Mode equivalency in PAT: Reading Comprehension

Jan Eyre, Melanie Berg, Jess Mazengarb, and Elliot Lawes

NZCER
Rangahau Mātauranga o Aotearoa

Mode equivalency in PAT: Reading Comprehension

Jan Eyre, Melanie Berg, Jess Mazengarb, and Elliot Lawes

2017

NZCER
Rangahau Mātauranga o Aotearoa

New Zealand Council for Educational Research
PO Box 3237
Wellington
New Zealand

ISBN 978-0-947509-64-4

Summary

As pen and paper assessments increasingly move online, questions about the equivalency of assessment using these formats emerge. Does an online assessment measure the same skills as a paper-based assessment? This report details the results of an investigation into a school-based assessment, PAT: Reading Comprehension, which has recently gone online. The investigation compared student achievement data from the online and paper-based modes of the assessment to find out whether the test items were equivalent across modes, and whether student scores were comparable across modes.

The results of the investigation suggest that while test items behaved in similar ways in the online and paper-based modes, students' scores were significantly[1] lower in the online mode. These findings are discussed in terms of their implications for assessment design.

1 In this paper, the word 'significant' is always used to refer to statistical significance. An alpha level of 0.05 was used for all statistical tests, unless otherwise stated.

Contents

Tables

Figures

1. Introduction

The move to digital assessment

Increasingly, assessments are being offered in a digital format. This move to digital assessment (also known as computer-based or online assessment) can be seen as a natural consequence of the ever-increasing use of technology both inside and outside of education. Although computer-based, online and digital assessment are not synonymous in general, we do not investigate the differences between them in this paper. Instead, we focus on the difference between paper-based assessment and these other modes of assessment. Therefore, in this report, we will mostly use the terms 'computer-based', 'online' and 'digital assessment' interchangeably.

Digital technology offers particular benefits for large-scale assessments. The online format streamlines administration through reducing the time needed to set up, grade and report on assessments. However, schools and students vary in their readiness and capability for digital assessment. As an interim stage, to cater for schools or students who are not yet ready to move to digital assessment, it is common to offer these large-scale assessments with a choice of modes: paper-based or computer-based.

When the same assessment is offered in two different modes, it is important to establish that the modes are comparable: that is, that the assessment administered on paper assesses the same construct, in the same way, as the assessment administered online. It is often assumed that the two modes are equivalent and can be scored on the same scale (known in the psychometric literature as 'measurement equivalence'). However, this assumption needs to be tested.

Over the past two decades there have been many studies that compare the results of computer-based and paper-based assessments. Generally, the differences between modes have been found to be small, sometimes favouring online assessment and sometimes favouring paper-based assessment. It is hard to generalise from these studies, as fast-moving changes in technology and in students' familiarity with computer-based devices mean that the conditions for online assessments are constantly changing. The majority of studies also involve assessments with a simple multichoice question format, where test-takers select a correct answer, rather than having to construct (write) their own answers.

Concerns about equivalence

Two recent large-scale assessments in the United States have raised concerns about the comparability of computer and paper-based assessments in which students write (construct) their responses. In 2012, fourth-grade students took a pilot version of a computer-based National Assessment of Educational Progress (NAEP) writing assessment. The results were compared with scores from a paper-based version of the assessment. The findings showed that high-performing students (those in the top 20%) scored 'substantively higher' on computer than on paper. Low and middle-performing students did not appear to benefit from using the computer. The authors conclude that "the use of the computer appeared to widen the achievement gap" (White, Kim, Chen, & Liu, 2015, p. vii). This aligns with findings from a review by Eyre (2015) of research relating to the validity and reliability of written assessments offered in dual modes. This review concluded that transferring written assessments to online platforms "without thoroughly investigating the possible mode effects … may create unfair situations that will widen the gap between low and high-achieving students" (Eyre, 2015, p. 20).

Differences in scores between modes were also found in the Partnership for Assessment of Readiness for College and Careers (PARCC) exams in 2014–15, which are aligned to the Common Core State Standards and are taken by students in grades 3–8+. Students tended to score more highly on paper than on computer. The advantages for paper were most noticeable in English and language arts and middle to upper grade mathematics. One reason put forward by PARCC to explain this is that students' level of familiarity with the computer-based system affected their scores in the online mode (Herold, 2016).

A range of factors are thought to impact on the comparability of paper-based and computer-based assessments. Besides student familiarity with computers and the systems used to deliver the assessments, these factors include the design of the digital interface; screen size and resolution; the amount of scrolling required; students' ability to comprehend text when it is presented on screen; and fluency of keyboarding skills (Karkee, Kim, & Fatica, 2010; Randall, Sireci, Li, & Kaira, 2012; both cited in Darr, 2014).

Given the potential for students to be advantaged or disadvantaged by a particular mode, it is important to design online assessments thoughtfully and to find out whether the two modes are equivalent. The equivalence of two modes must be carefully considered before moving a paper-based assessment to a digital platform.

> When tasks are moved from pen and paper to the computer, equivalence is often assumed, but this is not necessarily the case. For example, even if the paper version has been shown to be valid and reliable, the computer version may not exhibit similar characteristics. If equivalence is required, then this needs to be established. (Noyes & Garland, 2008, p. 1362)

In 2015, the New Zealand Council for Educational Research (NZCER) conducted and reported on a small-scale study that compared paper and computer-based versions of the Progressive Achievement Test of Mathematics (PAT: Mathematics). Four Year 8 classes took both versions of the test over a 2-week period. An analysis of the results showed very little difference between total scores in each mode, and differences in individual item difficulty between the two modes were also relatively small. Darr (2014, p. 63) noted that PAT: Mathematics online was designed to be as similar as possible to the paper-based version. It used only multichoice questions and "the only computer action required was to point and click".

The current study extends the work on PAT: Mathematics to another of the NZCER Progressive Achievement Tests—PAT: Reading Comprehension.

Assessing reading comprehension online

Assessments of reading comprehension usually involve reading passages of text and answering comprehension questions. Transferring a reading assessment to an online format is not straightforward. Most of us would agree that reading online and reading on paper are different experiences. A number of studies comparing online and paper-based reading assessment have confirmed this.

Some research studies have found that there is a negative effect on reading comprehension when we read online, compared to reading on paper (Kerr & Symons, 2006; Mangen, Walgermo, & Bronnick, 2013). For example, in their study of computer and paper-based reading assessments with students in two Norwegian schools, Mangen et al. (2013) found that "reading linear narrative and expository texts on a computer screen leads to poorer reading comprehension than reading the same texts on paper" (p. 67). Several reasons have been suggested for these differences. We know that reading from a computer screen is typically slower than reading from paper (Kerr & Symons, 2006; Kim & Kim, 2013; Noyes & Garland, 2008). It has also been suggested that computer-based assessments of reading involve a higher cognitive load and can be more tiring than their equivalent paper-based versions. This may be especially true for assessments that involve sophisticated tasks that require sustained attention. Bridgeman, Lennon, and Jackenthal (cited in Bridgeman, 2009) suggested that resolution of the monitor and amount of scrolling required by the test-taker may also affect performance.

The issue of scrolling is particularly relevant to reading comprehension assessments that involve lengthy passages of text. Bridgeman et al. found that students who could see the whole passage of text without scrolling performed better on a reading assessment than those who had to scroll to see the full passage (cited in Bridgeman, 2009). As Kingston (2008, p. 32) points out:

> Reading while scrolling is cognitively different than reading a page. While reading a page students can use spatial memory clues. They may remember they saw some information pertinent to answering a particular question in the upper right portion of the page and quickly return to that spot. Parallel clues are not available on a traditional computer-administration system because scrolling constantly changes the spatial frame of reference.

Schroeders and Wilhelm (2011) found a test of reading comprehension (for German high school students learning a foreign language) to be 'mode invariant'—scores for the online and paper-based versions were equivalent. However, the online version was carefully designed to be as similar as possible to the paper-based version. Only multichoice questions were used, and all the texts could be read without scrolling. These two considerations—amount of scrolling and type of questions—seem to be important factors in the design of online assessments, especially those that are also offered on paper. These two factors were taken into consideration when NZCER's PAT: Reading Comprehension was moved to an online format.

