*Nordic Journal of Literacy Research*

**Peer-reviewed article**

# Gender differences in Norwegian PIRLS 2016 and ePIRLS 2016 results at test mode, text and item format level

Katrin Schulz-Heidorf* and Hildegunn Støle
*University of Hamburg, Norwegian Reading Centre, University of Stavanger*

**Abstract**
Gender differences in reading are a common finding in international assessments with girls usually outperforming boys. This article investigates such gender differences by looking at test modes (paper-based versus digital assessments), reading purpose (literary versus informational), text features (associations between reading scores and how much students like a text) and item format characteristics (multiple choice versus constructed response items). All analyses are based on data of Norwegian fifth-grade students (n = 3610) from the most recent cycle of the Progress in International Reading Literacy Survey (PIRLS and ePIRLS) 2016. The results point towards a general mode effect between the paper-based and digital assessment for constructed response items. This effect seems to be less strong in boys, indicating that boys may be motivated to type responses on a keyboard as opposed to writing with a pen on paper. For text features, we found that boys might be disengaged from reading when the text shows female characteristics such as a female protagonist, leading to boys' lack of interest and, subsequently, to lower scores. The results are discussed in the light of the test design of PIRLS and ePIRLS.

**Keywords:** *Digital assessment; paper-based assessment; reading achievement; mode effect; school*

Received: May, 2018; Accepted: December, 2018; Published: December, 2018

## Introduction

The Progress in International Reading Literacy Study (henceforth PIRLS) has measured reading comprehension among 10-year old students in a number of countries since 2001, giving the participating nations insights into how well national reading results compare to achievement in other countries, and to trace any changes across time. Norway has taken part in all PIRLS cycles (i.e. in 2001, 2006, 2011 and 2016). These have all been paper-based, but in 2016 the traditional PIRLS-test was complemented by an additional test of "online informational reading", the electronic PIRLS (henceforth ePIRLS). The ePIRLS assessment gives us the opportunity to examine

*Correspondence to: Katrin Schulz-Heidorf, University of Hamburg, Faculty of Pedagogy, Von-Melle-Park 8, 20146 Hamburg, Germany. Email: katrin.schulz-heidorf@uni-hamburg.de


*Citation: K. Schulz-Heidorf & H. Støle. "Gender differences in Norwegian PIRLS 2016 and ePIRLS 2016 results at test mode, text and item format level" Nordic Journal of Literacy Research, Vol. 4(1), 2018, pp. 167–183. http://dx.doi.org/10.23865/njlr.v4.1270*

a number of questions about the role of reading modality (computer), such as its relation to gender. In most countries, the paper-based PIRLS assessments have documented gender differences in favour of girls in all three of the first cycles (Mullis, Martin, Foy, & Drucker, 2012). Also, PIRLS 2016 showed a gender gap in favour of girls in 48 of the 50 participating countries; the remaining two countries had no significant gender gap in either direction (Mullis, Martin, Foy & Hooper, 2017b). In none of the countries did boys outperform girls in the paper-based PIRLS 2016 assessment.

Fourteen countries participated in ePIRLS, and even though girls outperformed boys in many of these, it turned out that three countries had no significant differences. Interestingly, Danish students exhibited a significant gender gap in the paper-based PIRLS 2016, while having no such difference between boys and girls in ePIRLS. Table 1 is an exhibit taken from the international report on ePIRLS (Mullis, Martin, Foy & Hooper, 2017a) with countries ranked by their size of gender differences in online informational reading achievement.

*Table 1.* Girls and boys results in ePIRLS ranked by gender differences. The table includes mean score, standard error of measurement (SE) and significant differences.

| ePIRLS countries | Girls' results | SE | Boys' results | SE | Gender difference |
|---|---|---|---|---|---|
| Italy | 534 | 2.6 | 531 | 2.4 | n.s. |
| Portugal | 524 | 2.6 | 521 | 2.6 | n.s. |
| Denmark | 560 | 2.9 | 556 | 2.9 | n.s. |
| USA | 560 | 2.8 | 554 | 3.1 | 6 |
| Canada | 547 | 3.7 | 539 | 3.7 | 8 |
| Chinese Taipei | 551 | 2.3 | 541 | 2.2 | 9 |
| Ireland | 572 | 2.8 | 561 | 3.4 | 11 |
| Israel | 542 | 2.5 | 530 | 3.1 | 11 |
| Slovenia | 532 | 2.5 | 518 | 2.5 | 14 |
| Sweden | 567 | 2.6 | 552 | 2.7 | 15 |
| Georgia | 485 | 3.2 | 469 | 3.8 | 15 |
| **Norway** | **576** | **2.6** | **558** | **2.9** | **18** |
| Singapore | 599 | 3.2 | 578 | 3.3 | 21 |
| United Arab Emirates | 483 | 3.4 | 454 | 4.1 | 29 |
| International mean | 545 | 0.8 | 533 | 0.8 | 12 |

*Note: The abbreviation n.s. means that the gender difference is non-significant.*

*Source: Mullis, Martin, Foy & Hooper, 2017a, p. 84.*

We have marked the Norwegian student sample in bold in Table 1 as this population will be the focus of the present article. The mean score of Norwegian students is high and significantly above the international mean (545 point score; for international results see Mullis et al., 2017a), for both girls (576 points) and boys (558 points), but the gender difference at 18 points is among the largest of all countries. Only Singapore and the United Arab Emirates have larger gender differences. Among the Nordic countries, Sweden also has a large gender gap at 15 points in favour of girls, while there is no significant difference in Denmark in ePIRLS. As for the paper-based

PIRLS 2016, all three Nordic countries as well as Finland (who did not participate in ePIRLS) experience significant gender differences in reading (Gabrielsen & Hovig, 2017). In other words, gender gaps in reading achievement fluctuate in culturally similar countries like Norway, Sweden and Denmark in different test modalities. In the present article, we explore some factors suggested to contribute to the differences in boys' and girls' achievement in reading assessments.

