Are writing scores from online writing tests for primary students comparable to those from paper tests?

Centre for Education Statistics and Evaluation

September 2021

Centre for Education Statistics and Evaluation

The Centre for Education Statistics and Evaluation (CESE) undertakes in-depth analysis of education programs and outcomes across early childhood, school, training, and higher education to inform whole-of-government, evidence-based decision making. Put simply, it seeks to find out what works best. Its focus is all education in NSW. CESE's main responsibilities within the department are:

- qualitative and quantitative research, including data analysis, evidence papers and case studies that build understanding and uptake of evidence-based practice
- robust data collection to enable research and statistics for the education and training sector
- evaluating key policies and programs to strengthen quality delivery and student outcomes
- national engagement on research agenda and data strategy
- driving capability uplift in the use of data and evidence as part of everyday practice
- trialling innovative initiatives to improve student outcomes.

Authors

Lucy Lu, Margaret Turnbull, Wai-Yin Wan, Karen Rickard, Linda Hamilton
Centre for Education Statistics and Evaluation, September 2021, Sydney, NSW

Please cite this publication as:
Centre for Education Statistics and Evaluation (2021), **Are writing scores from online writing tests for primary students comparable to those from paper tests?** NSW Department of Education, education.nsw.gov.au/cese

For more information about this report, please contact:

Centre for Education Statistics and Evaluation
Department of Education
GPO Box 33
Sydney NSW 2001

info@cese.nsw.gov.au

+61 2 7814 1527
education.nsw.gov.au/cese

Acknowledgements

CESE would like to thank those who have contributed to the research, particularly the school principals, teachers and students who took the time to participate. Thanks also go to Goran Lazendic at ACARA for providing advice, and to Peter Siminski, Associate Professor in Economics at the University of Technology, Sydney, for reviewing the report and providing technical advice on the analysis conducted.

We acknowledge the homelands of all Aboriginal people and pay our respect to Country.

Table of contents

Executive summary	6
Introduction	8
Literature review	10
Comparability of test results from computer-based and paper-based tests	10
Cognitive load theory and writing performance	11
Other aspects of keyboarding that impact writing performance	12
Delivery of typing instruction	12
Impact of technology and assessment mode on teaching practice	13
Validity and equity issues	14
Research questions	14
Methodology	15
Sampling of schools and students	15
Writing genres and tasks	15
Test administration	16
Marking	17
Typing test and student survey	17
Teacher and student observations and interviews	17
Data treatment and variables used in analysis	18
Sampling weights	19
Analysis methods	19
Results and discussion	21
Summary of findings	37
Conclusions and recommendations	38
References	40

Appendix 1 – Technical details for MFRM 43

Appendix 2 – Technical details for mixed effects model for repeated measures 44

Appendix 3 – Test instructions relating to planning and editing, and an example of planning sheets 47

List of figures

Figure 1
Scatterplot of computer-based versus paper-based writing scores 22

Figure 2
Boxplots indicating Year 3 median typing speed, interquartile range and minimum/maximum typing speed, for schools with five or more Year 3 students participating in the study 28

Figure 3
Scatterplots of typing speed and computer-based writing scale scores for each year level 29

List of tables

Table 1
Sampled student demographics by genre of writing tasks undertaken 16

Table 2
Writing task mean scaled scores with standard errors (SE) (sampling weights applied) 21

Table 3
Mean differences in scaled scores (computer-based score – paper-based test score) and paired sample t-test results (sampling weights applied) 22

Table 4
Fit statistics and difficulty estimates for measurement facets from FACETS analysis 24

Table 5
Results from the mixed effects model for repeated measures 25

Table 6
Typing speed (words per minute) and typing accuracy rate by year level (sampling weights applied) 27

Table 7
Results from the mixed effects model for repeated measures including typing speed and accuracy 30

Table 8
Median typing speed (and interquartile range) by year level, gender and Aboriginal status (reduced sample, sampling weights applied) 32

Table 9
Estimated differences in writing scaled scores between computer- and paper-based tests for an average ability student in each year level 33

Executive summary

The move to online testing for NAPLAN in 2019 brought many benefits to teachers, schools and education systems. However, prior to implementation, concerns were raised amongst stakeholders relating to the validity, comparability, equity and fairness of online testing. One key issue was the online assessment of writing, for Year 3 students in particular. The main concern was whether Year 3 students would have sufficient typing skills to produce online texts in timed conditions that were a valid reflection of their underlying writing proficiency. This report on the research conducted in 2016 was used to inform decision making concerning the move of NAPLAN writing tests from paper to online tests.

Informed by a literature review of research into online assessment of writing, this study used a mixed methods approach to investigate whether primary students in NSW schools perform differently according to the mode of writing test (computer-based versus pen and paper based), and if there is a difference, whether it is uniform across different groups of students. In addition, the study examined the extent to which typing proficiency accounts for any differences observed in students' performance in a computer-based writing test versus in a pen and paper test.

Whilst some students performed better and some performed worse on the computer-based test than the paper-based test, statistical analysis indicated that:

- after holding constant the effects of various task, student and school level factors on students' writing performance, on average, students scored lower on computer-based writing tests than they did on paper-based tests, across all years examined (Years 2-5), and for both narrative and persuasive writing genres.
- for an average student, the estimated gap between paper-based and computer-based test results varied from 15 to 20 NAPLAN scaled points, depending on the student's scholastic year. This represents roughly 0.2 to 0.3 of one standard deviation in writing results.
- over above other factors, typing proficiency was significantly associated with the mode effect, such that the faster students could type, the smaller the difference between their computer-based and paper-based results. For an increase of 5 words per minute in typing speed, the gap is reduced by approximately 7.5 scaled points.
- after taking into account all other factors including typing, the mode effect was found to be slightly worse for students with higher literacy ability than those with lower ability. The mode effect was approximately 5 scaled points larger for students whose literacy level was one standard deviation above the average for their year.
- the mode effect also appeared to be smaller (in size) for boys than for girls, and for Aboriginal than for non-Aboriginal students, although the differences did not reach statistical significance. Larger studies are needed to confirm any differential mode effect for these demographic groups.

In addition, qualitative analyses of researcher observations and teacher and student interview responses indicate that:

- students responded positively to using computers but there was a disconnect between the mode students said they preferred and the mode that best supported students' performance in writing tests.
- students appeared to undertake less planning for the computer-based writing test compared to the paper-based test.

- there was considerable variability in technical readiness from school to school. A number of schools, particularly some in low SES areas, had an insufficient number of working computers, limited technical support available, and students who were not as familiar with using computers and accessing the internet. These issues are currently being addressed through the NSW NAPLAN online transition program, and are therefore expected to diminish over time.
- most trial schools do not explicitly teach keyboarding skills.

For Year 3 students, the study found that the median typing speed was 9 words per minute. The literature suggests this is lower than the handwriting speed for this age group, hence it is likely that many Year 3 students would struggle to produce online texts comparable to handwritten texts in a timed condition. Most research recommends that typing instruction is best commenced in the upper primary years, and there is evidence of cognitive and educational benefits of teaching handwriting to students in early learning years. For these reasons it is recommended Year 3 students continue to participate in NAPLAN writing tests on pen-and-paper.

Given the importance of typing proficiency for computer-based writing assessments, schools should consider their local contexts and identify an effective method for developing students' typing fluency and to monitor the development of their typing proficiency over time, for students beyond Year 3.

Finally, further research is planned in NSW schools to investigate how the teaching of writing, and the writing process itself, can be enriched using new technologies to further develop students' writing skills.

Introduction

Every year since 2008, Australian students in Years 3, 5, 7 and 9 have undertaken pencil-and-paper tests of literacy and numeracy under the National Assessment Program – Literacy and Numeracy (NAPLAN). Within each assessed learning domain (reading, writing, spelling, grammar and punctuation, and numeracy) the tests are vertically and horizontally equated so that they are comparable across calendar years and across scholastic years.

In 2014, education ministers agreed that the annual NAPLAN tests will transition from the traditional pencil-and-paper delivery format to online delivery, with state and territory jurisdictions opting in no later than 2019. This means that all students will be required to use a computer workstation, laptop or tablet to complete the NAPLAN tests by 2019. The transition will commence from 2018 with most jurisdictions adopting a phased transition strategy, meaning the assessments are likely to be run in dual modes in 2018.

The Australian Curriculum, Assessment and Reporting Authority (ACARA) is responsible for developing the NAPLAN online tests and has been undertaking comprehensive research and trialling to support the move to NAPLAN online[1].

The benefits of online testing include tailored test design, a broader scope of assessable material, reduced time to provide feedback to schools and teachers, and potential improved test engagement particularly for low achieving students (Boyd & McDowall, 2001; Morphy & Graham, 2012). However, a recent review of relevant literature has raised a number of potential issues associated with online testing, particularly those that require extended responses (Eyre, 2015). These issues include validity (whether there is sufficient evidence to substantiate the link between scores from online tests and the intended interpretation and use of these scores), comparability (whether online tests are assessing the same skills as the equivalent versions of paper based tests, a critical issue when assessments are conducted in dual modes) and equity and fairness (for example, whether online tests represent a greater barrier for some groups of students than others to perform to their full potential due to different levels of experience with computer use).

As a national assessment program, NAPLAN provides data and trends on students' progress in literacy and numeracy. The assessment data are heavily relied upon for important policy and funding decisions as well as for informing the implementation of specific intervention programs. It is therefore vitally important that there is sufficient evidence supporting the validity and reliability of these assessments, not just for results at the population level, but also for the at-risk subgroups of students (for example, for Aboriginal, low SES students and students in rural and remote areas). Additionally as NAPLAN tests move to dual-mode delivery next year with some schools doing tests online and some doing them on the paper, it is critical that there is sufficient evidence to support claims of the comparability of results from online and paper tests. From a pedagogical perspective, it is also important to understand how the mode change impacts on the writing process, as well as what schools are currently doing, to help inform changes to teaching and school practices that would ultimately lift the writing skills of all students.