PAT: Reading Comprehension

The Progressive Achievement Test of Reading Comprehension (PAT: Reading Comprehension) is a low-stakes, standardised assessment developed for use in New Zealand schools. It is designed to provide formative information about students' ability to make meaning from written text.

There are seven different tests, each targeted at a specific year level from Year 4 to Year 10 (Test 1 targets Year 4, Test 2 targets Year 5 and so on). Each test consists of a set of instructions; two example questions; and a series of reading passages, each with an associated set of multiple-choice questions. The tests cover a range of text types, including poems, narratives and transactional texts.

PAT: Reading Comprehension was first developed and used in schools in 1969. A second, revised edition was developed in 2008. In 2014, an online version of the assessment was released. In the paper-

based version, students are given a test booklet and a separate answer sheet. The text booklet has an introductory page with instructions and example questions. Each subsequent double-page spread has a passage of writing on the left-hand page, with a series of associated multiple-choice questions on the right-hand page. Students work their way through the booklet, recording their answers to each question on the separate answer sheet. Each test booklet has either seven or eight passages of writing with between three and seven related questions on each. The instructions tell students that they should answer each question and that they have 45 minutes to complete the test. Most students complete the test well within this time.

Similarities and differences between computer and paper-based modes

The online version of PAT: Reading Comprehension was designed to match the paper-based version as closely as possible. Students log in to the relevant test and work through the assessment screen-by-screen. The introductory screens feature the same instructions and examples as the paper-based version, with small modifications to reflect the online rather than paper-based environment. On each subsequent screen, a passage of writing appears on the left-hand side of the screen, with the associated questions on the right. Online and paper-based versions of the tests use the same passages of writing and associated questions.

While the online test was designed to be as similar as possible to the paper-based version, there are some differences. These differences are in presentation and in the tools available to students.

Differences in presentation of texts and questions

Scrolling

Some passages of writing do not display on a single screen; the student has to scroll down to access the whole text. A scrolling bar appears on these items[2] to indicate that students need to read down to the end of the text. The student also has the option of seeing an alternative text layout by clicking on an 'eye' icon at the top right of the screen. The questions then collapse and the text spreads wider across the screen. This reduces but does not entirely remove the need for scrolling.

Layout of text

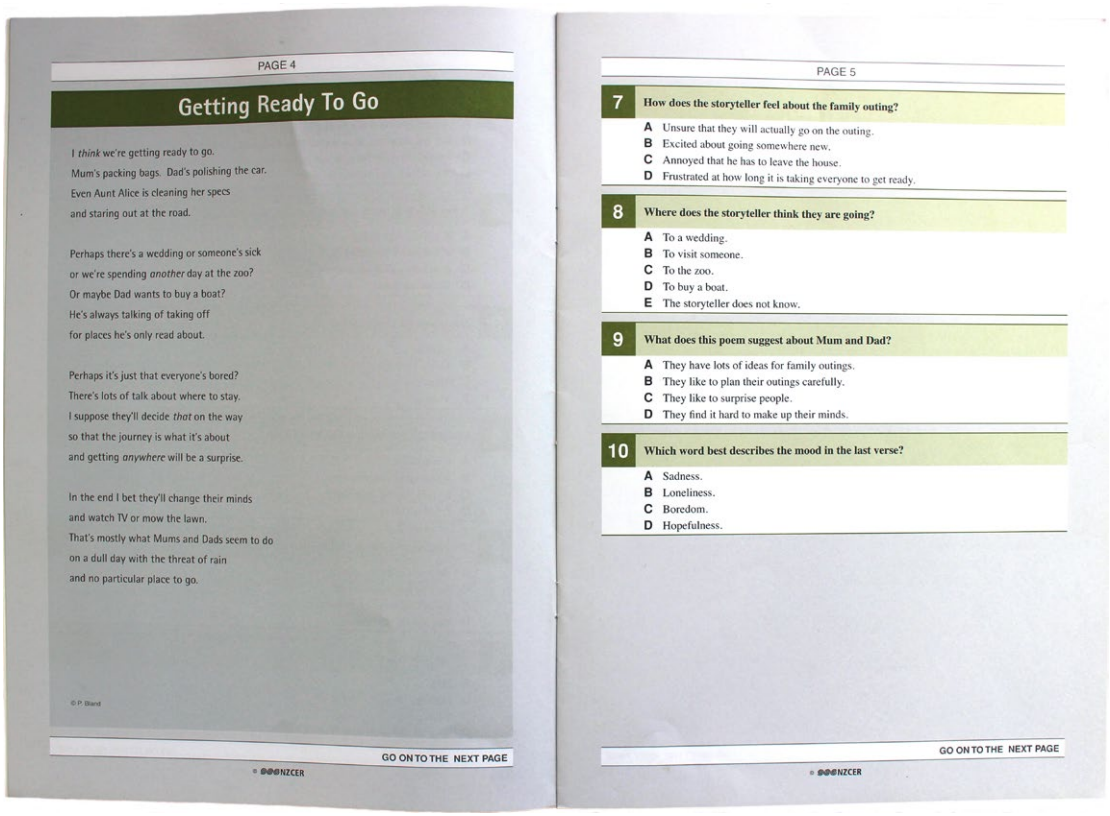
The passages of writing are narrower on screen than on the paper-based version (each line is shorter in length, meaning that the overall text is longer on screen). However, as mentioned above, there is also the option to display the text in a wider format. In this case, each line of text is wider than it is on paper. Paragraphs onscreen are separated by a clear line space without an indentation, whereas on paper, most paragraphs are indented.

Layout of questions

Students see one full question and its associated options at a time on the online version, whereas with the paper-based version students see the full list of questions and options associated with each text.

2 The terms 'question' and 'item' are used interchangeably in this report.

FIGURE 1 **PAT: Reading Comprehension Test 4, Q7 (paper version)**



PAGE 4

Getting Ready To Go

I *think* we're getting ready to go.
Mum's packing bags. Dad's polishing the car.
Even Aunt Alice is cleaning her specs
and staring out at the road.

Perhaps there's a wedding or someone's sick
or we're spending *another* day at the zoo?
Or maybe Dad wants to buy a boat?
He's always talking of taking off
for places he's only read about.

Perhaps it's just that everyone's bored?
There's lots of talk about where to stay.
I suppose they'll decide *that* on the way
so that the journey is what it's about
and getting *anywhere* will be a surprise.

In the end I bet they'll change their minds
and watch TV or mow the lawn.
That's mostly what Mums and Dads seem to do
on a dull day with the threat of rain
and no particular place to go.

© P. Bland

GO ON TO THE NEXT PAGE

PAGE 5

7 How does the storyteller feel about the family outing?

A Unsure that they will actually go on the outing.
B Excited about going somewhere new.
C Annoyed that he has to leave the house.
D Frustrated at how long it is taking everyone to get ready.

8 Where does the storyteller think they are going?

A To a wedding.
B To visit someone.
C To the zoo.
D To buy a boat.
E The storyteller does not know.