## Theory

Several factors may serve to explain why we find varying gender differences when assessing reading literacy in different settings. For example, it is suggested that males prefer working on computers to reading from print and that paper-based reading assessment discriminates against boys (Martin & Binkley, 2009). Further, it has been found that boys perform better on informational reading than on literary reading (Mullis et al., 2012) and engage differently in a text, depending on their interest in the topic (Oakhill & Petrides, 2007). Also, item formats (e.g. multiple choice versus constructed response) seem to account for gender differences with boys tending to skip constructed response items more frequently than girls (Solheim & Lundetræ, 2013). These test-related factors that may lead to gender differences in reading performance will be explicated in more detail below.

### Test mode and reading purpose in PIRLS and ePIRLS

Martin and Binkley (2009) claim that the test modality matters to boys, suggesting that boys are more motivated for working at computers rather than paper and pencil, and that this gives them a disadvantage in paper-based reading assessments. This claim seems to find support in studies that show that males have more positive attitudes towards computers than females. In a recent meta-analysis, Cai, Fan and Du (2017) found that the attitudinal gender difference persists even in an age when Information and Communication Technologies (henceforth ICT) is close to ubiquitous, used in homes, at school and in the workplace. Gender differences related to test mode have indeed been identified in some countries participating in another large-scale international assessment, the Programme for International Student Assessment (PISA) for 15-year olds. Both Drabowicz (2014) and Jerrim (2016) found that the gender gap in mathematics in favour of boys widened in many countries in the 2012 digital test compared to the paper-based test from the same year.

On the other hand, Punter, Meelissen and Glas (2017) revealed that 14-year old girls outperformed boys in computer and information literacy skills in most countries participating in a recent large-scale survey (ICILS 2013; Fraillon, Ainley, Schulz, Friedman & Gebhardt, 2014, p. 103). This finding contradicts the expectation that the more positive attitudes among males compared to females lead to better results on computer-based tests, as assumed by e.g. Martin and Binkley (2009). Cai et al. (2017) include a timely reminder that both genders mainly exhibit positive attitudes

to ICT, indicating that test performance across modes might not be a question of males liking computers and females reporting to dislike them.

Because the question of test mode and gender differences is far from settled, we explore whether the belief that boys are disfavoured in paper-tests and subsequently will perform better on computer-based tests, finds any support in the PIRLS versus ePIRLS results.

Further, gender gaps have been consistently larger in favour of girls regarding literary compared to informational texts in former PIRLS cycles (Mullis et al., 2012). Considering that ePIRLS assesses informational reading only, and therefore contains no literary texts, one might expect the gender gap to be smaller in ePIRLS than it is in PIRLS. We test this by comparing the results of Norwegian girls and boys on the two reading purposes in PIRLS and the single reading purpose in ePIRLS.

In sum, we have identified two reasons to expect that Norwegian boys might close some of the gender gap in ePIRLS compared to PIRLS. The computer test modality (versus paper) might motivate boys to try harder, and the single informational reading purpose of ePIRLS could also work to the advantage of boys compared to the literary reading purpose assessed together with informational reading in PIRLS. Nonetheless, there are a couple of other features of large-scale assessments theorised to affect gender differences, namely item format and text features.

### Item format

Large-scale reading assessments such as PIRLS and PISA measure reading literacy by having the students read two different texts each (from a varied selection of texts rotated among the individuals) before responding to items ("questions") to the texts. Two types of items probe the different reading processes (see Methodology): Multiple choice (MC) items require the respondent to tick one out of four options of which only one is correct, whereas for constructed response (CR) items the respondent has to write/type an answer to a question. Ticking a box for MC items, either by pencil or computer mouse, is an easier operation than producing a response for CR items (Solheim & Lundetræ, 2016) – getting it correct is another matter. In addition, some of the CR items demand several elements and/or some depth in the response for a full score of up to 3 points. In contrast, MC items do not entail partial credit scoring, as they are scored as either zero points for an incorrect response or one point for a correct response. In other words, skipping CR items or giving minimum responses, may cost a student several points on such assessments.

Solheim and Lundetræ (2013) inspected item types in light of gender differences in the Norwegian national reading tests, constructed on very similar principles to PIRLS and PISA, and employing the same item formats. They found that boys skipped such constructed response items more often than girls did.

Solheim and Lundetræ (2016) later compared Nordic PIRLS, PISA and PIAAC[1] results, finding evidence that CR items disfavour boys in international reading assessments, too. The varying gender differences reported for these three different age group assessments may be partly explained by their ratio of CR items (PIAAC includes no such items and there is no gender gap). In PIRLS, there are approximately an equal number of points to gain from CR and MC items (ibid.), and this feature of test construction contributes to the larger gender differences found in Norway in international studies of reading compared to the much smaller gender differences found in the Norwegian national reading test. CR items are relatively rare in the national reading test (varying from 10% to 20%) compared to MC items (c. 80%).