1 https://www.acara.edu.au/assessment/online-assessment-research

An additional key issue for NAPLAN online raised in public and at policy working groups concerns online assessment of writing, for Year 3 students in particular. The main concern is whether Year 3 students have sufficient typing skills to produce online texts in timed conditions that are a valid reflection of their underlying writing proficiency. Furthermore, there is a concern that the implementation of the Year 3 online test may detract from the teaching of handwriting to students at this early learning stage, which may have a detrimental effect on students' cognitive and early literacy development.

In order to inform policy decisions regarding the transition to NAPLAN online, a number of states as well as ACARA have investigated the effect of the change in mode of delivery on the test scores from writing assessments (that is, mode effects), particularly for Year 3 students, in 2016. This paper reports the findings from the NSW research, conducted in September, 2016, by the NSW Department of Education in collaboration with the Catholic Education Commission, Association of Independent Schools and the NSW Education Standards Authority.

The paper starts with a brief literature review of mode effects, focusing on writing assessments in the context of school education (K-12), followed by descriptions of the research methods, findings and implications of the findings for teaching, learning and assessment.

Literature review

Comparability of test results from computer-based and paper-based tests

Over the last two decades, a number of studies (for example, Horkay, Bennett, Allen, Kaplan & Yan, 2006; Morphy & Graham, 2012; Russell & Haney, 1997; Russell & Plati, 2001; White, Kim, Chen & Liu, 2015) have examined whether there are differences in performance between the test delivery modes for K-12 writing assessments. These studies were conducted against the backdrop of an increasing use of computers in delivering large-scale writing assessments for school students by international jurisdictions.

In general, studies have reported mixed results depending on the cohorts of students examined. For example, a large study by US Department of Education (White et al., 2015) examined whether fourth grade students can fully demonstrate their writing proficiency in computer tests, using data collected from 10,400 fourth graders in a 2012 pilot writing assessment. Whilst at a population level, the average scores from online and paper tests appeared to be comparable, there was a polarising mode impact pattern. High-performing students achieved substantively higher scores on computer than on paper testing (effect size = 0.56), whilst low-achieving students achieved even lower scores on computer (effect size = 0.16), suggesting the mode change had the potential to widen the achievement gap. The differential mode effects for high and low performing students were found to relate to students' prior exposure to computer use, prior exposure to writing on the computer and typing and editing skills. Computer/internet access and experience varied across student demographic groups. The study also reported that fourth graders had a low average typing speed of 12 words per minute, as compared to 30 words per minute for 8th graders. This could explain another finding from the study that on average the fourth graders wrote fewer words on computer than paper.

Horkay et al. (2006) examined the comparability of USA NAEP (National Assessment of Educational Progress) writing assessments administered on computer versus on paper, using data collected from large representative samples of eighth grade students (13-14 years old) in USA. They found that, whilst at the population level, and for most demographic subgroups investigated (for example, gender, race/ethnicity, parents' educational levels), there were no significant mean score differences between paper and computer delivery, students from urban/large town areas performed statistically worse on computer than on paper. Furthermore, students with high computer familiarity (a combined measure of typing speed, typing accuracy, basic word processing skills) wrote better on computer than students with lower computer familiarity but with equivalent paper writing skills. Based on the study results, Horkay and colleagues (2006) cautioned against an interpretation which finds no difference in mean scores at the population level, given the potential for population means to mask the differential impact of mode changes on subgroups of students.

Two smaller studies (Barkaoui, 2014; Zou & Chen, 2016) focusing on students learning English as an additional language (EAL) also reported that the level of computer familiarity impacted on the quality of online written texts produced, for EAL students, over and above their general English language proficiency and level of writing skills.

This link between keyboarding skills and students' online writing performance is also discussed and confirmed in other studies (for example, Alves, Castro & Olive, 2008) including those from Australia (for example, Christensen, 2004; Fluck, Pullen & Harper, 2009), with Christensen (2004) suggesting the link between skills and performance may be explained by the interplay of cognitive load and children's orthographic-motor processes.

Cognitive load theory and writing performance

Cognitive load theory (Sweller, 1994; Sweller, van Merrienboer & Paas, 1998) emphasises the limits of working memory and the importance of these limits for instructional design, particularly when processing novel information. Generally, the more pieces of information the brain is required to process in the working memory, the higher the cognitive load[2].

Similar to cognitive load theory, theories of writing as a cognitive process suggest that writers need to draw on a limited number of cognitive resources during the writing process to complete various higher order global functions such as planning, generating and organising ideas, and revising (for example, Douglas, 2000; Flower & Hayes, 1981). Graham, Berninger, Abbott, Abbott and Whitaker (1997) provide evidence for this amongst students in the lower primary grades. In their study of the mechanics of composing written texts, six hundred primary aged students were tested on the quantity and quality of handwritten work as impacted by handwriting fluency and spelling skills. Graham and colleagues found for both the younger (years 1 to 3) and older (years 4 to 6) cohorts that handwriting fluency directly impacted the quantity and quality of students' written work, suggesting that for students with limited handwriting fluency, more cognitive resources (or cognitive load) may be required to produce text (for example,, letter production, motor skills) which could lead to less cognitive resources available to attend to higher-order processes such as planning, translating and revising.

Bisschop, Morales, Gil and Jiménez-Suárez (2016) suggest that automation of lower order writing skills can allow the working memory to process more higher order writing skills such as generating ideas, planning, and revision. More specifically Christensen (2004) suggests students should first develop 'proficiency' in orthographic-motor skills (both being able to incorporate the visual representation of letters and these letters as words; as well as the motor skills to generate those representations) to allow students to process higher order writing skills in their working memory. Yet, Sweller and colleagues (1998) suggest automaticity is generally not achieved until a person has had extensive practice and warn that "performance is likely to be slow and clumsy" (p. 258) without automation.

The addition of keyboarding could add to the cognitive load for any students who have not achieved at least a basic proficiency in using computers to compose text. Young students may develop some keyboard proficiency through computer use at home and at school however it's unlikely that their typing has become automatic due to limited systematic instruction in typing (Alstad et al., 2015). Insufficient keyboarding skills prevent writers from fully engaging in the writing processes as they have to devote valuable cognitive resources to mechanical activities (for example, locating keys on the keypads), leaving fewer resources available for higher-order writing processes such as those "involved in ideation, syntactic and

2 For details of cognitive load theory and its impact for teaching and learning, see a detailed literature review by CESE at https://education.nsw.gov.au/about-us/educational-data/cese/publications/literature-reviews/cognitive-load-theory.

semantic monitoring and pragmatic awareness" (Christensen, 2004, p. 561). Thus, the working memory of young students may be loaded with how to type, as opposed to what to type.

The increase in automaticity with age may also explain why some studies report higher writing performance when composing text online vs pen and paper. For example, a meta-analysis of 27 studies comparing computer-based vs paper-based writing of weaker writers found keyboarding created advantages in written quality, length, organisation, mechanical correctness and student motivation (Morphy & Graham, 2012). However, the majority of these studies focussed on secondary and upper primary students, and only four studies with students in Years 1 to 3 were included.

Other aspects of keyboarding that impact writing performance

Similar to handwriting fluency, keyboarding skills have been shown to directly impact text length (a strong predictor of text quality) in time conditioned tests, particularly for younger students (Horne, Ferrier, Singleton, & Read, 2011; Russell & Plati, 2000; White et al., 2015). Connelly, Gee & Walsh (2007), in a study of 300 Year 5 and 6 students, found that keyboarding speed was consistently behind handwriting speed, and keyboarded scripts were up to 2 years behind handwritten scripts in development. Likewise, Berninger, Abbott, Augsburger and Garcia (2009) found that while Year 2, 4 and 6 students could type more letters from the alphabet on a keyboard than by hand, these students wrote faster and longer essays when written by hand, and year 4 and 6 students also write more complete sentences by hand.

Greater variation in typing skills than handwriting skills (Horne et al., 2011) could also explain the greater variability observed in online test scores than in paper scores (for example, Horkay et al., 2006). Given keyboarding skills can be a barrier to students demonstrating their writing proficiency fully, researchers (for example, Connelly et al., 2007; MacArthur, 1999; Poole & Preciado, 2016) call for explicit keyboarding instruction to develop students' keyboarding proficiency, before students are required to sit high-stakes assessments online.

Delivery of typing instruction

As a psychomotor skill, fluent typing requires coordination. While it is generally accepted that children possess fine-motor control by 7 years old (Chwirka, Gurney & Burtner, 2002) an appropriate age range for students to commence keyboarding has been debated. In a literature review of past studies on handwriting versus keyboard skills of students, Freeman, MacKinnon, and Miller (2005) indicate that automation of keyboarding skills occurs later in age than the automation of handwriting skills; evidence of 'slow and clumsy' performance is observed in the literature which suggest young children make use of the hunt-and-peck technique when typing. In a study of sixty-six Year 2 students, Chwirka and colleagues (2002) concluded that Year 2 students were able to achieve a keyboarding proficiency of five words per minute only after using keyboarding software for 15 minutes per day over an eight month period. Freeman and colleagues (2005) conclude that for most students to effectively use keyboards, they should be able to type as fast as they can write. In addition, they suggest students should be given an opportunity to automate keyboarding skills by learning how to touch-type and be provided with instruction and time to do so. Additionally they recommended the upper primary grades as an appropriate age for students to learn to type: "there is general agreement that students at the upper elementary level are ready to acquire keyboarding skill **in a reasonable amount of time**" (p. 140, emphasis added).