9 What does this poem suggest about Mum and Dad?

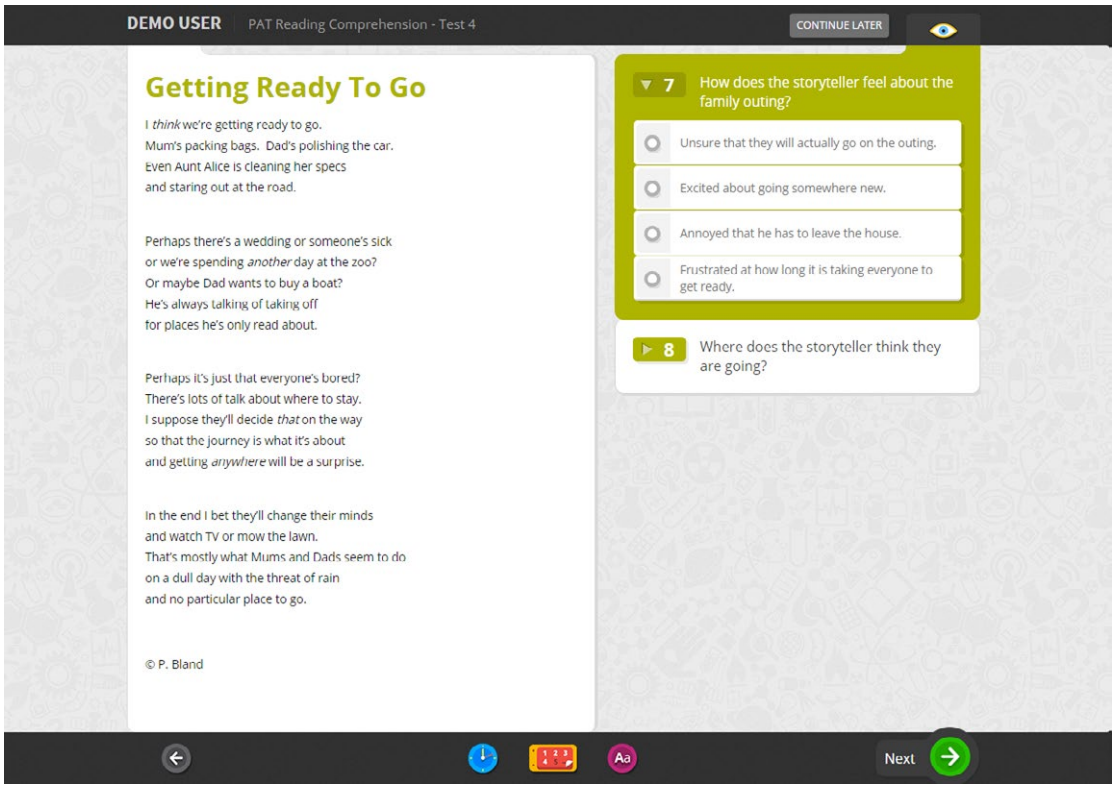
A They have lots of ideas for family outings.
B They like to plan their outings carefully.
C They like to surprise people.
D They find it hard to make up their minds.

10 Which word best describes the mood in the last verse?

A Sadness.
B Loneliness.
C Boredom.
D Hopefulness.

GO ON TO THE NEXT PAGE

FIGURE 2 **PAT: Reading Comprehension Test 4, Q7 (online version)**



Differences in tools

The online version includes several tools that are not available in the paper-based version.

Clock

An on-screen clock shows students how much time (of 45 minutes) is remaining, along with a progress bar showing how many questions they have answered.

Customised screen display

Students are able to select from three different fonts, including a 'dyslexia help' font. They are also able to increase or decrease the size of the font and alter the colours of both the text and background of the on-screen display. The colour choices are dark text against a light background, white text against a black background, and dark text against a pale brown background. The default font is different from the font used in the paper-based version.

These differences in design and tools, and the fact that reading online may be different from reading on paper, have the potential to impact on the equivalence of the two modes of assessment. Since its launch in 2014, many schools have moved to the online version of PAT: Reading Comprehension. Others are still using the paper-based version. The current study draws on data from both modes and uses statistical analysis to answer questions about the equivalency between paper-based and computer-based modes of the assessment. As with Darr's earlier small-scale study of PAT: Mathematics (2014), there were two main questions guiding the research:

- Are test items equivalent across the two modes?
- Are student scores comparable across the two modes?

2.

Methodology and results

This section details the methodology used to answer each of our research questions, the data we used and the results of our analyses.

Our data consisted of PAT: Reading Comprehension assessment records for students in 2014 and the start of 2015, stored in the NZCER Marking Service database. We downloaded all online and paper-based PAT: Reading Comprehension assessment data from the NZCER Marking Service. The Ministry of Education's 2014 Schools Directory was used to match assessment records with school demographic information (e.g. school roll, school decile).

Investigating our research questions necessitated two different analysis methods. To compare the difficulty of the test items in each mode we used the Rasch measurement model (Rasch, 1980). To compare student assessment scores across the two modes we used multilevel modelling. Details about these analyses are given in their respective sections.

We used the software environment R version 3.2.0 for all data management and data cleaning (R Core Team, 2015). We used the R package lme4 for multilevel modelling (Bates, Maechler, Bolker, & Walker, 2015), and the software package Winsteps for Rasch analysis (Linacre, 2012a).

Data and data cleaning

This section describes the data we used for our analyses, and the decisions made to resolve errors and inconsistencies that would affect the validity of those analyses.

The analysis techniques we used to approach each of our two research questions meant that we derived two different datasets from the PAT: Reading Comprehension data: one suitable for the analysis of items and one suitable for the analysis of student assessment scores.

Invalid assessment records

Some of the assessment records were not valid for our analyses and were therefore excluded from both datasets. These included:

- records created by users trying out the NZCER marking services using the demo site
- records created by NZCER developers carrying out site testing
- duplicate records created for school administrative purposes
- a small number of records that appeared to be labelled with an incorrect test number
- records from overseas schools.

PAT: Reading Comprehension item analysis

Data

Student ethnic group and gender

The assessment data contained two variables for student ethnic group: one administrative, recorded in the school's student management system (SMS); and one student specified. The administrative variable was used in item analysis because it was more complete and more consistent between records for the same individual.

If student ethnic group or gender were missing from an assessment record, they were back-filled where possible, using other assessment records associated with the same National Student Number (NSN).

Incomplete assessment records

For the purpose of item calibration, items not attempted were treated as incorrect. The assessment experience of students with a large number of missing responses has limited validity for use in item analysis, and so we excluded those students' assessment records. For all seven PAT: Reading Comprehension tests, assessment records with more than three missing responses were discarded to ensure reliable estimation of item difficulties. This cut-off point (three missing responses) was decided on by inspecting the distribution of number of questions completed for each test. Less than 5% of assessment records had missing responses for more than three items.

Data description

The final data set used for the item analysis consisted of 188,624 assessment records. Of these, 168,918 tests were done on paper and 19,706 were done online. Assessment records were spread across the seven PAT: Reading Comprehension tests, with slightly greater numbers of assessment records for Tests 4 and 5. The distribution of online assessment records across the test levels was similar to that of paper-based assessment records. There were multiple test records for individual students; however, as the analysis in this section is focused on item properties, it was not necessary to account for the fact that any student might be assessed several times.

Table 1 shows the composition of the data set by the variables gender and ethnic group. The administrative ethnic group variable allowed for student identification with multiple ethnic groups. Accordingly, the percentages of assessment records by ethnic group do not add to 100%. There were no notable differences between paper-based and online records in gender or ethnic group composition.

TABLE 1 **Item analysis data: Assessment records by gender and ethnic group**

Student characteristics	Online records		Paper records		All records	
	n	%	*n*	%	*n*	%
Gender						
Female	10,308	52	83,572	50	93,880	50
Male	9,398	48	83,956	50	93,354	50
Ethnic group						
NZ European/Pākehā	12,413	63	108,249	64	120,662	64
Māori	3,834	19	31,040	18	34,874	18
Pasifika	1,873	10	13,710	8	15,583	8
Asian	2,113	11	16,549	10	18,662	10
Other	1,167	6	6,791	4	7,958	4

Table 2 shows the composition of the data set by the variables decile, school roll and urban category. School roll was grouped to give four categories of size from small to large. This allows for a rough comparison of the distribution of online and paper assessments amongst schools of varying rolls. School decile was grouped into quintiles.

Again, paper-based and online records generally show similar proportions across these variables. Worth noting is the slightly greater proportion of online records from rural schools, when compared with paper records.