Similarly, Schwabe, McElvany & Trendtel (2014) explored data from German 10-year old PIRLS and 15-year old PISA respondents, finding that girls in both age groups have an advantage over boys in CR items, regardless of motivation. The authors suggest that the increased ratio of CR items in international reading tests discriminates against boys.

The role of item format on gender differences in reading is worth inspecting in a comparison of results from PIRLS and ePIRLS. We calculate differences between item formats, CR versus MC, and test mode, paper versus computer. PIRLS and ePIRLS employ similar ratios of CR and MC items.

### Text features

Boys have been found to be more affected than girls by their interest (or lack of) in the topic of the text on comprehension tests (Oakhill & Petrides, 2007). On a similar note, Logan and Medford (2011), suggest that motivation may influence the effort that boys make in reading. Roe and Vagle (2012) propose that boys have difficulty relating to female protagonists in stories, and therefore underachieve on comprehension tasks to such texts. Further, Solheim and Lundetræ (2013) suggest that the possible disengagement of boys when reading about female characters, not only applies to literary texts, but may also affect boys' achievement on informational texts focussing on girls.

ePIRLS contains one text of particular interest: "Dr. Elizabeth Blackwell" is an informational text presenting the story of the first female doctor. It is also the only text with a female protagonist. The students reported how much they liked working with the tasks (texts and items) in ePIRLS. Table 2 illustrates the percent of students who liked a text a lot or a little (contrasting those students who did not like a text very much or not at all). Interestingly, boys report to like the Blackwell text much less than girls do.

[1] The Programme for the International Assessment of Adult Competencies (PIAAC) measures literacy, numeracy and problem solving in technology-rich environments among 16–65 year olds.

*Table 2.* Reported likes in percent and error of measurement in parenthesis on all texts in ePIRLS.

| | Percent of students who liked the task a lot or a little | | |
|---|---|---|---|
| ePIRLS text | Overall | Girls | Boys |
| Mars | 88 (0.3) | 87 (0.4) | 89 (0.3) |
| Rainforests | 93 (0.2) | 94 (0.3) | 92 (0.3) |
| **Dr. Elizabeth Blackwell** | 83 (0.3) | **88 (0.4)** | **78 (0.5)** |
| Zebra and Wildebeest Migration | 92 (0.2) | 93 (0.3) | 92 (0.3) |
| The Legend of Troy | 89 (0.3) | 89 (0.3) | 90 (0.3) |
| Average percent | 89 (0.1) | 90 (0.2) | 88 (0.2) |

*Source: Mullis, Martin, Foy & Hooper, 2017a, p. 6.*

Table 2 shows that girls and boys like most texts equally well, with the exception of "Dr. Elizabeth Blackwell". While all differences in the students' reported likes are significant, but differ by one to two percentage points for four of the five texts, there is a full ten-percentage point difference between the genders for the Blackwell-text. Of the girls that read the Blackwell-text, 88 percent reported to like working with it a little or a lot while a significantly less 78 percent of the boys reported the same. We test the hypothesis that boys underachieve on tasks that do not engage them, by calculating their scores on this text compared to the other texts in ePIRLS.

To summarise the findings presented in this chapter, we expect varying gender differences in the reading assessments PIRLS and ePIRLS based on test mode (paper-based versus digital assessment), reading purpose (informational versus literary purpose), item format (constructed response versus multiple choice) and text features (female protagonist).

## Research Questions And Hypotheses

The following hypotheses and research questions about gender differences in PIRLS and ePIRLS in the Norwegian student sample will be examined in this article:

1. Hypothesis: Motivated by the digital test mode, boys make more of an effort leading to higher achievement in ePIRLS compared to PIRLS. Research question: Do boys close some of the gender gap in ePIRLS compared to PIRLS?
2. Hypothesis: Boys perform better on informational reading tasks than on literary ones. Research question: Do boys perform significantly better on the informational reading purpose (both in PIRLS and ePIRLS) than they do on the literary reading purpose (PIRLS only)?
3. Hypothesis: Boys skip more CR items than MC items, and they perform less well than girls on CR items. However, this may not be true in a computer-based test condition. Research question: Do boys get significantly better scores on CR items in the computer-based ePIRLS than in the paper-based PIRLS?
4. Hypothesis: Text features such as female protagonists may disengage boys to such a degree that they underachieve on tasks connected to the text. The ePIRLS

Blackwell text is an example of such a text. Research Question: Do boys perform significantly worse on the Blackwell text than on the other informational texts in ePIRLS which they like better?

## Methodology

The innovative online informational reading assessment, ePIRLS, was introduced in 2016 as an addition to the paper-based PIRLS. Fifty countries participated in PIRLS 2016, with fourteen of them choosing to take part in the first round of ePIRLS. The assessment takes place every fifth year. ePIRLS only focusses on informational text reading while PIRLS tests reading literacy from both informational and literary texts. There is no overlap of the tasks in PIRLS and ePIRLS.

Both the PIRLS and the ePIRLS assessment follow the same definition and framework of reading literacy:

> *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, Martin & Sainsbury, 2015, p. 12).