Impact of technology and assessment mode on teaching practice

The impact of assessment on classroom practice is well acknowledged in the literature (Broad, 2006; Herrington & Moran, 2006). One area of policy significance that requires further investigation is the potential impact of online assessment of writing on the teaching of handwriting in the early years. Recent studies (Dinehart & Manfra, 2013; Longcamp, Zerbato-Poudou & Velay, 2005) suggest that handwriting is positively correlated with phonological awareness and hence with spelling, as well as being positively associated with letter memorisation and early word learning. Therefore online testing of writing for younger students (that is, Year 3) might detract from the teaching of handwriting in the early years.

Another important question for educators is how teaching pedagogy should be adapted to harness the potential of technology to transform teaching and learning (Edwards-Groves, 2011; Kervin, Verenikina, Jones, & Beath, 2013; Walsh, 2010). Many researchers view technology in schools as a lever for overcoming inequality, and believe it can provide low SES students with powerful learning tools that can help them overcome educational disadvantage (Warschauer, 2006). Three meta-analyses (cited in Eyre, 2015) conducted between 1993 and 2007 examined how functions of word processing, (for example, the ability to move, copy, paste, delete and insert text online) can support teaching of writing. All found that such skills have the potential to lift the quality of students' writing. The effects tend to be larger for middle and high school students than for elementary students (Goldberg, Russell & Cook, 2003) and greater for low-achieving writers in the context of adolescents' writing (Graham & Perin, 2007). However, the research is also clear that in order to maximise the benefits of technology, students need to be explicitly taught how to use word processing to improve writing, and more importantly, students need to be taught how to use these tools at all stages of the writing process (MacArthur, 2009). More recent research has given increased attention to writing practices in contemporary digital environments, recognising that digital composition is becoming a central part of what teachers and students do in the literacy classroom (Mills & Exley, 2014; Walsh & Simpson, 2013).

However, it's worth noting that very little information is available as to what Australian schools are actually doing to help students to better use technology to improve the writing process. This is despite the Australian Curriculum expecting Year 5 students to use technology to 'compose with increasing fluency, accuracy and legibility' (ACELY 1707)[3]. The lack of Australian research in this area points to the need for future studies that examine the nature of the typing instruction delivered by schools, as well as the ways schools are using ICT to support and enhance learning of writing as part of effective teaching pedagogy.

3 www.australiancurriculum.edu.au

Validity and equity issues

The validity of online writing scores is challenged by research findings that online writing performance is potentially affected by factors such as keyboarding skills which are unrelated to the underlying skills being assessed (that is, writing skills). Unless the definition of the construct being assessed by the online writing tests is changed to 'the ability to write on computer', existing evidence points to the mode change potentially introducing construct-irrelevant variance (Messick, 1989) into the measurement process, thus weakening the link between test scores and the intended score interpretations.

Moreover, findings that online writing assessments have the potential to widen achievement gaps (Horkay et al., 2006; White et al., 2015) have also led researchers to question the fairness and accessibility of online writing assessments (Eyre, 2015). Despite the ubiquity of computers in modern society, not all demographic groups possess the same levels of computer access and familiarity (Jerrim, 2016). This has the potential to further exacerbate existing differences in student performance across demographics. As noted by Eyre (2015), "[t]hose who have low levels of computer familiarity and are also struggling writers are likely to be doubly disadvantaged in computer-based assessments of writing" (p. 17).

Research questions

With the findings from this literature review in mind, the present study attempts to investigate the following questions:

1. Do primary students perform differently according to the mode of writing test (computer-based versus pen and paper based)? If there is a difference, is it uniform across different student groups?
2. To what extent does the mode effect vary for students with different levels of typing proficiency?
3. What is the impact of the mode change on the writing process (for example, planning, editing, reviewing)?
4. Are there any other factors potentially impacting on students' computer-based writing performance highlighted through teacher and student interviews?

Methodology

Sampling of schools and students

This study used a two-stage sampling design. Firstly, sixty-nine schools in NSW were selected under a stratified random sampling procedure across sectors, locations and levels of socio-educational advantage (as approximated by the Index of Community Socio-Educational Advantage (ICSEA))[4].

Secondly, each school nominated two classes, representative of the school population, to participate in the study, one with Year 2 or 3 students, the other with Year 4 or 5 students. Year 2 students were included in the study as, at the time of research (that is, Term 3, 2016), they were two terms away from undertaking the Year 3 NAPLAN tests.

Writing genres and tasks

To understand the mode effect for different genres of writing, the same two genres that are tested in NAPLAN were assessed in this study: narrative writing and persuasive writing. For each genre, two different writing prompts were selected from previous NAPLAN writing tests; the narrative writing tasks from 2008 and 2009, and the persuasive writing tasks used in 2011 and 2012. The benefit of using previous NAPLAN writing tasks was that the marking criteria and rubrics already exist and trained NAPLAN markers could be used for greater marking consistency.

Schools were randomly assigned to one genre, so that all students participating in the study at each school undertook both a paper-based and a computer-based writing task **of the same genre**, a few days apart. Having the same student attempting tasks of the same genre was designed to remove the effect of genre on study results. Schools were assisted via a purpose-built website to develop a test schedule so that the order in which paper-based and computer-based tasks were administered was alternated for class 1 and class 2 at each school to minimise the confounding effect of task sequence on the study results.

Some data attrition occurred as a result of technical issues, computer and internet access issues, student absenteeism on the test day, and parents opting out of the research. The final data set comprised a total of 1,651 students with results for both the paper-based and computer-based writing tests. Table 1 provides demographic details of the student sample.

4 See explanatory notes for ICSEA at the ACARA website https://acaraweb.blob.core.windows.net/resources/Fact_Sheet_-_About_ICSEA.pdf.

Table 1

Sampled student demographics by genre of writing tasks undertaken

	Narrative	Persuasive	Total
Total students	**721**	**930**	**1,651**
Male	369	446	815
Female	352	484	836
Aboriginal students	44	45	89
Year			
Year 2	163	141	304
Year 3	236	325	561
Year 4	215	264	479
Year 5	107	200	307
Sector			
Government	465	576	1041
Catholic	214	283	497
Independent	42	71	113
Location			
Metropolitan	544	713	1257
Provincial	150	129	279
Remote	27	88	115

Test administration

Participating schools were supplied with writing task cover sheets, pre-printed with each student's name and unique student identifier, and planning sheets for both the paper-based and computer-based tasks to assist students with planning their responses. For the paper-based tasks, printed instructions, writing prompts, and response booklets were also provided. For the computer-based tasks, ear buds were supplied.

The computer-based tests were administered via the Department's online assessment platform, and were designed and administered consistent with the NAPLAN online research studies conducted by ACARA. For example, the online writing task screens were locked down so that spelling and grammar checking was disabled during the online tests.

All writing tests were administered under standard NAPLAN writing test conditions, with students given 40 minutes to read the stimulus, plan, write and edit their scripts. After both classes had completed both writing tasks, the paper-based writing responses and the planning sheets for both tasks were returned to the researchers for analysis. For marking purposes, students' online writing responses were printed and attached to their cover sheets and planning pages.

Marking

Students' writing scripts were marked by 20 experienced NAPLAN markers according to standard NAPLAN marking rubrics[5], and standard marking quality assurance procedures whereby two senior markers monitored the marking process and double marked 10 per cent of the scripts. Each marker examined only one genre (narrative or persuasive), with scripts de-identified and mixed by year level and school. Markers received training similar to the 2016 ACARA online trial study. This training assisted markers to identify any text characteristics pertinent to typed scripts particularly relating to spelling, layout and punctuation, and to discount any capitalisation or punctuation or spelling errors that could clearly be judged as typographical errors.

A random selection of 124 handwritten scripts were reproduced as typed scripts and re-marked by a different marker to investigate any marking bias due to text appearance.

Typing test and student survey

Following completion of both writing tasks, students completed a short online survey about their preference and use of computers for writing, followed by a 3 minute online typing test, for which they were instructed to copy the text, typing as accurately and quickly as possible. Texts for the typing test were carefully selected to be age appropriate. A different text was used for Years 2 and 3 than for Years 4 and 5.

The typing test automatically timed out after 3 minutes. For students who completed the typing test in less than 3 minutes, the duration of each student's typing test was recorded by the online platform.

Teacher and student observations and interviews

Seventeen researchers from across the sectors each visited a school and observed the two classes completing the paper and computer based writing tests with the aim of identifying any differences in students' planning, drafting and editing processes.

The researchers also conducted follow-up interviews with participating teachers and one or two randomly selected students to capture information about students' exposure to typing instruction, typing and computer experiences at home and school, and teacher and student confidence and engagement with technology.

5 See http://www.nap.edu.au/_resources/Amended_2013_Persuasive_Writing_Marking_Guide_-With_cover.pdf for an example of the NAPLAN marking guides for persuasive writing.

Data treatment and variables used in analysis

In order to improve the interpretability of the results, the total raw scores (sum of all trait scores for each response) for each writing task were transformed to NAPLAN scaled scores (scores on the same NAPLAN writing scale) based on advice specifically provided by ACARA for this research. This ensures that scores from the paper based and online tests for the same student are comparable as well as scores from the narrative tasks and those from the persuasive tasks are also comparable with similar population variances[6]. In part this helps to remove confounding task and genre effects.