TABLE 2 **Item analysis data: Assessment records by decile, roll and urban category**

School characteristics	Online records		Paper records		All records	
	n	%	*n*	%	*n*	%
Decile						
1–2	2,592	13	17,339	10	19,931	11
3–4	2,017	10	16,297	10	18,314	10
5–6	2,547	13	29,344	17	31,891	17
7–8	3,612	18	38,002	23	41,614	22
9–10	8,938	45	67,650	40	76,588	41
School roll						
<101	399	2	4,636	3	5,035	3
101–200	1,337	7	16,247	10	17,584	9
201–350	4,126	21	30,025	18	34,151	18
>350	13,844	70	118,010	70	131,854	70
Urban category						
Main urban	15,251	77	133,495	79	148,746	79
Secondary urban	972	8	10,078	8	11,050	6
Minor urban	1,555	5	14,162	6	15,717	8
Rural	1,927	10	10,093	6	12,020	6

Methodology

Differential Test Functioning (DTF) and Differential Item Functioning (DIF) analyses were carried out to compare the psychometric properties of the PAT: Reading Comprehension assessment administered online with the assessment administered in the paper mode. In DTF, two separate item hierarchies were defined (one for the assessment records in each test mode) and the difficulty of each item was compared by test mode. DIF investigates the items in a test one at a time. We used the item difficulties defined for the online test mode, and investigated items for a relationship with student characteristics.

Analyses were carried out with PAT: Reading Comprehension Tests 1, 4 and 7. These tests were selected to provide coverage of the age range for which PAT: Reading Comprehension is generally used.

Differential Test Functioning: are the tests behaving differently overall?

Each test was rescaled separately with records of assessments completed online and records of assessments completed on paper. This produced two separate sets of item calibrations—one for each mode (online and paper).

Item calibrations were compared across the two modes using scatter plots. Items outside the approximate 95% confidence interval around the line of commonality[3] were investigated (see Figure 3).

Differential Item Functioning: are any of the items behaving differently for any student subgroups, between online and paper?

For each test, student subgroups for the separately scaled paper-based and online assessment modes were compared to see whether any of the items behaved differently between groups. To do this, each item difficulty for each subgroup was estimated while holding all the other item difficulties and student ability measures constant.

The subgroups analysed were student gender and student ethnic group (for ethnic groups Pasifika, Māori, NZ European/Pākeha and Asian). Each ethnic group was compared to all the students who were not identified as part of that group (i.e. all students identified as Māori were compared to all students not identified as Māori).

Items that were easier or harder for one subgroup compared to another for the online test were selected for further investigation, if that difference was not also apparent in the paper-based test mode.

Results

Differential Test Functioning

DTF analysis was carried out for PAT: Reading Comprehension Tests 1, 4 and 7. The results were very similar across the three tests, with no notable differences between modes in any case. The results for Test 4 are presented for illustration (see Figure 3).

Figure 3 displays the item difficulties for paper-based and online assessment modes. Each individual point represents an item; each item is plotted as the question number given to that item in the test. The middle dotted line is a best-fit line through the mean of both sets of items, and the two lines on either side form an approximate 95% confidence interval (approximate because each point has its own confidence interval). The items in red show where there is a significant[4] difference between the difficulty

3 A line of commonality is a trend line through the item difficulties for PAT: Reading Comprehension in the paper and online mode. This line is equally good at predicting the item difficulties for online from paper, and item difficulties for paper from online (Linacre, 2012b).

4 An alpha level of 0.001 was used for these significance tests.

for that item in the paper test mode compared to the difficulty for that item in the online test mode. The relationship between the item difficulties for the online and paper-based tests is linear, with a correlation of 0.97.

The position of an item on the horizontal axis represents the item difficulty for the test taken online, and the position of an item on the vertical axis represents the item difficulty for the test taken on paper. Items that fall below the trend line are items that were relatively more difficult in the online test, and items that fall above the trend line are items that were easier in the online test, when compared with the paper-based test.

The differences between item difficulties for the paper and online item calibrations were calculated. The median absolute difference in item difficulty between the two modes was 0.11 logits[5]; a difference of this size would not be significant.

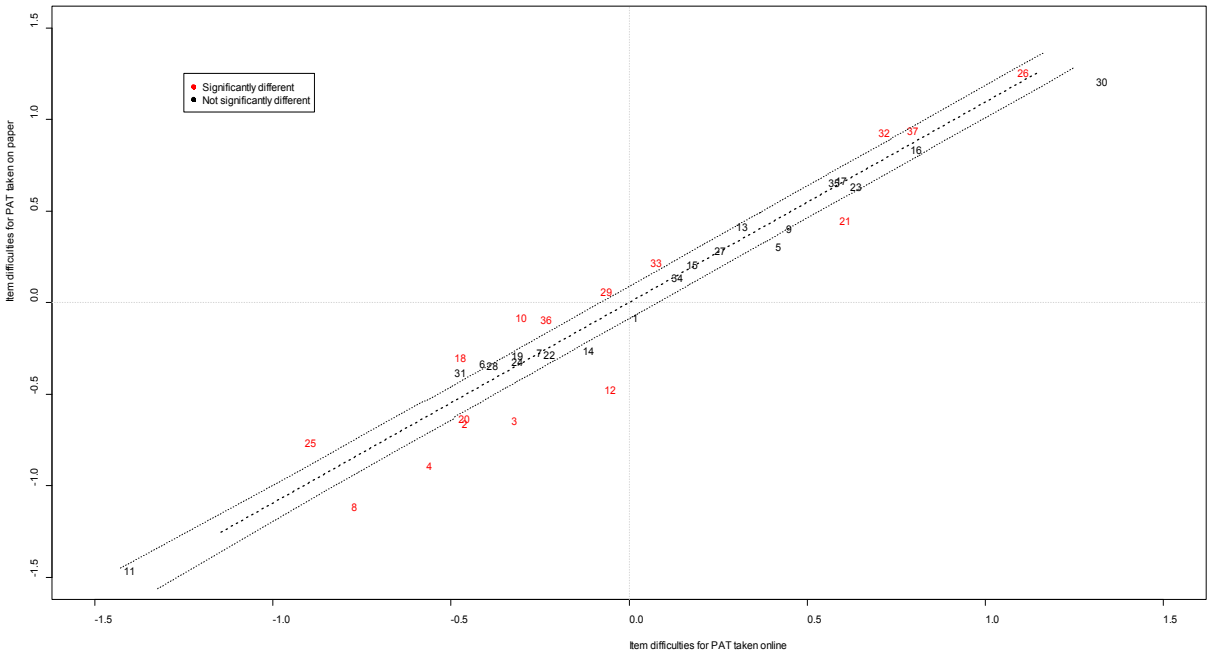
The maximum difference in item difficulty between the modes was 0.4 logits, with item 12 being 0.4 logits more difficult in the online test than in the paper-based test. In practice, this means that if a student had a 50% chance of answering item 12 correctly in the online test, the same student would have a 60 percent of answering the same item correctly in the paper-based test.

Items that were significantly more or less difficult for online test-takers compared to paper test-takers were investigated for item bias: that is, differences in item difficulties that could be explained by a consistent difference in the way assessment items are presented and interacted with online compared to paper assessments.

The possibility of item bias was explored by looking at possible sources of difference between the difficulty of items assessed in online and paper-based versions of the test. No consistent patterns were found. For example, some questions that required scrolling on screen were slightly easier online than on paper, and some were slightly more difficult. The same was true for other possible sources of difference that we investigated, such as question type (inference, retrieval etc.) and differences in layout between the two modes. It was not possible to find a clear explanation for the small differences in difficulty of these items between modes.

5 A logit, or a log-odds unit, is the mathematical unit of Rasch measurement. It is used to describe both person ability and item difficulty.

FIGURE 3 **Item difficulties for PAT: Reading Comprehension Test 4 by paper and online mode**



Differential Item Functioning

DIF analysis was carried out for PAT: Reading Comprehension Tests 1, 4 and 7 to see if any items behaved differently for any subgroups of students. As with the DTF analysis, the results for the three tests analysed were similar. There were no notable differences in item functioning by subgroup found for any of the three tests. Results for the DIF analysis by gender for Test 4 are shown for illustration (Figure 4).