The framework operationalises this definition of reading literacy by including test items that measure four cognitive processes believed to be central to reading comprehension (Mullis et al., 2015). Two of these cognitive processes belong to lower-order reading ability, e.g. finding and retrieving explicitly stated information, and two are considered higher-order reading abilities, such as reflecting on and evaluating texts. The two tests use the same type of items to probe the four reading processes. Multiple choice (MC) items give the respondent four (in the digital assessment, sometimes up to six) options out of which only one is correct, whereas constructed response (CR) items require the respondent to write or type an answer to a question. The latter item type is considered to be particularly salient in order to probe higher-order reading processes.

The representative sample consists of 3610 Norwegian students in grade 5 that took both the PIRLS- and ePIRLS-assessment during the same week, with one to three days between them. The students also answered context questionnaires, as did their parents, teachers and school principals. A number of students took only the paper-based (N = 622) test. For the analyses that focus on comparing student responses across both assessment modes, these students were subsequently excluded.

Table 3 highlights the differences between the excluded students and the overall sample in key aspects such as language spoken at home (dichotomised in "always Norwegian" and "almost always, sometimes or never Norwegian"), the amount of books at home (dichotomised in "100 books or less" and "101 books or more"), gender and achievement results in the PIRLS overall score, the score on literary texts only and the score on informational texts only. For the latter three, plausible values were used.

*Table 3.* Differences between included and excluded student samples in language spoken at home, amount of books at home, gender and PIRLS achievement (overall, literary purpose only, informational purpose only) in percentage and points, including error of measurement (SE).

| | | Included students | | Excluded students | |
|---|---|---|---|---|---|
| | | % | SE | % | SE |
| Language spoken at home | always Norwegian | 87.4 | 1.0 | 86.6 | 1.6 |
| | almost always to never | 12.6 | 1.0 | 13.4 | 1.6 |
| Amount of books at home | 100 books or less | 43.5 | 1.3 | 48.5 | 3.2 |
| | 101 or more books | 56.5 | 1.3 | 51.5 | 3.2 |
| Gender | Girl | 50.5 | 1.2 | 48.4 | 2.6 |
| | Boy | 49.5 | 1.2 | 51.6 | 2.6 |
| PIRLS Achievement Overall | | 560 points | 2.3 | 554 points | 5.0 |
| PIRLS Achievement Literary | | 561 points | 2.5 | 555 points | 5.4 |
| PIRLS Achievement Informational | | 560 points | 2.4 | 552 points | 5.7 |

As can be seen in Table 3, the excluded student sample of 622 Norwegian fifth-graders does not show significant differences to the rest of the sample apart from the amount of books at home. This was expected as attrition in ePIRLS was mainly due to technical failure at the schools when ePIRLS was implemented (Gabrielsen & Strand, 2017, p. 20) and was not linked to individual student characteristics. However, as the amount of books at home has shown a strong effect on student's test results in reading (Mullis et al., 2017a) and is regarded as an indicator for the cultural capital of a family (Bourdieu, 1983), this slightly positive bias in the sample will need to be taken into account when discussing the results of the analyses.

There are five different texts (= blocks) with corresponding items (= tasks) in ePIRLS and twelve texts in PIRLS (six literary and six informational texts), of which students responded to two in each assessment. The amount of items and item formats (MC and CR, see above) varies somewhat between the different texts: There are 17 items on average per text with an average of 8 MC items and 9 CR items. In both PIRLS and ePIRLS, students had 40 minutes to read and complete each of these text-based tasks. They had a short break between each part of the test.

For the analyses, Generalised Linear Mixed Models (GLMM) were specified in SPSS. While such regression models allow for the dependent variable to show a non-normal distribution, they can also account for random effects. In PIRLS and ePIRLS, the individual reading scores may vary without this being due to the independent variable (e.g. item format, gender and mode). Instead, it might be linked to "random" conditions in test behaviour. This is adequately modelled by specifying random effects in GLMM. As a result, test scores in the two assessments can be compared without differences being due to random test conditions. All analyses were weighted by the overall student sampling weight, known as TOTWGT in the PIRLS international databases (Martin, Mullis & Hooper, 2017).

Both the amount of items per block as well as the amount of items in the two different item formats (multiple choice and constructed response) vary between the text

blocks (see above). Also, students did not respond to all available blocks but received two blocks per assessment. Hence, individual test scores cannot be compared directly. Therefore, percent correct scores were computed, indicating the percentage of the maximum score a student could receive for his/her text block combination. Missing values on individual items were coded as incorrect (0 points). On CR items, student responses may be credited up to 3 points, depending on the quality of the answer. The maximum scores for each test block take these differing scores in CR items into account. To control for the varying psychometric characteristics of the different sets of tasks, all score-based analyses were controlled for by item difficulty.

When comparing student scores across the genders in order to link variance in gender differences to test features such as item format, the analyses were controlled for by general student ability, meaning boys' and girls' test scores in both tests (internationally computed plausible values). As described above, the paper-based PIRLS-assessment tests both informational and literary reading, whereas the digital ePIRLS-assessment tests informational reading only.

As described above, the paper-based PIRLS-test consists of tasks related to both informational and literary texts, while the digital ePIRLS-assessment only tests informational reading. For comparisons across test modes (paper versus digital), scores on informational tasks were used only.

## Results

### Test mode and reading purpose

First, we explore hypothesis 1, whether the gender gap in favour of girls decreases when the test mode is computer (ePIRLS) rather than paper (PIRLS).