Given previous studies have found evidence of mode change exacerbating the gap in writing results between high and low performing students (for example, Horkay et al., 2006), a research interest for this study was whether the mode change affects students at various ability levels differently. A proxy measure of general literacy ability was generated from students' NAPLAN results in reading, writing, spelling, and grammar and punctuation. For students in Years 3 and 5, NAPLAN results from 2016 (two terms prior to the study) were used. For students in Years 2 and 4 in 2016, NAPLAN results from the following year (2017; two terms after the study) were obtained once they became available. This resulted in NAPLAN results being available for 96% of students in the study sample. A measure of general literacy ability was generated by first standardising students' results within each year level for each literacy domain, then averaging the results across the four literacy domains, and re-standardising the average literacy scores within each year level.

Measures of typing ability included typing speed and typing accuracy. Typing speed was measured in words per minute (wpm) by dividing the total number of words typed by the duration of the typing test. Typing accuracy was measured as the percentage of words typed correctly.

A number of technical issues, including computer and internet access issues, resulted in either loss or corruption of some of the online writing scripts and typing test responses. To ensure the robustness of the study findings, online scripts that were known to have been impacted by technical issues based on teacher feedback, or which received a score that was 115 NAPLAN scaled score points (or roughly 2 NAPLAN bands) lower than the scaled score points received by the same student from the paper-based test were excluded from the analysis. For similar reasons, typing test results for students who typed 5 or fewer words were also excluded from the analysis. The final data set included a total of 1,651 students from Years 2-5 with both paper-based and computer-based writing test scores. Typing test results were available for a total of 1,323 students, with 879 students having both writing test scores plus typing test results available for analysis.

6 Every year, a pairwise equating process is used by the national testing authority to equate writing results from that year to the national writing scale. Details of that process are published in the technical reports at the https://www.nap.edu.au/results-and-reports/national-reports. For this research, raw scores from 2011 and 2012 NAPLAN tasks were converted to the scaled scores using the equating information provided by ACARA released to states and territories each year. Raw scores from 2008 and 2009 were converted to the scaled scores using specific equating information provided by ACARA, as the current writing scale is only used for reporting of writing results from 2011 onwards.

Sampling weights

Appropriate sampling weights were applied for the estimation of the mode effect (that is, differences in the writing scores) for all students and subgroups of students. The weighting methodology was specific to the two-stage stratified sampling procedure and similar to that used for international tests adopting similar sampling designs such as TIMSS (Foy, 1995). School weights were calculated as the total number of schools in each stratum divided by the number of sampled schools. Class weights were unable to be calculated as the total number of classes per year level per school was not centrally available. Instead, students in Years 2 and 3, and those in Years 4 and 5, were treated as two virtual groups of students. Weights were applied for each virtual group by dividing the total number of students in the relevant year levels within each school by the number of sampled students in those years. The final weight applicable to each student in the final dataset was the product of the relevant school and virtual group weights.

Analysis methods

The overall difference between the paper-based and computer-based writing scores (that is, the mode effect) was first examined by paired sample t-tests (sampling weights applied) to determine the mode change effect sizes for all students, and for each year level, demographic group and writing genre. For analysis by location, the small numbers of remote students were combined with provincial students to form a "non-metropolitan" group for analysis purposes.

To estimate the net effects of the mode change on writing performance accurately, more sophisticated statistical methods are needed to account for a range of confounding factors that may impact on the assessed performance levels from writing tests (for example,, Carr, 2000; Huot, 1990; Leckie & Baird, 2011; Lim, 2009; Schoonen, 2005; Weigle, 2002). These factors can be categorised as those related to tasks (for example, prompts, genres, order of the tests if multiple tests involved), to the marking process (for example, marking rubrics used and marking inconsistency), to student and school related factors (for example, gender, student and school SES) and to the interactions amongst these factors that exist across different hierarchical levels.

A measurement analysis was initially conducted using the Many Facet Rasch Measurement (MFRM) model (details see Linacre, 1989; Linacre & Wright, 2002), which is an expanded form of the Rasch rating scale model[7]. This model conceptualises that writing assessment scores are simultaneously influenced by a number of facets from the measurement process (for example, person ability, the task attempted, the genre of the task and the mode in which the essay is produced). The model calibrates these facets in one reference framework, enabling the effect size of the different facets to be estimated (when others are accounted for) and compared (see Appendix 1 for more technical details). The relevant analysis was carried out using the FACETS computer program (Linacre, 2008).

7 Note that Rasch modelling is also used for test calibration and scaling of the NAPLAN tests.

To further account for the interactions amongst the different factors as well as the nested structure of the data (for example, students within schools), a repeated measures analysis using a mixed effects model was then conducted to investigate the net mode effect associated with task related factors as well as those related to students and schools, when all explanatory variables and interactions between pairs of variables are considered in one statistical model (see Appendix 2 for more technical details).

This model was first run with all explanatory variables with the exception of typing proficiency, due to the loss of cases when the typing variables are added to the analysis. The first analysis therefore provides more precise estimates for the mode effect for different student groups, when typing proficiency was not considered.

The model was then run on the reduced dataset with all explanatory variables and the typing proficiency variables plus their interactions, to explore the impact of typing on the mode effect.

Further qualitative and descriptive analyses using information collected from classroom observations, teacher and student surveys and interviews were carried out to investigate research questions 3 and 4.

Results and discussion

Research question 1 – Do primary students perform differently according to the mode of writing test (computer-based versus pen and paper based)? If there is a difference, is it uniform across different student groups?

Descriptive and t-test statistics

The means and standard errors of the weighted scaled writing scores from the paper-based and computer-based writing tasks were first examined. As shown in Table 2, the mean scaled scores for the computer-based writing tasks were lower than for paper-based writing tasks for all year levels and both writing genres.

Table 2

Writing task mean scaled scores with standard errors (SE) (sampling weights applied)

	Narrative writing task					Persuasive writing task				
		Paper-based		Computer-based			Paper-based		Computer-based	
Year	*n*	Mean	SE	Mean	SE	*n*	Mean	SE	Mean	SE
Year 2	163	415.1	5.1	381.0	5.3	141	388.6	4.5	373.9	4.6
Year 3	236	442.7	4.4	424.8	4.4	325	440.6	5.6	426.0	4.9
Year 4	215	444.9	3.9	433.5	4.8	264	456.3	4.6	439.4	5.2
Year 5	107	493.0	4.2	474.0	5.4	200	496.0	4.1	479.7	5.2

Figure 1 shows the scatterplot of students' computer-based vs paper-based results. Whilst the mean computer-based results were lower than the mean paper-based results, Figure 1 shows that there were some students who scored higher on the computer-based test than the paper-based test. Also as expected, there was a moderately strong relationship between paper-based and computer-based writing test scores, as shown in Figure 1. The correlation between paper-based and computer-based test scores was consistently around $r = 0.7$ within each year level.

Figure 1

Scatterplot of computer-based versus paper-based writing scores

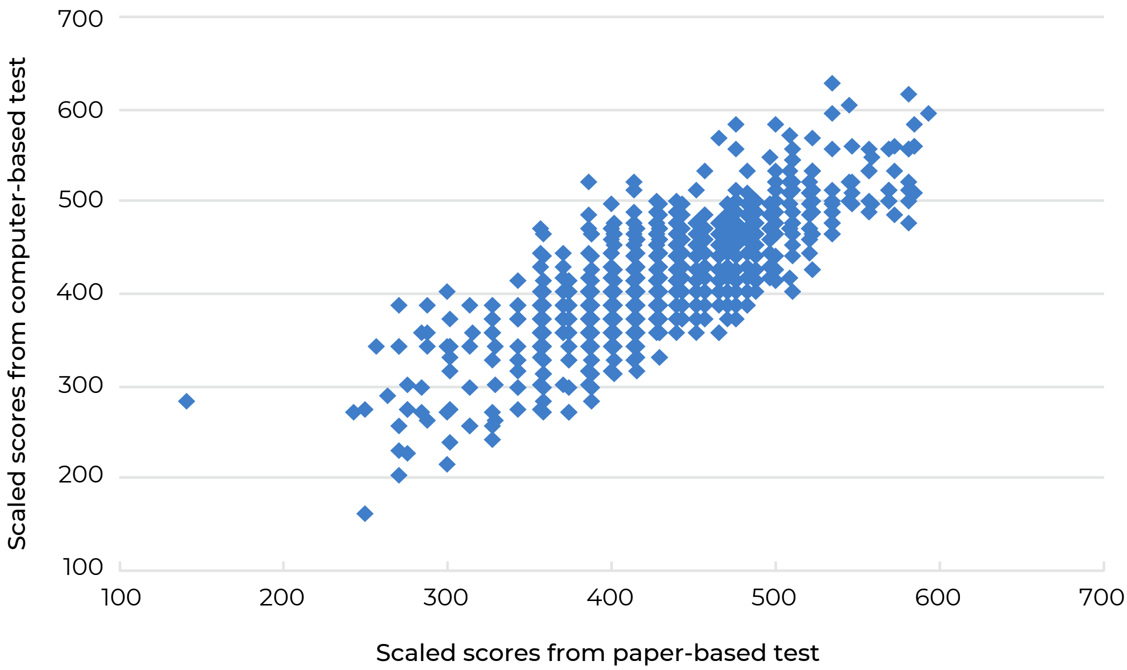


Table 3 presents the mean differences in the scaled scores for the two writing tasks by genre, year level, gender, and location, along with the results of a series of paired sample t-tests, after sampling weights have been applied. With an overall difference of -17.6 NAPLAN scaled score points, writing scores for computer-based tasks were significantly lower than those from the same students' paper-based tasks ($t(1638) = -12.8$, $p<.001$).