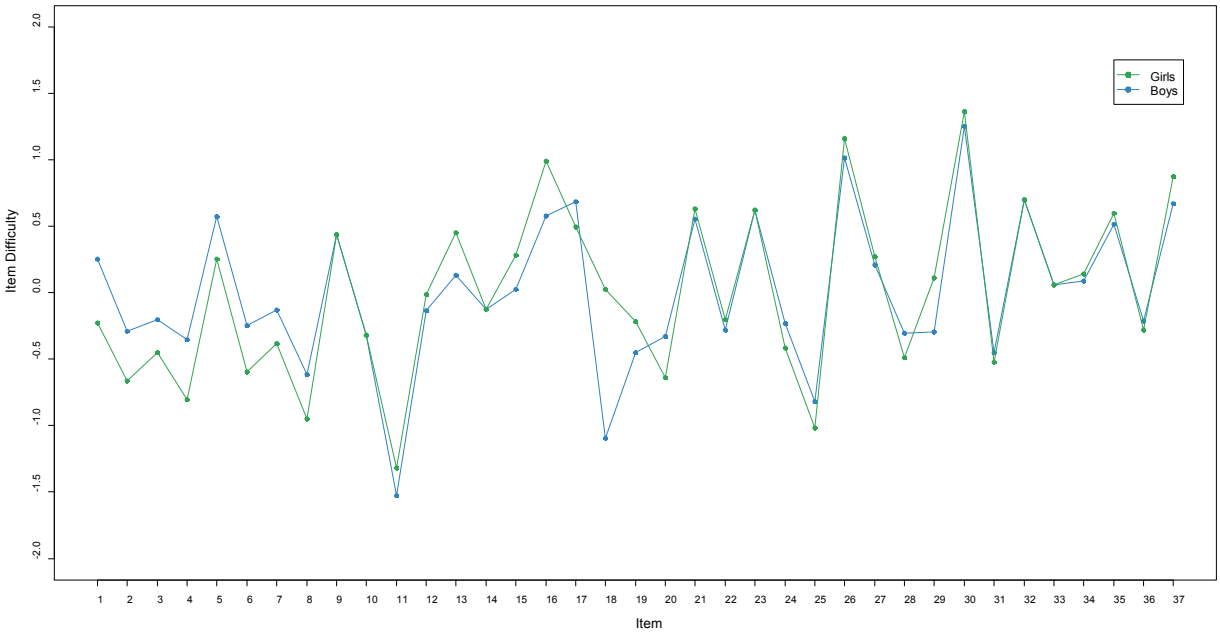
Contrasts between items for boys and girls in the online test mode were compared to the same contrasts for the paper-based test, in order to determine whether the test mode was having an impact on differential item functioning. These contrasts were very similar to one another, indicating that mode did not have an effect on differential item functioning by gender.

The same process was undertaken to look for any effect of mode in differential item functioning by ethnic group. Again, no significant differences were found between the online and paper-based assessments.

Figure 4 shows the difficulty for each item in Test 4, for girls and boys, when test mode is online; the horizontal axis is numbered by test item in the order it was given in the test, and the vertical axis is the item difficulty in logits. The blue line shows the item difficulties for girls, and the green line for boys. Item 18 stands out, being 1.12 logits more difficult for girls than for boys. However, this difference was seen similarly in the DIF analysis for gender in the paper mode and therefore cannot be attributed to assessment mode.

Overall, in both modes some items appear to be more difficult for girls, while others appear to be more difficult for boys, with no indication of a consistent gender-based advantage either way.

FIGURE 4 **Differential item functioning for online assessment mode: Comparing item difficulties by gender**



PAT: Reading Comprehension score analysis

Data

Student matching

While the presence of multiple assessment records for one student was not accounted for in the item analysis, it was considered important in the analysis of student scores. The results of assessments completed by the same student are not independent, and failure to take this into account increases the probability of erroneously detecting a difference. In many cases, these multiple records were easily matched using the National Student Number (NSN). However, in some cases student NSN data were missing. Extensive matching of records, using student name together with other variables such as year level, was carried out with 2014 and 2015 PAT: Reading Comprehension data for the purposes of the Depicting Learners' Progress project which ran concurrently to this one, resulting in the matching of further assessment records with individual students.

These data were cleaned for consistent student gender and ethnic group records across assessment instances. Students with inconsistent gender were assigned either the gender most commonly recorded or, in the case of an equal number of both genders recorded, assigned a gender randomly (this was for a very small number of students). No gender was assigned in the case of missing gender for all assessment records associated with an individual. A student was assigned all ethnic groups they had ever identified with: that is, if a student had one assessment record where the 'Other' ethnic group was indicated and two assessment records where the 'Asian' ethnic group was indicated, all assessment records for that student had both Asian and Other ethnic groups indicated.

Other exclusions

A small number of records were discarded due to inconsistent year level and calendar year combinations. Where a student had more than one assessment record at the same time (within the same school term) the average of those records was taken. Note that both records were kept if the tests were in different modes, or if the records were at different schools.

Data description

The final data set used for the analysis of student scores consisted of 177,437 assessment records, 157,288 of which were from paper-based tests and 20,149 from online tests. Overall, the characteristics of the data set by assessment, student and school level variables were very similar to that of the data set used for item analysis (see Table 1 and Table 2). Unlike the item analysis, however, the student score analysis took into account the hierarchical structure of the data: each assessment instance occurring for a particular student at a particular school. The 177,437 assessment records that comprised this data set were identified as belonging to 128,764 students in 700 schools. Table 3 and Table 4 show the characteristics of these students and schools. School decile and roll have been grouped as described below Table 1 (see p. 9).

TABLE 3 **Score analysis data: Students by gender and ethnic group**

Student characteristics	Percentage of students
Gender	
Female	50
Male	50
Ethnic group	
NZ European/Pākehā	49
Māori	14
Pasifika	6
Asian	8
Other	5

TABLE 4 **Score analysis data: Students and schools by decile, roll and urban category**

School characteristics	Percentage of students	Percentage of schools
Decile		
1–2	10	14
3–4	10	13
5–6	16	19
7–8	24	22
9–10	41	32
School roll		
<101	3	14
101–200	9	22
201–350	18	23
>350	70	40
Urban category		
Main urban	80	65
Secondary urban	6	7
Minor urban	8	10
Rural	6	18

Methodology

The PAT: Reading Comprehension data have a multilevel hierarchichal structure where the levels consist of assessment instances, students and schools. Assessment instances are grouped within students, which are in turn grouped within schools. Assessment instances are referred to as level 1 of the data, students are referred to as level 2 of the data and schools are referred to as level 3 of the data. Furthermore, the multilevel structure of the PAT: Reading Comprehension data is 'cross-classified'. That is, each assessment instance is associated with one student and one school, but individual students can be associated with different schools. An assessment instance is linked to a school through the student who completed it.

A multilevel model allows us to describe PAT: Reading Comprehension scores as a function of other variables, where some variables vary for one level of the data (such as gender, for individual students) and other variables for another level of the data (such as decile, for different schools). A linear mixed model is an extension to an ordinary linear model with the addition of *random effects*. We add random effects to account for the dependence in PAT: Reading Comprehension scores due to the data's hierarchical structure: random effects account for the variability in test scores attributable to individual students and the variability in test scores due to the school the student sat the test in. In other words, we are taking account of variability in PAT: Reading Comprehension scores between students (or within schools), and the variability in test scores between schools. By accounting for dependencies to the hierarchical structure of the data, we reduce the risk of a Type I error (i.e. finding spurious associations in our data).

Analysis for an effect of assessment mode on PAT: Reading Comprehension score

Initial exploration of this data consisted of basic summaries and graphs of all available data. These were used to indicate relationships worth including in the model-fitting process. This exploration revealed potential differences in PAT: Reading Comprehension score between assessment modes by student ethnic group, school decile and school location by urban categories.