*Table 4.* Gender differences of Norwegian 5th grade students in PIRLS and ePIRLS in score points with error of measurement (SE).

| | PIRLS | | ePIRLS | |
|---|---|---|---|---|
| | Score | SE | Score | SE |
| Girls' results | 571 | 2.8 | 576 | 2.6 |
| Boys' results | 549 | 2.6 | 558 | 2.9 |
| Gender difference | 22 | 2.3 | 18 | 3.2 |

*Source: Mullis et al., 2017a, p. 84.*

Table 4 shows that while girls score 18 points more than boys in ePIRLS, they score 22 points more than boys in the overall PIRLS assessment. However, the difference is insignificant, meaning that even if the gender gap seems smaller by numbers of points, the computer mode does not give boys a motivational advantage sufficient to close the gender gap among Norwegian 10-year olds. Hypothesis 1 is not supported.

Next, we check hypothesis 2, whether the reading purpose, literary versus informational, affects gender differences in the Norwegian sample. Table 5 illustrates this.

*Table 5.* Average score in the Norwegian 5th grade sample by gender, text type and test modality with error of measurement (SE).

| | PIRLS | | | | | | ePIRLS | |
|---|---|---|---|---|---|---|---|---|
| | Overall | | Literary | | Informational | | Informational | |
| | Score | SE | Score | SE | Score | SE | Score | SE |
| Girls' results | 570 | 2.8 | 571 | 2.7 | 568 | 2.8 | 576 | 2.6 |
| Boys' results | 548 | 2.6 | 550 | 3.2 | 549 | 2.9 | 558 | 2.9 |
| Gender difference | 22 | 2.3 | 21 | 3.1 | 19 | 3.0 | 18 | 3.2 |

*Sources: Mullis et al., 2017a, p. 20 and Mullis et al., 2017b, p. 141.*

Table 5 shows that there is but a slight difference in scores between Norwegian boys and girls on the informational text items in PIRLS and ePIRLS: Girls on average score 19 more points than boys on informational items in the paper-based assessment, while the difference on informational items in the digital setting is 18 points in favour of girls. For literary texts (paper-based only), the difference is bigger with 21 points in favour of girls. However, the differences are non-significant. The hypothesis that Norwegian boys perform (significantly) better on informational reading tasks is refuted.

## Gender effects of item format

Our third hypothesis concerns the different item formats in the reading tests. We compare boys' and girl' results on MC versus CR items. Boys might "lose" points on CR items for two reasons: They might skip them or they may tend to write shorter, more superficial responses than girls do, thus missing out on points even when they do respond to CR items. Hence, the general gender gap might be explained by girls tending to skip items less often and writing more elaborate answers, thus getting more points on CR items (in both test modes).

First, we examined whether boys tend to skip constructed response (CR) items more often than girls do. However, preliminary analyses showed that in both modes, the rate of items being skipped (either by girls or boys) was extremely low with 4.6% (3.2% for both MC and CR items). As this indicates that CR items are rarely skipped, we did not base any further analyses on it.

While we do not find that Norwegian girls or boys skip a considerable amount of items, we do find gender differences in their scores. Even when controlled for reading ability, girls perform better than boys both on CR and MC items, as can be seen from Table 6 illustrating the GLMM-analyses of the data. This finding supports hypothesis 3 that girls outperform boys on CR items in both test modes.

*Table 6.* Average percent correct scores of Norwegian 5th graders on CR and MC items in both PIRLS (informational texts only) and ePIRLS by gender, controlled for by item difficulty and student ability.

| | CR items | | | MC items | | |
|---|---|---|---|---|---|---|
| | PIRLS | ePIRLS | Mode difference | PIRLS | ePIRLS | Mode difference |
| Girls' results | 79.9 | 73.1 | –6.8 | 75.9 | 78.1 | +2.2 |
| Boys' results | 76.3 | 70.6 | –4.7 | 74.0 | 76.1 | +2.1 |
| Gender difference | 3.6 | 2.5 | –1.1 | 1.9 | 2.0 | +0.1 |

The gender difference on CR items in the paper-based PIRLS (informational texts only) is significant ($F(92, 25.733) = 17.5$, $p < .001$) with 3.6 percentage points between boys and girls (see Table 6, controlled for general student ability in reading). The percent correct score for girls on this item type is 79.9 percent correct, whereas for boys it is 76.3 percent correct.

The analysis of similar (but not identical) constructed response items in the digital ePIRLS reveals that girls again outperform boys by 2.5 percentage points between 73.1 percent correct for girls and 70.6 percent correct for boys (Table 6), and again, the difference is significant ($F(66, 45.113) = 32.1$, $p < .001$) (Table 6). If we interpret the difference between how well girls perform on CR items in the paper-based PIRLS compared to the digital ePIRLS as a mode effect, we find it to be significantly stronger in girls (6.8 percentage points in favour of paper, $F(79, 35.956) = 25.3$, $p < .001$) than in boys (4.7 percentage points in favour of paper; $F(79, 34.890) = 21.5$, $p < .001$; significant gender difference with $F(158, 70.874) = 20.8$, $p < .001$).

These analyses reveal that girls indeed show higher achievement on CR items than boys in both modes, even when controlling for student ability. We find little support for hypothesis 3, stating that Norwegian boys perform better or make more of an effort on constructed response items in the computer mode of ePIRLS than they do in the paper-based PIRLS. Boys' mean is 76.3 percent correct on paper and 70.6 percent correct on computers. The difference of 4.7 percentage points is in favour of paper, even for boys.