Table 3

Mean differences in scaled scores (computer-based score – paper-based test score) and paired sample t-test results (sampling weights applied)

	n	Mean difference	Standard error	*t*	Effect size
All	1651	-17.6	1.4	-12.8*	-0.31
Narrative	721	-20.2	2.0	-10.1*	-0.37
Persuasive	930	-15.7	1.9	-8.5*	-0.28
Year 2	304	-24.9	3.2	-7.8*	-0.45
Year 3	561	-16.0	2.1	-7.8*	-0.33
Year 4	479	-14.1	2.5	-5.6*	-0.25
Year 5	307	-17.4	3.2	-5.4*	-0.31
Female	836	-22.0	1.9	-11.9*	-0.41
Male	815	-13.5	2.0	-6.8*	-0.24
Metropolitan	1257	-17.7	1.5	-11.4*	-0.32
Non-metropolitan	394	-17.3	3.0	-5.8*	-0.29

Note. Negative differences indicate the average computer-based test scores were lower than paper-based test scores.

* all significant at $p<.001$.

As presented in Table 3, paired t-test results indicated that students' computer-based writing results were statistically significantly lower than their paper-based writing results across all year levels (Years 2 to 5), across both genders, and across both narrative and persuasive writing genres. The difference was greater for students in Year 2 (-24.9), than for students in other years (differences ranging from -14.1 to -17.4). Boys appeared to be less adversely affected by the mode change than girls (differences of -13.5 and -22.0 respectively), and students who attempted persuasive tasks appeared to be less affected than those who attempted narrative tasks (differences of -15.7 and -20.2 respectively).

Across all students, the estimated effect size (d) associated with the mode change was 0.31, signifying a moderate effect (Cohen, 1988), with the effect size across subgroups and genres ranging from 0.24 to 0.45.

FACETS measurement analysis

A FACETS measurement analysis of students' writing scores was also performed, separately for the narrative and persuasive writing tasks, to examine and compare the effects of a number of task related factors including task mode on the writing results. As indicated in Appendix 1 which provides details of the model specifications, this analysis conceptualises the test score achieved by a student as a function of person ability as well as the task difficulty associated with task-related factors such as the task mode, the task prompts, and the task order (that is, whether it was done as the first or the second test). The difficulty of each measurement facet is estimated and reported in Table 4.

Table 4 shows that for both genres, the computer-based mode is more difficult than the paper-based mode, with the difference in the difficulty estimates between the two modes estimated at 0.56 logit values for narrative writing and 0.44 logit values for persuasive writing. This translates to an effect size[8] of 0.28 for narrative writing and 0.25 for persuasive writing. In contrast, other factors have much smaller effect sizes. For example, on average, students achieved slightly better results for the second test than they did for the first test, indicating some test practice effect. However, the difference in the difficulty estimates for the two test orders (done first vs done second) is 0.08 logits for both genres, which is equivalent to an effect size of 0.04 and 0.05 respectively for the narrative and persuasive tasks.

8 Effect size is calculated as the ratio of the difference in the difficulty estimate to the standard deviation of the person ability estimates.

Table 4

Fit statistics and difficulty estimates for measurement facets from FACETS analysis

Narrative tasks (Person logit SD = 2.00)		Difficulty estimate	S.E.	Infit mean square	Outfit mean square	Difficulty difference	Effect size
Mode	Paper based Computer based	-0.28 0.28	0.03 0.03	1.02 1.00	1.00 0.99	0.56	0.28
Task prompt	Found Box	-0.03 0.03	0.03 0.03	1.01 1.01	1.00 1.00	0.06	0.03
Task order	Done first Done second	0.04 -0.04	0.03 0.03	1.01 1.01	1.00 1.00	0.08	0.04
Persuasive tasks (Person logit SD = 1.75)		**Difficulty estimate**	**S.E.**	**Infit mean square**	**Outfit mean square**	**Difficulty difference**	**Effect size**
Mode	Paper based Computer based	-0.22 0.22	0.02 0.02	1.01 1.00	1.00 0.98	0.44	0.25
Task prompt	Cook Toys	-0.05 0.05	0.02 0.02	0.99 1.01	0.97 1.01	0.10	0.06
Task order	Done first Done second	0.04 -0.04	0.02 0.02	1.01 1.00	0.98 1.00	0.08	0.05

*Interpretation of the fit statistics are provided in Appendix 1.

In summary across the two genres, the mode change from paper to computer-based testing has a negative effect on students' writing results, with the effect size estimated in the range of 0.25 to 0.28, after accounting for task related factors. This effect is much larger than that associated with other factors, such as the prompt effect and task order effect.

Repeated measures mixed effect analysis (Model 1)

In order to more appropriately account for the factors at multiple levels and interactions between factors, a repeated measures analysis using a mixed effects model on the writing test scores was then performed. This analysis takes into account the nested structure in the data (time nested within students nested within schools), as well as various explanatory factors, interactions between time and other explanatory variables, and interactions between task mode and other explanatory variables.

The dataset for this analysis was slightly smaller (1,581) than for the previous analyses (1,651) due to a small number of students for whom a measure of general literacy ability could not be generated due to missing NAPLAN results.

The results are reported in Table 5 with small and non-significant interactions removed except for those of research interest. Technical Appendix 2 provides information about the model specifications and results from diagnostic procedures checking model assumptions.

Table 5

Results from the mixed effects model for repeated measures

Dependent variable: Writing test score	Coef.	Robust Standard Error	*z*	*P>z*	95% Conf. Interval	
					CI lower	CI upper
Constant	405.7	24.2	16.8	0.000	358.4	453.1
Student characteristics						
Student literacy ability	36.9	1.6	23.1	0.000	33.8	40.1
Gender (Ref: Girls)						
Boys	-13.1	2.3	-5.7	0.000	-17.6	-8.6
ATSI status (Ref: Non-Aboriginal)						
Aboriginal	-12.1	6.0	-2.0	0.043	-23.8	-0.4
Year (Ref: Year 2)						
Year 3	33.9	8.2	4.1	0.000	17.8	50.0
Year 4	48.6	7.0	6.9	0.000	34.8	62.4
Year 5	82.9	8.4	9.8	0.000	66.3	99.4
School characteristics						
ICSEA	0.009	0.023	0.4	0.685	-0.035	0.054
Location (Ref: Metro)						
Non Metro region	-1.3	3.0	-0.4	0.670	-7.3	4.7
Writing test characteristics						
Task mode (Ref: Paper-based test)						
Computer-based test	-28.0	7.9	-3.5	0.000	-43.5	-12.4
Genre and Task prompt (Ref: Narrative 2008 Found)						
Narrative 2009 Box	-1.6	2.3	-0.7	0.488	-6.2	3.0
Persuasive 2011 Toys	-14.8	3.7	-4.0	0.000	-22.1	-7.5
Persuasive 2012 Cook	-8.8	3.6	-2.4	0.016	-16.0	-1.7
Time order of test (Ref: Time 1)						
Time 2	-3.0	7.7	-0.4	0.702	-18.1	12.2
Interactions						
Task mode#Student literacy ability	-0.8	1.5	-0.5	0.595	-3.8	2.2
Task mode#ATSI status						
Computer-based test#Aboriginal	12.1	5.7	2.1	0.033	1.0	23.1
Task mode#Gender						
Computer-based test#Male	8.5	2.6	3.3	0.001	3.4	13.6
Task mode#Year						
Computer-based test#Year 3	7.3	8.2	0.9	0.374	-8.8	23.5
Computer-based test#Year 4	10.1	8.7	1.2	0.247	-7.0	27.2
Computer-based test#Year 5	10.1	7.4	1.4	0.174	-4.5	24.7

Dependent variable: Writing test score	Coef.	Robust Standard Error	*z*	*P>z*	95% Conf. Interval CI lower	CI upper
Time order of test#Year						
Time 2#Year 3	7.8	8.2	1.0	0.339	-8.2	23.8
Time 2#Year 4	11.7	7.9	1.5	0.136	-3.7	27.1
Time 2#Year 5	13.0	9.2	1.4	0.157	-5.0	31.0
Number of observations	**3162**					
Number of students	**1581**					
Number of schools	**69**					

Note. The model incorporates student sample weights and school sample weights appropriately at the student and school levels.

As shown in Table 5, the main effects from the repeated measures analysis for student level factors were as expected. For example, students with lower literacy skills (relative to their peers in the same year level) performed worse on writing tests than students with higher literacy skills (ß = 36.9, $p < .001$). Boys, on average, scored around 13 scaled score points lower than girls (ß = -13.1, $p < .001$), and Aboriginal students scored lower on average than their counterparts (ß = -12.1, $p < .05$). As expected, students in Years 3, 4 and 5 achieved higher writing results than students in Year 2.

The main effects for the two school level factors, average level of socio-educational advantage (ICSEA) and location (metro vs non-metro), while in the expected direction, were not statistically significant after controlling for student general literacy ability and other student level factors. This means that there is no evidence that students with the same level of literacy ability performed differently on the writing tests in high vs low SES schools or in metro vs non-metro schools.

The results also suggest that students undertaking the persuasive writing tasks scored lower than those undertaking the narrative writing tasks. While students performed similarly on the two narrative tasks (ß = -1.6, $p = .488$, for narrative task "Box" compared to narrative task "Found"), students performed lower on the persuasive task "Toys" (ß = -14.8, $p < .001$) and on the persuasive task "Cook" (ß = -8.8, $p < .05$) than on the narrative reference task "Found".