The level 1 equation[6] for our model was:

$$patc = \beta_0 + \beta_1 age + \beta_2 mode + e$$

Here age is a combination of student year level and the school term that the assessment took place in. Although age is a student characteristic, it is a variable at the assessment level as it can change with subsequent assessments a student sits. The unit for PAT: Reading Comprehension score is *patc*.

At level 2 our model had equations:

$$\beta_0 = \gamma_{00} + \gamma_{01}\text{gender} + \gamma_{02}\text{Māori} + \gamma_{03}\text{Pasifika} + \gamma_{04}\text{Asian} + \gamma_{05}\text{Other} + u_0$$

$$\beta_1 = \gamma_{10} + \gamma_{11k}\text{gender} + \gamma_{12}\text{Māori} + \gamma_{13}\text{Pasifika} + \gamma_{14}\text{Asian} + \gamma_{15}\text{Other}$$

$$\beta_2 = \gamma_{20} + \gamma_{21k}\text{gender} + \gamma_{22}\text{Māori} + \gamma_{23}\text{Pasifika} + \gamma_{24}\text{Asian} + \gamma_{25}\text{Other}$$

At level 3 our model had equations:

$$\gamma_{00} = \delta_{000} + \delta_{001}\text{Quint2} + \delta_{002}\text{Quint3} + \delta_{003}\text{Quint4} + \delta_{004}\text{Quint5} +$$

$$+ \delta_{005}\text{Secondary} + \delta_{006}\text{Minor} + \delta_{007}\text{Rural} + v_{00}$$

$$\gamma_{10} = \delta_{100} + \delta_{101}\text{Quint2} + \delta_{102}\text{Quint3} + \delta_{103}\text{Quint4} + \delta_{104}\text{Quint5} +$$

$$+ \delta_{105}\text{Secondary} + \delta_{106}\text{Minor} + \delta_{107}\text{Rural}$$

Each coefficient can be interpreted in terms of the corresponding variable (or category name) written with it. For example, β_1 represents the expected change in PAT: Reading Comprehension score as a student gets older by 1 year; γ_{01} represents the expected differences in PAT: Reading Comprehension score associated with a student being male compared to female; and δ_{004} represents the expected difference in PAT: Reading Comprehension score associated with a school being 'quintile 5' (decile 9 or 10), compared with being 'quintile 1' (decile 1 or 2).

6 Note that the equations describing our model include every possible interaction between the variables included, whereas the final model included only the interaction terms seen in Table 5. The equations were written this way for clarity.

Results

We fitted a three-level multilevel mixed model to describe PAT: Reading Comprehension score for paper and online tests with:

At the test level

Age: a combination of student year level and school term the test took place in.

Mode: test mode.

At the student level

Gender: student gender.

Ethnicity: student ethnic group: Māori, Pasifika, Asian or Other.

At the school level

Quint: school decile, grouped into quintiles (i.e. quintile 1 is deciles 1 and 2).

Urban: school location in a main urban, secondary urban, minor urban or rural setting.

As our research question was aimed at explaining potential differences in PAT: Reading Comprehension score due to test mode, in the model-building process we focused on interactions between test mode and higher level variables. We found no association between assessment score and school roll. Table 5 shows the results of our final model fit. The table is split into *fixed* effects and *random* effects, and within fixed effects into test-specific, student-specific and school-specific variables. The intercept is the overall average score for female New Zealand European/Pākehā students with paper test mode, in a quintile 1 school in a main urban location. Below the intercept, each value in the 'estimate' column is the effect of every variable in PAT: Reading Comprehension scale score points. Similarly to the interpretation of the intercept estimate, coefficients are interpreted with respect to the 'baseline' value of the variables in the model.

TABLE 5 **Parameter estimates for model**

Effect	Estimate	SE	Sig.
Fixed effects			
Intercept	27.58	0.401	***
Level 1 (test-specific)			
Year level	7.92	0.0269	***
Mode (online)	-4.13	0.369	***
Level 2 (student-specific)			
Gender (boys)	-3.31	0.0695	***
Māori	-2.86	0.272	***
Pasifika	-2.62	0.308	***
Asian	1.45	0.484	**
Other	-2.81	0.684	***
Level 3 (school-specific)			
Deciles 3–4	4.42	0.521	***
Deciles 5–6	6.20	0.480	***
Deciles 7–8	8.04	0.466	***
Deciles 9–10	10.45	0.442	***
Secondary urban	-1.39	0.468	**
Minor urban	-0.76	0.408	
Rural	-1.05	0.336	**

Effect	Estimate	SE	Sig.
Cross-level interactions			
Mode x Māori	-0.63	0.184	***
Mode x Other	0.69	0.261	**
Mode x Deciles 3–4	1.84	0.432	***
Mode x Deciles 5–6	1.40	0.423	***
Mode x Deciles 7–8	2.71	0.423	***
Mode x Deciles 9–10	1.50	0.382	***
Māori x Deciles 3–4	-2.01	0.351	***
Māori x Deciles 5–6	-2.33	0.332	***
Māori x Deciles 7–8	-1.17	0.331	***
Māori x Deciles 9–10	-0.46	0.326	
Pasifika x Deciles 3–4	-2.26	0.464	***
Pasifika x Deciles 5–6	-3.10	0.452	***
Pasifika x Deciles 7–8	-3.07	0.441	***
Pasifika x Deciles 9–10	-3.16	0.416	***
Asian x Deciles 3–4	-1.49	0.648	*
Asian x Deciles 5–6	-2.63	0.577	***
Asian x Deciles 7–8	-2.98	0.543	***
Asian x Deciles 9–10	-2.52	0.508	***
Other x Deciles 3–4	2.11	0.884	*
Other x Deciles 5–6	2.50	0.787	**
Other x Deciles 7–8	1.84	0.744	*
Other x Deciles 9–10	1.50	0.708	*
Random effects			
School-level variance	7.80	-	-
Student-level variance	107.56	-	-
Residual variance	37.13	-	-

*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Interactions between test mode and student-level variables were explored to see whether the effect of mode differed for different groups of students. Likewise, interactions between test mode and school-level variables were explored to see if test mode differed for different types of schools.

The interaction between students identifying as Māori or Other ethnic groups with test mode was significantly associated with PAT: Reading Comprehension score. However, the effect for both ethnicity groups was small (less than one PAT: Reading Comprehension scale score point).

There was a significant interaction between test mode and school decile. This tells us that the effect of doing a PAT: Reading Comprehension test online compared to paper is not the same for students at decile 1 and 2 schools compared to schools of different deciles.

On average, the overall effect of sitting a test online compared to paper is –4.13 scale score points. That is, the expected score of a test taken online is 4.13 scale score points lower than the expected score of a test

taken on paper. However, this interpretation is not straightforward due to the presence of interactions—the moderating effects of students identifying as Māori or Other, and the decile of a student's school.

As an example of how to interpret the estimates in Table 5, the model predicts an expected PAT: Reading Comprehension scale score for an Asian girl at the start of Year 4 sitting a test on paper in a decile 4 main urban school:

$$27.6 + 1.4 + 4.4 - 1.5 = 31.9\, patc$$

This can be compared to the average score for a student with the same characteristics sitting a test online:

$$27.6 + 1.4 + 4.4 - 1.5 = 31.9\, patc$$

The first number in bold is the effect of the test being online compared to paper. Note the second bold number—the addition of almost 2 scale score points for the interaction between test mode and school decile (where decile is 3–4). This is an example of the overall effect of test mode not being applicable due to the moderating effect of another variable: school decile.