While the main focus of these analyses is on how well students perform on CR items in both test modes, we find a different picture for student performance on MC items. Controlling for student ability (and item difficulty), the scores of boys and girls vary slightly less between the genders with a significant gender gap of 1.9 percentage points in the paper-based PIRLS ($F(80, 23.298) = 16.0$, $p < .001$) and a very similar gender difference of 2.0 percentage points for the digital ePIRLS ($F(72, 50.415) = 43.2$, $p < .001$), both in favour of girls. This is significant ($F(152, 73.713) = 29.2$, $p < .001$), but from a practical perspective irrelevant. The difference of 0.1 percentage points in how the gender gap varies between the modes strongly points towards a general mode effect that is equally strong for both genders. This can indeed be attested by the similar difference in how well girls perform on MC items in the two modes (2.2 percentage points in favour of ePIRLS, significant with $F(76, 37.287) = 30.8$, $p < .001$) compared to how well boys perform on MC items in the two modes (similar 2.1 percentage points in favour of ePIRLS, significant with $F(76, 36.426) = 27.5$, $p < .001$). We sum up the findings (all controlled for student ability in reading and item difficulty):

(1) Girls outperform boys on CR items, both in the paper-based PIRLS and the computer-based ePIRLS assessment.
(2) Boys do not perform significantly better on CR items in the digital assessment compared to the paper-based one.
(3) On CR items, both girls and boys perform better in the paper-based assessment.

(4) On MC items, both girls and boys perform better in the digital assessment.
(5) The mode effect in how well students perform on these different item formats in paper versus digital is
    (a) generally stronger for CR items and weaker for MC items,
    (b) stronger for girls on CR items compared to boys on CR items and
    (c) about the same for boys and girls on MC items.

## Text features

Our last hypothesis remains: We explore gender differences in reading performance in relation to how well boys like a text. Text features such as female characters may negatively affect boys' motivation for reading a text, and subsequently, how well they achieve on items connected to the text. One of the informational texts in ePIRLS, "Dr. Elizabeth Blackwell"[2], tells the story of a woman. As the students reported how well they liked working with the texts (see Table 2), we know that the Blackwell-text did not appeal to boys as much as the other texts did (none of which include female protagonists).

We explored whether boys' relative disfavouring of the Blackwell text is reflected in their scores on the associated items. For this analysis, presented in Table 7, we again calculated percent correct scores but included both MC and CR item formats.

*Table 7.* ePIRLS percent correct scores of MC and CR items by gender and text, including significance testing by T-tests for independent samples.

| Text | Boys' results | Girls' results | Gender difference | Significance of gender difference |
|---|---|---|---|---|
| *Mars* | 71.7 | 74.5 | 2.9 | T(7950) = 2.868, p < .001 |
| *Rainforests* | 71.1 | 74.0 | 2.8 | T(7930) = 2.831, p = .005 |
| ***Dr. Elizabeth Blackwell*** | **64.3** | **72.3** | **8.0** | **T(5686) = 6.512, p < .001** |
| *Zebra and Wildebeest Migration* | 73.6 | 77.2 | 3.6 | T(8046) = 3.703, p < .001 |
| *The Legend of Troy* | 62.0 | 66.2 | 4.2 | T(6990) = 3.691, p < .001 |

The results for the five texts in ePIRLS (Table 7) give strong support for the hypothesis that boys' reading achievement is affected by their text preference. They score an average of 64 percent correct on the Blackwell-text while scoring significantly better[3] on almost all other texts, ranging from 71 to 73 percent correct. The only exception is the text "The Legend of Troy". There is no significant difference between how well boys score on the Blackwell-text compared to "The Legend of Troy" (T(6327)

[2] "Dr. Elizabeth Blackwell" is one of two texts that are not to be used in later ePIRLS test cycles, and thus was published in December 2017. The other text is "Mars".

[3] T-test for dependent samples; comparison of the percent correct of boys for *Blackwell* with *Mars*: T(6778) = 6.470, p < .001 for a difference of 7.4 percentage points, *Blackwell* with *Rainforests*: T(6684) = 5.931, p < .001 for a difference of 6.8 percentage points, *Blackwell* with *Zebra and Wildebeest Migration*: T(6817) = 8.285, p < .001 for a difference of 9.3 percentage points.

= 1.871, p = .061 for a difference of 2.3 percentage points). However, girls' results on this text reveal that this is not a gender difference, but rather that the Troy-text[4] is difficult for both boys *and* girls.

Table 7 also illustrates that the difference between boys' and girls' results on items connected to the Blackwell-text is much larger with 8 percentage points than on any of the other texts. This strongly reflects the gendered responses boys and girls gave to how well they liked the texts in ePIRLS (Table 2). It supports our hypothesis 4 that boys may be disengaged by the text topic and the gender of the protagonist and hence seem to underachieve when text topics do not motivate them.

## Discussion

The analyses of the Norwegian PIRLS and ePIRLS results have contributed to the field of gender and mode differences in reading assessment. First, we found no evidence that Norwegian boys close the gender gap in a computerised test, ePIRLS, assessing informational reading only. The optimistic view of Martin and Binkley (2009) that digitised reading assessment leads to smaller gender differences is thus not supported. Even though males still seem to have more positive attitudes towards ICT (Cai et al., 2017), this seems not to affect the reading achievement of young boys in ePIRLS, at least not in Norway. Norwegian boys do not perform significantly better in ePIRLS than they do in PIRLS, even when literary reading is accounted for.