The main effect for the time order of the writing tasks was not statistically significant, nor were the interactions between time order and year level, however, the coefficients for the interaction terms are sufficiently substantial to warrant some comment. For the reference group (Year 2), the small, non-significant main effect for time order (ß = -3.0, $p = .702$) indicates that the Year 2 students performed similarly on the second writing task than on the first task. However, the interaction terms between year level and time order indicate that older students seemed to perform better on the second task than on the first task, suggesting a possible task practice effect for older students (ß's ranging from 7.8 to 13; all $p > .10$).

Important coefficients pertinent to this research are those associated with task mode. Results indicate that for the reference group (Year 2 non-Aboriginal girls) computer-based tests scores were 28 scaled score points lower than paper-based test scores (ß = -28.0, $p < .001$). The interaction terms between year level and mode were not statistically significant, although the direction of the coefficients suggest that the mode effect for Year 3 to 5 students might be less than that for Year 2 students.

Statistically significant interactions included task mode with gender and task mode with Aboriginality, suggesting that the mode effect was less for boys (ß = 8.5, $p = .001$) and for Aboriginal students (ß = 12.1, $p = .033$). The interaction between task mode and students' literacy skills was small and not statistically significant (ß = -0.8, $p = .595$), indicating no evidence from this analysis of the mode effect being different for students with high or low literacy skills.

However, as computer familiarity (including typing proficiency) is found to contribute to the mode effect in earlier studies, there is a need to re-examine the effects for different student groups taking into account the potentially varying typing proficiency levels amongst these student groups.

Research question 2 – To what extent does the mode effect vary for students with different levels of typing proficiency?

Descriptive analysis of typing proficiency

Before typing proficiency variables are included in further statistical modelling, descriptive analysis of the distribution of typing proficiency and the relationship between typing ability and students' computer-based writing performance is examined.

Typing test results were available for 1,323 students. As expected, the median typing speed increased across the year levels, from a median of 5.3 words per minute (wpm) for Year 2 students to 15.4 wpm for Year 5 students (Table 6). Typing accuracy rates, however, were similar across the year levels, with median rates around 95 per cent. Overall typing speed and typing accuracy were correlated at r=0.33.

Table 6

Typing speed (words per minute) and typing accuracy rate by year level (sampling weights applied)

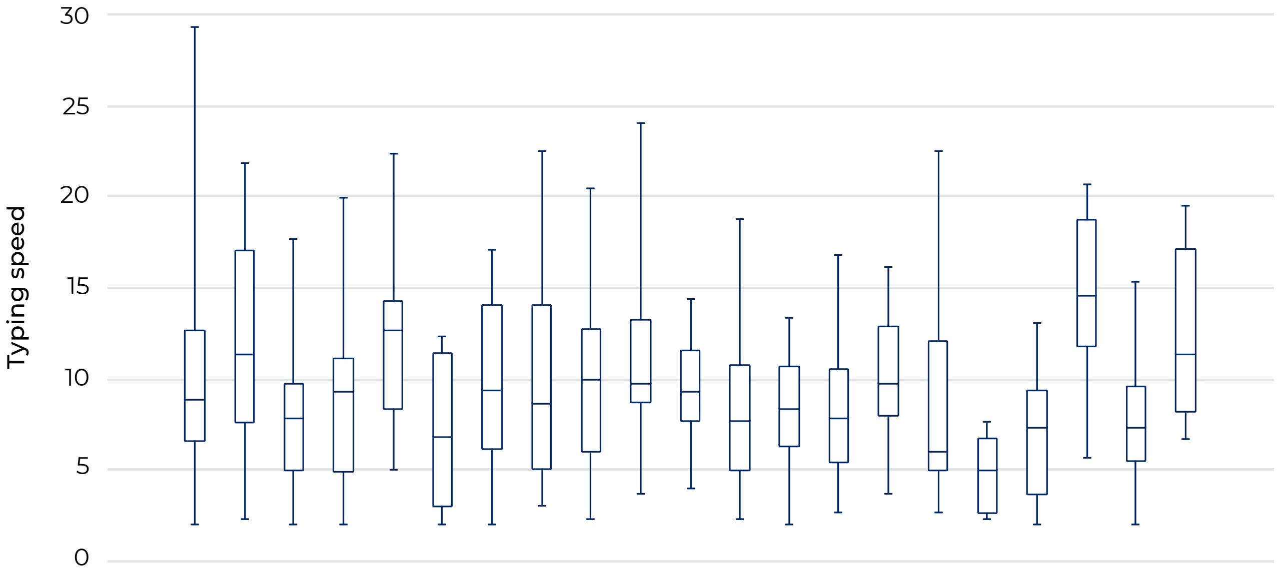
		Typing speed (wpm)		Typing accuracy rate (%)	
Year level	***N***	**Median**	**Inter-quartile range**	**Median**	**Inter-quartile range**
Year 2	242	5.3	3.3 - 7.7	93.7	83.3 - 100
Year 3	412	8.7	5.7 - 12.7	95.7	89.5 - 98.5
Year 4	306	10.3	6.4 - 14.1	94.2	90.0 – 97.7
Year 5	363	15.4	10.5 - 18.4	94.9	90.9 – 97.9

Note. Sampling weights were recalculated for this sample with typing data.

However, there was considerable variation in typing ability from school to school, and from student to student within each class at each school. Figure 2 depicts the variability in typing speed across schools for Year 3 students.

Figure 2

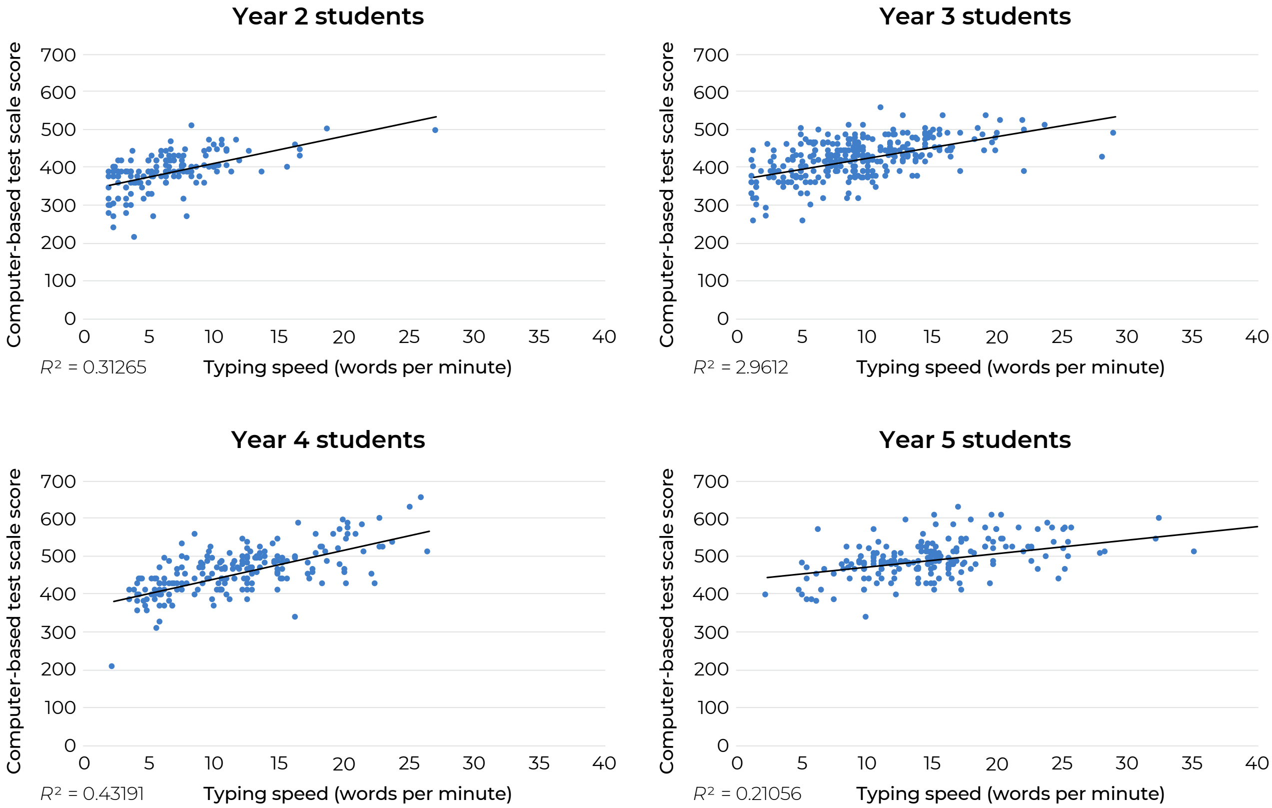
Boxplots indicating Year 3 median typing speed, interquartile range and minimum/maximum typing speed, for schools with five or more Year 3 students participating in the study



As expected, there was also a moderate relationship between typing speed and students' computer-based writing scores, with the correlations in the range of *r* =0.46 to *r* =0.66 for each year level, as indicated in Figure 3. This suggests that typing proficiency might be a contributing factor to the mode effect observed in the previous sections.

Figure 3

Scatterplots of typing speed and computer-based writing scale scores for each year level



Repeated measures mixed effect analysis including typing proficiency (Model 2)

The same repeated measures mixed effects model was used to examine whether the mode effect varies for students with different levels of typing proficiency, using the same set of explanatory variables as in the previous analysis with two additional explanatory variables – typing speed (words per minute) and typing accuracy rate, plus interaction terms between mode and typing variables.