Figure 5 shows the difference between expected test score when the PAT: Reading Comprehension test is taken online compared to on paper, for students with four different sets of characteristics. The estimated average score differences are:

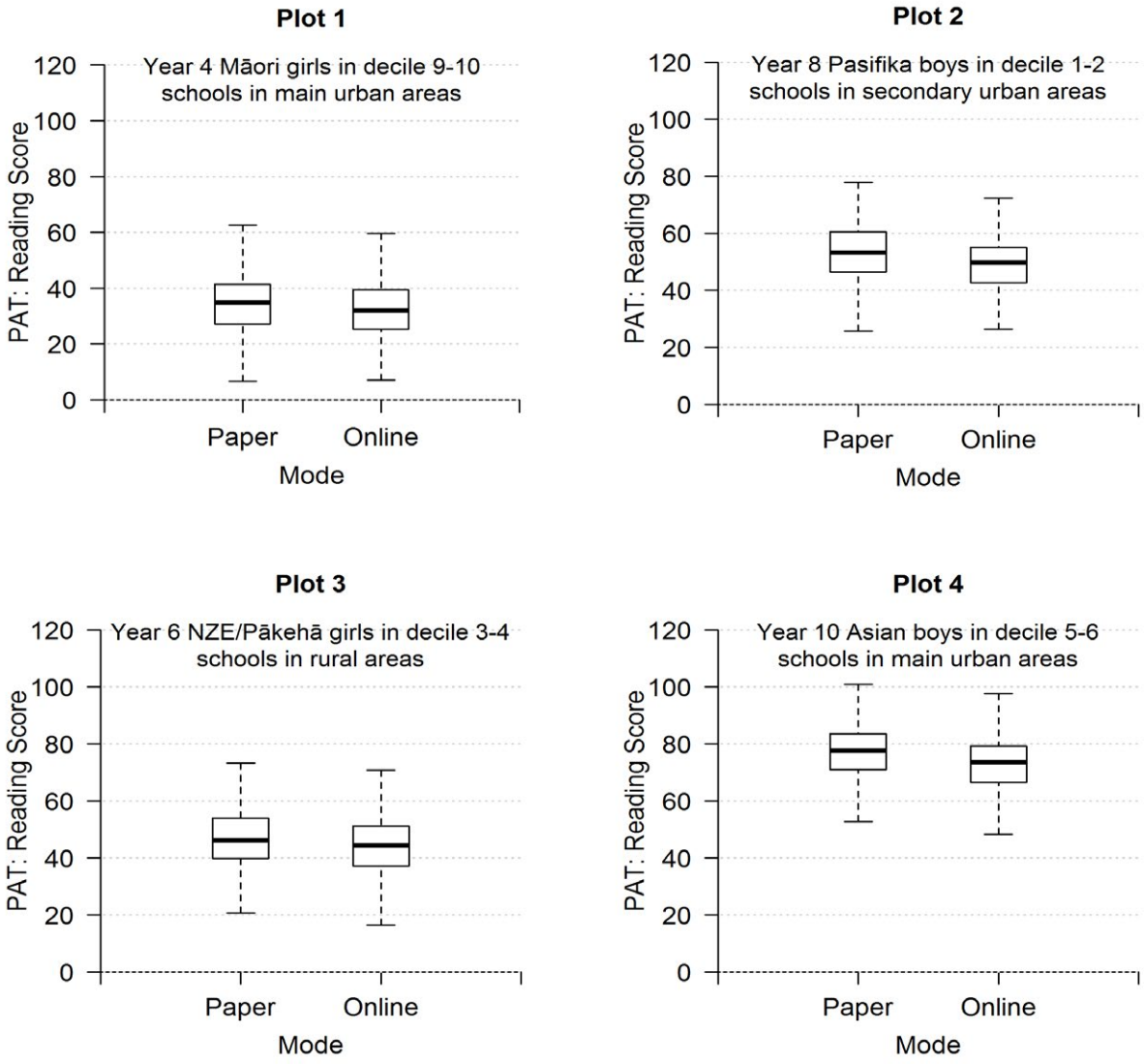
Plot 1: 3.3 scale score points higher for paper than for online

Plot 2: 4.1 scale score points higher for paper than for online

Plot 3: 2.3 scale score points higher for paper than for online

Plot 4: 2.7 scale score points higher for paper than for online.

FIGURE 5 **Example distributions of PAT: Reading Comprehension scores by mode**



3. Discussion

Our primary purpose in this study was to compare results from online and paper-based versions of PAT: Reading Comprehension in order to investigate equivalence between the two modes. There were two main findings:

1. There was very little difference between how items behaved online and how they behaved on paper. That is, the relative difficulties of items (their locations on the scale) stayed the same, regardless of whether the student read and answered the items online or on paper.
2. While the difficulty level of items relative to each other remained constant across modes, students taking the assessment online scored lower on average than those who took the assessment on paper.

Individual items behaved similarly in either mode

The finding that individual items behaved similarly in either mode is consistent with studies which have found that careful design of online items minimises differences in item properties between online and paper-based modes (Schroeders & Wilhelm, 2011). The literature tells us that two factors with the potential to create differences in the computer-based mode are question type and amount of scrolling. Controlling these factors is therefore likely to minimise mode effects.

Question type

The more sophisticated the question type in terms of its computer use, the more likely it is that the computer-based test will present extra requirements for the test-taker. For example, entering text via a keyboard is likely to place greater demands on the test-taker than selecting the correct answer from a list. PAT: Reading Comprehension uses only multichoice questions, and answering the questions online requires a simple point-and-click response. This simple format is likely to have contributed to the consistency in item behaviour across modes.

Scrolling

The amount of scrolling required to read texts online has also been found to influence equivalence of computer and paper-based assessments (Bridgeman, Lennon, & Jackenthal, 2003, cited in Bridgeman, 2009). PAT: Reading Comprehension was carefully designed to minimise the effects of scrolling.

Although some texts in the PAT: Reading Comprehension tests do require students to scroll down to read the whole text in the standard view, the option to switch to an alternative view (by clicking on an 'eye' icon) largely eliminates this need (one or two of the longest texts still require minimal scrolling in the alternative view). It is likely that this design also contributed to the close match between the behaviours of individual items in either mode.

Analysis of small differences

In our analysis of the small differences in relative difficulty of some items across modes, we found no evidence that scrolling was a factor. Some items that required scrolling were slightly more difficult in the online mode, but some were slightly easier. There was also no evidence that other design-related factors were related to these differences (such as position of the question in relation to the text, or differences in layout features such as line length of text). Question content (retrieval, local inference or global inference) was also unrelated to the differences.

These findings suggest that the careful design of the computer-based display was successful in ensuring that students found the same items difficult or easy, in the same patterns, in either mode (that is, that the items discriminated student performance in similar ways in either mode).

Students assessed online scored lower than those assessed on paper

Our use of statistical modelling enabled us to partially control for the possibility that the group of students taking the assessment online were systematically different from those taking the assessment on paper. Therefore, the second finding, that students' overall scores were lower on average in the online mode, suggests that there was something about the *experience* of taking the assessment online that interfered with students' reading comprehension. We can speculate on the factors that might have contributed to this, drawing on existing research. Possible contributing factors include:

- cognitive load
- online reading behaviour
- influence of multimodal, hyperlinked, interactive online text
- mismatch between classroom and assessment experience.

Cognitive load

A number of studies have found that reading on screen adversely affects comprehension when compared with reading on paper. For example, Wästlund, Reinikka, Norlander, and Archer (2005, p. 389) found that:

> The consumption of information, measured by a test of reading comprehension, is more difficult when the assignment is presented on a VDT [Visual Display Terminal] than upon paper. VDT presentation led to fewer correct responses, to a greater level of experienced tiredness and an increased feeling of stress.

Wästlund et al. (2005) theorised that reading on computer involves a greater cognitive load than reading on paper, as the reader must both comprehend the text and also cope with the demands of the computer interface. This results in a 'dual task' situation, where increased cognitive demands also result in increased tiredness.

With advances in computer technology and in students' levels of computer experience, the 'dual task' effect as reported by Wästlund et al. (2005) may be diminishing. However, others have speculated that increased exposure to online text has brought further potential for 'dual task' effects. Handling the computer equipment may no longer be such a problem, but familiar and habitual ways of reading online, which are different from ways of reading on paper, have the potential to interfere with traditional assessments of reading comprehension.