Secondly, it was found that girls score significantly better than boys on CR items in both test modes, even when controlling for general test performance. The superiority of girls' results on CR items compared to boys' results, supports previous research (Solheim & Lundetræ, 2013, 2016; Schwabe et al., 2014), showing that girls have an advantage over boys when CR items are plentiful in reading assessments. This is not explained by boys skipping more CR items than girls in either PIRLS or ePIRLS. There were very few missing responses, an indication that most students had sufficient time and motivation to complete the tests.

We also found that for both girls and boys, CR items seem to be more difficult in the digital assessment than on paper. This could be due to Norwegian children generally finding it more difficult to type answers on a keyboard than to write them by hand.

A different explanation is that children treat digital reading and writing less seriously than they treat writing with a pen on paper. It has been suggested that computer reading behaviour is fast and shallow compared to deep paper reading (e.g. Baron, 2015; Wolf, 2007). Additionally, children use digital devices (computers, tablets, and increasingly, mobile phones) for entertainment and socialising (Livingstone,

[4] While we do not find a significant difference for how well boys score on the Blackwell- compared to the Legend of Troy-text, we find it for girls (T(6349) = 5.200, p < .001 for a difference of 6.1 percentage point).

Haddon, Vincent, Mascheroni & Olafsson, 2014); activities that typically do not involve much elaborate writing. At school, children may still write by hand when they compose texts in various subjects. It is possible that school children in Norway connect serious writing to the paper mode, while associating computer-based activities, even when these involve typing, with less school-like activities.

We have not explored a possible third explanation: We strongly suspect that the digital ePIRLS yields the opportunity to copy and paste responses to CR items. If this is the case, it is possible that Norwegian children opt for the easier solution to copy a minimal response from the text instead of typing more elaborate answers by themselves, thus missing out on top scores on CR items graded scoring. Copy and paste behaviour may occur in combination with the argument above, that digital text presentations lead to a shallow reading. However, this must remain as speculation until further research is performed concerning exactly how students respond to CR items in ePIRLS.

While we find that girls outperform boys on CR items in both test modes, with CR items being more difficult for all students in a digital assessment, we also find a slightly smaller gender difference for them in the digital assessment than on paper. When comparing the percent correct points girls and boys "lose" across the modes we find it to be less strong in boys than in girls: On average, boys receive 4.7 percent correct points less in the digital assessment compared to the paper-based one while girls receive 6.8 percent correct points less in it. This overall "drop" in scores can be interpreted as a general mode effect for CR items that seem to be more difficult in the digital assessment. The mode effect also seems to be less strong for boys. While we did not test males' positive attitudes to computers (as suggested by Cai et al. (2017) for the explanation of gender differences), we do find that boys seem to be less effected by responding to CR items on a computer than girls are which may be linked to motivational aspects, such as typing on a keyboard as opposed to writing with a pen on paper.

For MC items, we found the opposite to be true. While girls still outperform boys on MC items in both tests (even while controlling for reading ability), we only find a very small change in gender differences across test modes and – interestingly – a positive mode effect with both girls and boys similarly performing better in the digital assessment. While we hypothesised that the mode effect for CR items is less strong in boys because they might be more motivated to respond to questions by typing on a keyboard than using a pen on paper, it might not be computers *per se* that motivate boys (as suggested e.g. by Martin & Binkley, 2009). If so, we would have expected the mode effect for MC items (that differ in clicking with the computer mouse versus ticking boxes with a pen on paper) to be equally different for boys and girls. Summarising these results from PIRLS and ePIRLS, we find that boys are not motivated by taking reading assessments on a computer to such an extent that they perform equally well or better than girls in Norway. However, they seem to profit from it when typing answers on a keyboard (as opposed to writing them with a pen on paper). This

strongly points towards more in-depth analyses on what mode-related factors lead to varying test performances of boys and girls.

Finally, there is the hypothesis regarding boys' weaker performance on texts that do not interest them. On a post-reading comprehension test, Oakhill and Petrides (2007) found that boys scored significantly better on the questions to the one out of two texts they were most interested in, whereas girls performed equally well on questions to both texts, regardless of their interest in the topic prior to reading. It has been found that boys' test effort depends on their interest and motivation (Oakhill & Petrides, 2007), and these factors may in turn depend on whether the reader can identify with the main character of a text (Roe & Vagle, 2012; Solheim & Lundetræ, 2013).

The students' reports revealed that boys liked the Blackwell-text less than they liked the other texts in ePIRLS (Table 2). Its partly narrative structure and its female protagonist are both factors that according to research (Roe & Vagle, 2012; Solheim & Lundetræ, 2013) may reduce text appeal and motivation for boys. We found that boys performed significantly worse on Blackwell than they did on all but one of the other texts (with a non-significant difference in performance to "The Legend of Troy"). While we find gender differences in favour of girls for all texts in ePIRLS, the difference between boys and girls is the strongest for the Blackwell-text. This supports previous research (ibid.) that Norwegian boys are disengaged by and perform less well on the items related to texts with female main characters. If a young boy has difficulty identifying with the main character of a story, this might influence his test performance, as reading effort depends more on interest and motivation for boys than it does for girls (Logan & Medford, 2011; Oakhill & Petrides, 2007).

To test whether boys score worse on a text than girls can be attributed to the topic or even the gender of the protagonist, a control study should be conducted consisting of two identical texts that only vary in the gender of the protagonist (or the general topic). However, our findings already indicate a strong relationship between boys' relating to a female protagonist in stories and their achievement on comprehension tasks to such texts.