Table 7 shows the results of this analysis with small and non-significant interactions removed except for those of research interest. The analysis was based on a total of 840 students who had typing data as well as writing results from both tests (plus non-missing literacy ability estimates). Whilst this was a reduced sample, analysis indicates that the distribution of the typing abilities of this reduced sample is very similar to those for the larger pool of students with typing results. This analysis was run with the new sampling weights calculated based on the reduced sample of students. Appendix 2 provides evidence from model diagnostic checks that indicate model assumptions were satisfied.

Table 7

Results from the mixed effects model for repeated measures including typing speed and accuracy

Dependent variable: Writing test score	Coef.	Robust Standard Error	*z*	*P>z*	95% Conf. Interval	
					CI lower	CI upper
Constant	336.8	41.5	8.1	0.000	255.5	418.1
Student characteristics						
Student literacy ability	32.8	2.7	12.2	0.000	27.5	38.1
Gender (Ref: Girls)						
Boys	-12.4	3.1	-4.0	0.000	-18.6	-6.3
ATSI status (Ref: Non-Aboriginal)						
Aboriginal	-15.1	9.7	-1.6	0.119	-34.2	3.9
Year (Ref: Year 2)						
Year 3	29.6	10.3	2.9	0.004	9.4	49.7
Year 4	39.9	11.4	3.5	0.000	17.7	62.2
Year 5	78.5	11.3	6.9	0.000	56.3	100.7
School characteristics						
ICSEA	0.03	0.03	1.0	0.343	-0.03	0.09
Location (Ref: Metro)						
Non Metro region	3.3	3.5	1.0	0.344	-3.5	10.2
Writing test characteristics						
Task mode (Ref: Paper-based test)						
Computer-based test	-30.8	24.4	-1.3	0.207	-78.6	17.0
Genre and Task prompt (Ref: Narrative 2008 Found)						
Narrative 2009 Box	0.3	3.2	0.1	0.938	-6.1	6.6
Persuasive 2011 Toys	-16.1	4.8	-3.4	0.001	-25.6	-6.7
Persuasive 2012 Cook	-11.6	4.8	-2.4	0.015	-21.0	-2.3
Time order of test (Ref: Time 1)						
Time 2	-9.1	7.7	-1.2	0.236	-24.2	6.0
Typing skills characteristics						
Typing accuracy	0.5	0.2	2.0	0.047	0.0	0.9
Typing word count per minute	0.7	0.3	2.3	0.025	0.1	1.4
Interactions						
Task mode#Student literacy ability	-5.2	1.9	-2.7	0.006	-9.0	-1.5
Task mode#ATSI status						
Computer-based test#Aboriginal	9.8	9.6	1.0	0.309	-9.1	28.6
Task mode#Gender						
Computer-based test#Male	5.8	3.4	1.7	0.090	-0.9	12.4

Dependent variable: Writing test score	Coef.	Robust Standard Error	*z*	*P>z*	95% Conf. Interval	
					CI lower	CI upper
Task mode#Year						
Computer-based test#Year 3	-9.8	8.3	-1.2	0.236	-26.1	6.4
Computer-based test#Year 4	-8.8	9.2	-1.0	0.339	-26.8	9.2
Computer-based test#Year 5	-16.5	7.5	-2.2	0.028	-31.2	-1.8
Time order of test#Year						
Time 2#Year 3	15.3	8.4	1.8	0.069	-1.2	31.8
Time 2#Year 4	15.9	10.0	1.6	0.110	-3.6	35.4
Time 2#Year 5	17.6	9.4	1.9	0.061	-0.8	36.0
Task mode#Typing accuracy	0.03	0.25	0.1	0.901	-0.46	0.52
Task mode#Typing word count per minute	1.5	0.3	4.4	0.000	0.8	2.2
Number of observations	**1680**					
Number of students	**840**					
Number of schools	**46**					

Table 7 shows that both typing speed and typing accuracy have a positive relationship with writing performance, which is not surprising given the likely relationship between typing proficiency and student overall literacy ability and that between student overall literacy ability and writing performance.

More importantly, the interaction between mode and typing speed is statistically significant (ß = 1.5, *p* < .001), although the interaction between mode and typing accuracy is not (ß = 0.03, *p* = .901). This means that, holding everything else equivalent, the negative mode effect (that is, the difference between paper-based and computer-based test scores) is reduced by approximately 7.5 score points if a student's typing speed is increased by 5 words per minute.

After typing variables were taken into account, the size of the previously observed differences in the mode effects for boys vs girls (ß = 5.8, *p* = .09) and for Aboriginal vs non Aboriginal students (ß = 9.8, *p* = .309), were reduced somewhat, and were no longer statistically significant, although the interaction between gender and mode approached statistical significance. This suggests that some of the difference between paper-based and computer-based test results for these demographic groups may be due to differences in typing speed between the respective groups. This is confirmed in Table 8 which shows that typing speeds for Aboriginal students and boys were lower than their respective counterparts, for each scholastic year.

Table 8

Median typing speed (and interquartile range) by year level, gender and Aboriginal status (reduced sample, sampling weights applied)

Subgroup	Year 2	Year 3	Year 4	Year 5
Female	6.7 (5.0-8.7)	10.4 (6.7-14.1)	11.7 (8.0-17.8)	15.4 (10.3-17.7)
Male	5.3 (3.4-7.7)	9.3 (7.0-12.7)	9.3 (5.7-13.0)	15.1 (11.7-17.1)
Non-Aboriginal	6.3 (3.7-7.9)	9.7 (7.0-13.4)	10.7 (6.3-15.0)	15.4 (11.3-17.7)
Aboriginal*	4.7 (2.0-6.0)	5.8 (5.8-12.4)	10.1 (6.3-10.7)	8.7 (6.7-15.7)

* **Note.** As the number of Aboriginal students in each year level is relatively small, results for Aboriginal students should be interpreted with caution.

An additional finding, in contrast to the repeated measures analysis without typing proficiency, is that the negative gap between computer-based and paper-based tests is larger (worse) for students with higher literacy skills (relative to their peers in the same year level). Holding everything else equal, including typing proficiency, a one standard deviation increase in a student's literacy level on the scale constructed for this analysis is associated with an increase in the size of the gap by 5.2 scaled points ($ß = -5.2$, $p = .006$). This differential effect is small in magnitude, nonetheless it is in contrast to research by White and colleagues (2015) whereby high ability fourth grade students in the US performed better on computer-based writing tasks than on paper-based tasks. Differences in methodology and analytical approaches between the two studies may explain these contrasting results. Given this study investigated a wider age range of students (with the inclusion of Year 2 and 3 students whom, as this study showed, have generally slow typing speeds) than the study by White and colleagues (in which Year 4 students took part), perhaps typing ability has a larger impact on the performance of younger high ability students.

Given the interaction between typing speed and mode, the interaction terms between year level and task mode from Table 7 are less straightforward to interpret as typing speed co-varies with year level. To assist interpretation, Table 9 provides the predicted difference between computer-based and paper-based tests, by year level and gender, for students with average literacy levels for their year, when they are at the median typing speed for their year. Note that the difference in the mode effect for boys vs girls is close to statistical significance and is non-negligible in size.

Also included are the predicted differences for students with faster (+5 wpm) and slower (-5 wpm) typing speeds than the median for their year level. For the Year 2 students, for whom the median typing speed is close to 5, the predicted difference is only calculated for the median typing speed for that year.

Table 9 shows that the gap between computer-based and paper-based writing performance is quite similar across the year levels, for students at the median typing speed for their year. The results in Table 9 also demonstrate the reduction in the gap (of around 7.5 scale points) for students who can type an additional 5 words per minute faster than the median for their year.

Table 9

Estimated differences in writing scaled scores between computer- and paper-based tests for an average ability student in each year level

	Female			Male			All		
Typing speed	-5 wpm	median wpm	+5 wpm	-5 wpm	median wpm	+5 wpm	-5 wpm	median wpm	+5 wpm
Year 2		-17.2			-13.6			-14.8	
Year 3	-29.1	-21.4	-13.7	-25.0	-17.3	-9.7	-27.1	-19.5	-11.8
Year 4	-26.0	-18.3	-10.7	-23.9	-16.2	-8.6	-24.5	-16.9	-9.2
Year 5	-28.0	-20.4	-12.7	-22.7	-15.1	-7.4	-25.0	-17.4	-9.7

Note. The gap for the different level of typing speed is provided for intervals of +/- 5 words/minute. For Year 2 students this is provided for the median typing speed only. When predicting the average difference between paper-based and computer-based writing results, all other variables included in the model (for example, literacy level) were held at the weighted sample mean/median levels.

In summary, the above analysis shows that after holding constant the effects of various task, student and school level factors on students' writing performance, on average, students scored lower on computer-based writing tests than they did on paper-based tests, across all years examined (Years 2-5), and for both narrative and persuasive writing genres.

Over above other factors, typing proficiency was significantly associated with the mode effect, such that the faster students could type, the smaller the difference between their computer-based and paper-based results. For an increase of 5 words per minute in typing speed, the gap is reduced by approximately 7.5 scaled points.

Additionally, after taking into account all other factors including typing, the mode effect was found to be slightly worse for students with higher literacy ability than those with lower ability. The mode effect was approximately 5 scaled points larger for students whose literacy level was one standard deviation above the average for their year level.

For an average student (for example, average literacy ability and average typing speed), the estimated gap between paper-based and computer-based test results varied from 15 to 20 NAPLAN scaled points, depending on the student's year. This represents roughly 0.2 to 0.3 of one standard deviation in writing results.