Online reading behaviour

A range of studies have shown that we read screen-based text in different ways than we read text presented on paper. In a discussion of issues associated with reading digital text, Mangen (2008) suggests that the 'intangibility' of digital text (compared to the tactile experience of handling a physical, paper-based text) contributes to "making us read in a shallower, less focused way" (p. 408). She supports this by referring to a number of studies that show that "we tend to *scan* text on screen" rather than reading word-by-word (p. 409). This notion of scanning aligns with evidence from eye-tracking studies reported by Neilsen (2006). These show that when reading a webpage, we typically scan the page in an 'F' shaped pattern. That is, we fixate on the top-left portion of the screen, move across to the right, move down the left-hand side, scan to the right again, then move down to the bottom of the screen. Large areas of the screen thus receive minimal attention.

Influence of multimodal, hyperlinked, interactive text

Furthermore, when we engage with digital technology in our daily lives, we habitually encounter texts that include features such as hyperlinks. These interactive features allow us to move quickly between content and constantly change what we see on the screen. Most websites, for example, allow us to follow our own path by clicking on links. Mangen (2008) suggests that the possibility of interactivity can distract us and lead to superficial reading. When our attention begins to wander, we seek to rekindle it by moving to a new area of the site: "a click with the mouse immediately changes the visual input so that attentional focus can be maintained" (p. 410). It is possible that this expectation of interactivity transfers to any online reading experience. The experience of hopping from screen to screen when reading online means that our attention is divided when reading large blocks of text such as those in PAT: Reading Comprehension—we are always looking to move on to the next click, rather than being deeply immersed in the text before us.

Mismatch between classroom and assessment experience

Interactive technology can be deeply engaging, and the majority of students prefer computer-based to paper-based assessment (Barnes, 2010; Darr, 2014; Noyes & Garland, 2008). However, it may be that there is a mismatch between the type of reading being assessed and the use of the computer to assess it. Much of our daily online reading consists of short passages of information that are suitable for skimming and scanning (e.g. websites or social media posts). It is fair to say that many of us prefer to read longer texts that require focused attention on paper. If students usually read the linear expository, narrative or literary texts that are typically used in tests of reading comprehension in a print-based form, there is a mismatch between their classroom experience and their assessment experience when reading comprehension is assessed online.

While acknowledging the affordances of technology for other types of immersive experiences such as computer simulations and games, Mangen (2008) concludes that "the computer, as a reading device, seems to be poorly suited for the contemplative and deeply focused reading we associate with the book" (p. 410). It is possible that the students who completed PAT: Reading Comprehension online had trouble reading the passages of text in a focused way, and applied similar strategies of skimming and scanning that they might use for other online media. It may also be that the interactivity offered by the digital platform affected the way in which they engaged with the texts. The ever-present possibility of clicking on to a new screen or changing the view of the text might have contributed to a scattering of attention, which is at odds with focused and deep reading. This in turn could have contributed to increased cognitive load and impaired reading comprehension.

4.
Conclusion

The results of this study suggest that for PAT: Reading Comprehension, the mode in which the assessment is taken affects the distribution of scores. When linear passages of text are presented in PAT: Reading Comprehension online, comprehension of them is lower than when they are presented on paper. This finding has several implications for assessment design and for the use of technology for assessment purposes.

Firstly, the results suggest that possible effects on student achievement must be considered when moving assessments of reading comprehension to an online format. This is particularly pertinent when assessments are offered in dual mode, where there is the potential for students to be disadvantaged if they complete the assessment in the computer-based mode. It is also important to recognise that students who have previously completed their assessments in a paper-based mode might achieve at a lower level when they move to a computer-based assessment.

Secondly, the results suggest that the digital environment presents challenges for traditional assessments of reading comprehension. The distractions associated with the online mode may contribute to a greater cognitive load and impaired comprehension. It is important to consider whether the impact of these distractions can or should be minimised: for example, by using short blocks of text with fewer questions associated with each block. This would allow learners to move more quickly through the assessment in a manner that more closely resembles typical online reading behaviour. That is, assessment developers could make the passages of text resemble the kinds of texts that we normally read online, and thus replicate online reading behaviour more closely. However, this changes the focus of the construct being assessed from the ability to comprehend longer passages of text to the ability to comprehend short blocks of text.

Technology has enabled new forms of literacy, including online and multimodal texts. There is growing interest in how these 'new literacies' may be integrated with traditional literacy practices in the classroom (Lankshear & Knobel, 2006). There is a need for discussion on ways of assessing these new literacies. If both traditional linear texts and new multimodal texts are valued in education, appropriate ways of assessing them need to be found. It may be that an online platform is particularly suited to the assessment of comprehension of online, multimodal texts, while paper-based modes are more suited to assessment of comprehension of traditional linear texts. This is an area for further research, discussion and reflection.

In the meantime, while assessments of reading comprehension are offered in dual mode, it is important to realise that students' results may be affected by the mode in which they completed the assessment.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi: 10.18637/jss.v067.i01

Barnes, S. K. (2010). *Using computer-based testing with young children. Northeastern Educational Research Association (NERA) Conference Proceedings 2010*, paper 22. Retrieved from http://digitalcommons.uconn.edu/nera_2010/

Bridgeman, B. (2009). Experiences from large-scale computer-based testing in the USA. In F. Scheuermann & J. Bjornsson (Eds.), *The transition to computer-based assessment* (pp. 39–44). Luxembourg: Office for Official Publications of the European Communities.

Darr, C. (2014). Computer-administered vs paper-and-pencil tests: Is there a difference? *set: Research Information for Teachers, 3*, 61–64.

Eyre, J. (2015). Dual modes for written assessments: Examining validity and reliability. *Assessment Matters, 9*, 4–24.

Herold, B. (2016, February 10). PARCC scores lower on computer exam: Discrepancy raises questions about fairness. *Education Week*, p. 1.

Kerr, M. A., & Symons, S. E. (2006). Computerized presentation of text: Effects on children's reading of informational material. *Reading and Writing, 19*, 1–19. doi: 10.1007/s11145-003-8128-y

Kim, H. J., & Kim, J. (2013). Reading from an LCD monitor versus paper: Teenagers' reading performance. *International Journal of Research Studies in Educational Technology, 2*(1), 15–24.

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. doi: 10.1080/08957340802558326

Lankshear, C., & Knobel, M. (2006). *New literacies: Everyday practices and classroom learning* (2nd ed.). Maidenhead, UK: Open University Press.

Linacre, J. M. (2012a). Winsteps® (Version 3.75.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/

Linacre, J. M. (2012b). *Winsteps Rasch tutorials: Tutorial 4 differential item functioning and dimensionality*. Retrieved from http://www.winsteps.com/tutorials/

Mangen, A. (2008). Hypertext fiction reading: Haptics and immersion. *Journal of Research in Reading, 31*(4), 404–419. doi: 10.1111/j.1467-9817.2008.00380.x

Mangen, A., Walgermo, B. R., & Bronnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Education Research*, *58*, 61–68.

Neilsen, J. (2006). *F-shaped pattern for reading web content*. Nielsen Norman Group. Retrieved from https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics Retrieved, 51*(9), 1352–1375.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Revised and expanded edition. Chicago: The University of Chicago Press (original work published 1960).

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*(5), 849–869. doi: 10.1177/0013164410391468

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 Behaviour, 21,* 377–394.

White, S., Kim, Y. Y., Chen, J., & Liu, F. (2015). *Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment: Scores, text length, and use of editing tools*. National Center for Educational Statistics. Retrieved from http://nces.ed.gov/nationsreportcard/subject/writing/pdf/2015119.pdf

NZCER
Rangahau Mātauranga o Aotearoa

New Zealand Council for Educational Research
facebook.com/nzcer
@NZCER
www.linkedin.com/NZCER
www.nzcer.org.nz