverk-for-grunnleggende-ferdigheter/>
- The Norwegian Media Authority. (2020). *Barn og medier 2020* [Children and Media 2020]. Retrieved from <https://medietilsynet.no/globalassets/publikasjoner/barn-og-medier-undersokelser/2020/200402-delrapport-3-gaming-og-pengebruk-i-dataspill-barn-og-medier-2020.pdf>
- Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26, 1–12.
- Wästlund, E., Reinikka, H., Norlander, T., & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior*, 21, 377–394.
- Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41(1), 67–85.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.