## Research Perspective

Comparing the scores on CR items across the modes, we found a tendency towards a general mode effect when it comes to constructing and then writing responses on a computer versus on paper with both boys and girls performing significantly lower on CR items in ePIRLS. The possible mode effect seems to affect girls in particular. In this cycle of PIRLS and ePIRLS 2016, a definite mode effect will be difficult to prove as both tests consist of different texts and related items, with none of them overlapping between the modes. Studies such as TIMSS and PISA are now being switched to digital-only assessments; a process that is accompanied by bridging-studies that allow for such testing as they mostly consist of the same items both on paper and

in the digital test environment. Even in such elaborate designs developed to test for mode effects, analyses are complex due to the fact that mode-specific characteristics such as screen size and resolution, scrolling and using a mouse instead of a pen cannot be avoided, even when items are supposed to be exactly the same on paper as on a PC, laptop or tablet. With ePIRLS and PIRLS not including overlapping items, attributing differences in scores to the mode is hence challenging to test for, even when controlling for varying item difficulties. The finding that boys and girls show lower scores on CR items when they write on a computer than on paper should therefore be validated with a study that is designed for such comparisons as it may shed further light on reading and writing behaviour in a digital environment.

## References


Bourdieu, P. (1983). Ökonomisches Kapital, kulturelles Kapital, soziales Kapital [Economic capital, cultural capital, social capital]. In *Soziale Ungleichheiten. Sonderband 2 der Zeitschrift Soziale Welt* (pp. 183–198). Göttingen: Schwarz.

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

Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: A meta-analysis. *Computers & Education*, *105*. 1–13.

Drabowicz, T. (2014). Gender and digital usage inequality among adolescents: A comparative study of 39 countries. *Computers & Education 74*, 98–111.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014*). Preparing for Life in a Digital Age. The IEA International Study of Computer and Information Literacy: International Report*. Springer Open.

Gabrielsen, E. & Strand, O. (2017). Rammer og metode for PIRLS 2016 [Framework and methods for PIRLS 2016]. In E. Gabrielsen (Ed.), *Klar Framgang! Leseferdighet på 4. og 5. trinn i et femtenårsperspektiv* (pp. 13–31). Oslo: Universitetsforlaget.

Gabrielsen, E., & Hovig, J. (2017). Hovedresultater fra PIRLS 2016 i Norden [Main results for PIRLS 2016 in the Nordic countries]. In E. Gabrielsen (Ed.), *Klar Framgang! Leseferdighet på 4. og 5. trinn i et femtenårsperspektiv* (pp. 32–49). Oslo: Universitetsforlaget.

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.

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.

Logan, S., & Medford, E. (2011). Gender differences in the strength of association between motivation, competency beliefs and reading skill. *Educational Research*, *53*(1), 85–94.

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. 75–81). JRC Scientific and Technical Reports.

Martin, M.O., Mullis, I.V.S., & Hooper, M. (2017). *Methods and Procedures in PIRLS 2016.* TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA). Retrieved from https://timssandpirls.bc.edu/publications/pirls/2016-methods/P16_Methods_and_Procedures.pdf

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011. International results in reading.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Foy, P., & Hooper, M. (2017a). *ePIRLS 2016. International results in online informational reading.* TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA). Retrieved from http://timssandpirls.bc.edu/pirls2016/international-results/wp-content/uploads/structure/CompletePDF/P16-ePIRLS-International-Results-in-Online-Informational-Reading.pdf

Mullis, I.V.S., Martin, M.O., Foy, P., & Hooper, M. (2017b). *PIRLS 2016. International results in reading.* TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International

Association for the Evaluation of Educational Achievement (IEA). Retrieved from http://timssandpirls.bc.edu/pirls2016/international-results/wp-content/uploads/structure/CompletePDF/P16-PIRLS-International-Results-in-Reading.pdf

Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2015). PIRLS 2016 reading framework. In I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2016 assessment framework* (2nd ed., pp. 11–30). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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. doi:10.1348/000712606X117649.

Punter, R.A., Meelissen, M.R.M, & Glas, C.A.W. (2017). Gender differences in computer and information literacy: An exploration of the performances of girls and boys in ICILS 2013. *European Educational Research Journal*, *16*(6). 762–780.

Roe, A., & Vagle, W. (2012). Kjønnsforskjeller i lesing – et dybdedykk i resultatene fra nasjonale prøver på åttende trinn fra 2007 til 2011 [Gender differences in reading – a dive into the results from national reading tests in the 8th grade from 2007 to 2011]. *Norsk Pedagogisk Tidsskrift 6*(96), 425–441.

Schwabe, F., McElvany, N., & Trendtel, M. (2014). The School Age Gender Gap in Reading Achievement: Examining the Influences of Item Format and Intrinsic Motivation. *Reading Research Quarterly*, *59*(2). 219–232.

Solheim, O.J., & Lundetræ, K. (2016). 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.

Solheim, O.J., & Lundetræ, K. (2013). Prøveutformingens betydning for rapporterte kjønnsforskjeller [How test construction may influence gender differences]. In E. Gabrielsen, & R. G. Solheim (Eds.), *Over kneiken? Leseferdighet på 4. og 5. trinn i et tiårsperspektiv* (pp. 61–76). Oslo/Trondheim: Akademika forlag.

Wolf, M. (2007). *Proust and the squid: The story and science of the reading brain.* New York: Harper Collins.