Other factors that may (directly or indirectly) contribute to mode effects

There are some limitations to the analysis conducted. Firstly, marking inconsistency impacts on assessment scores and has not been explicitly accounted for in the analysis. However, the impact of general marking inconsistency on the study's results has been minimised to a large extent through random assignment of markers across students and online/paper scripts. Another potential source of bias related to marking inconsistency is text appearance. That is, markers may be more lenient or harsher in their marking of handwritten versus typed scripts, and if the inconsistency is systematic for one mode it will introduce bias to the mode effect estimates. This concern is further explored in the next subsection below.

Possible marking bias due to text appearance (handwritten vs typed)

Some of the gap between computer-based and paper-based marking scores could be as a result of bias in human marking due to text appearance (that is, the tendency of markers giving handwritten scripts higher or lower scores than the same scripts that are typed). Studies investigating this type of bias have indicated mixed findings (Hughes & Akbar, 2010; Mogey, Paterson, Burk, & Purcell, 2010; Tao & Russell, 2004).

This study analysed the 124 handwritten writing responses that were typed and re-marked and found that the majority (111 out of 124) received scores that were within the expected range of normal marking variability, as specified by the national testing authority for marking quality assurance purposes. For the remaining few responses, some were given a lower score and others were given a higher score for the typed responses. Notwithstanding this, the examination of the distribution of the differences in the scores suggests there might be a small tendency of markers favouring handwritten responses. Although it is difficult to estimate the size of this potential bias given the small sample size, it is unlikely to have had a material impact on this study results. It is, however, recommended that further research be undertaken to investigate this potential marking bias as it could impact the comparability of NAPLAN writing scores during the NAPLAN online transition period, when some schools will be participating online and others will still have pen and paper based tests.

Text length

One way in which typing ability impacts computer-based writing performance is in the length of text primary students are able to produce in the time allowed, given evidence from international studies that, in primary years, students' typing fluency generally lags their handwriting fluency, and the gap is greater for younger students.

A comparison of the total word count for a random subset of 82 students' computer-based and paper-based writing responses indicated that, on average, students wrote 19 words less in the computer-based test than they did in the paper-based test. Year 2 and Year 3 students appear to be affected to a greater extent than Year 4 and 5 students, with the younger cohorts both writing 43 words less, on average, in the computer-based test than they did in the paper-based test. For the older cohorts, there was very little difference in the average word counts for the computer-based and paper-based tests. Further studies with larger sample sizes are needed to examine whether these results are replicable.

As with slow handwriting, slow typing speed constrains students' writing quality in multiple ways which includes limiting the cognitive resources available for higher order processes. Whilst there is no minimum level of typing speed required for online writing assessment discussed in the literature, some researchers (Jackowski-Bartol, 2001) suggest that students with low handwriting speeds (below 15 wpm) warrant additional time in examinations.

Low typing speeds are generally associated with the inability to touch type. The student interviews provided evidence of students' limited typing proficiency with 23 out of 37 students interviewed reporting the use of one or two fingers to type. The researchers' observations corroborated this evidence reporting that most students used one or two fingers and very few students used all fingers on both hands.

Research question 3 – What is the impact of the mode change on the writing process (for example, planning, editing, reviewing)?

Whilst there are divergent views amongst educators on the importance of planning and editing as essential skills in the teaching of writing, they are recognised as a part of the writing process and students are encouraged to allocate 5 minutes each to planning and editing during a NAPLAN style writing test. See a relevant extract of the test administration manual used for this research in Appendix 3, which was consistent with the standard NAPLAN writing test manual and that used by ACARA in its online writing test research. For this research, as per the usual test practices, students were provided with planning sheets for both paper-based and computed-based tests and researchers were asked to observe the amount of planning students did for both tests. Editing and reviewing weren't able to be investigated via the methodologies used in this research.

Researchers' observations indicated that there was less evidence of planning by students when undertaking the computer-based tests compared to paper-based tests; however, when students took part in the computer-based testing they were informed that planning on the booklet would not be counted towards marks, whereas any words typed in the pane would be marked. Most students did planning for the paper-based tests. For the computer-based tests, researchers observed that few students did planning on the computer. A manual comparison of the planning sheets used by the same students for both computer-based and paper-based tests indicated that, for many students, there was considerably less planning provided on the planning sheets for the computer-based tests than for the paper tests. An example of the contrast between the planning sheets used by a Year 5 student is provided at Appendix 3.

The contrast appears to be more significant for students who were of a higher writing ability level and who did a substantial amount of planning on paper for the paper tests but considerably less for the computer-based tests. For students of a lower ability level, there was less of a difference in the amount of planning on the planning sheets because they did much less planning for either test. Note these observations were based on a manual comparison of planning sheets and a systematic analysis of the impact on students' scores was not conducted.

Less planning, either directly on the computer or on paper, for the computer-based tests might be a reflection of how teachers have used technology to teach writing online in primary classrooms. Teachers reported through interviews that while they used technology in the classroom regularly, it was often for a whole class collaborative process of researching ideas prior to writing. Students would compose using pen and paper and then use computers individually in the publishing phase. In other words, the teaching focus has not been about teaching students how to use technology to compose a text on the computer in a solitary process, from planning, drafting, editing to reviewing. In students' interviews with researchers, it was also clear that students preferred using the computer for editing their work, but preferred pen and paper for the planning and organising potential.

This raises the broader question of the changing nature of writing and the process of text creation given new communication forms and evolving digital technologies, and the impact this has on the teaching of writing. Further research is planned to investigate how the teaching of writing, and the writing process itself, can be enriched using new technologies to further develop students' writing skills.

Research question 4 – Are there any other factors potentially impacting on students' computer-based writing performance highlighted through teacher and student interviews?

a. School and student technical readiness

Researchers noted considerable variability in technical readiness from school to school. A number of schools, particularly some in low SES areas, had an insufficient number of working computers, limited technical support available, and students who were not as familiar with using computers and accessing the internet. These schools required much greater test preparation time than expected, and these issues combined to impact negatively on the online testing experience for teachers and students.

b. Test environment novelty effects

Teachers and researchers also noted the impact of the novelty of the online test. Student familiarity with the pen and paper writing test environment meant that generally students behaved as expected in a test situation and were able to settle quickly and work quietly for a sustained period of time. For the online test, teachers and researchers reported that students were easily distracted by the unfamiliar environment including the physical layout of the room, the visibility of other students' screens, the test instructions, the computer functions (audio, scrolling, editing) and noise from keyboard typing. With greater exposure to computer-based writing and online testing, students' focus and sustained concentration may improve.

c. The extent to which typing skills are taught in schools

Results from teacher interviews indicated that most schools (25 out of 33) do not explicitly teach keyboarding skills. Of the few schools (8) that indicate some teaching of keyboarding skills, four had implemented touch-typing programs, two had used online typing games and two provided in-class instruction.

Research suggests the earliest time to start formal keyboarding instruction is in the upper primary grades, from age 10 onwards (for example, Erthal, 1998; Freeman et al, 2005). The implication is that the impact of cognitive load and orthographic-motor skills would be too demanding for many Year 3 students with low typing speed to be sufficiently proficient in typing for the NAPLAN test in May.

Given the large proportion of Year 4 and Year 5 students with apparently low typing abilities in this sample, it also suggests that schools should implement typing instruction between Years 3 and 5 so that online writing assessments for students in Year 5 could be a more valid indicator of their writing ability.

d. Disconnect between stated mode preference and performance

Nearly 1,500 participating students responded to an online survey, administered before the typing exercise. Sixty per cent of these students indicated that they thought they could write stories better with computers than with pen and paper. Around 80% indicated it was either easy or very easy to write stories using the computer/laptop/tablet. However, results from the analyses indicate a conflicting picture – that on average these students performed worse in online tests than on paper. This is not necessarily surprising, as other studies also reveal a 'Performance Paradox' (Oviatt, Arthur, & Cohen, 2006; Oviatt & Cohen, 2010) or a mismatch between the mode people say they prefer and the mode that best supports their performance. The conflicting picture indicates that great care needs to be exercised when interpreting students' stated preference for the test mode as this might not be an indicator of how ready students are for online tests, despite how they feel.

Summary of findings

The results of the various statistical analyses conducted in this study all confirm that, on average, primary students in NSW government schools score lower on computer-based writing tests than they do on paper-based tests. This mode effect was observed for both writing genres, and for students across all year levels examined (Years 2-5), both genders, and Aboriginal status. While initial results suggested that the mode effect was not as great for males and Aboriginal students as for their counterparts, these differences were reduced when typing ability was taken into account. The mode effect was also found to be somewhat less pronounced for students of low literacy ability and was more pronounced for students of high literacy ability, after controlling for typing ability.

In addition, the study found that typing speed was significantly related to the difference between students' computer-based and paper-based writing scores. For students with similar literacy ability, the faster that they could type, the smaller the difference between their computer-based and paper-based results.

Overall, for an average student (with average literacy and typing ability for their year level), the gap between paper-based and computer-based test results varied between -15 and -20 scaled points depending on the year level. This represents roughly 0.2 to 0.3 of one standard deviation in writing results.

One difference observed between students' computer-based and paper-based writing is the level of planning that students undertake prior to writing. Analysis of student planning sheets and researcher observations both seemed to indicate that students undertook less planning for their computer-based writing task than they did for their paper-based task.

In summary, the results of the study suggest that computer-based writing results may not be a reliable indicator of primary students' overall writing ability, as they are not necessarily comparable to paper-based writing results, unless students can adequately deal with the demands of composing in a digital environment.

Conclusions and recommendations

The move to online assessment of writing brings many advantages. At a national level it provides more manageable administration of large-scale standardised testing, reduced time for test set up, marking and reporting. At a school level, it has potential to provide more timely, targeted, personalised feedback about student writing – engaging students and teachers more meaningfully in the writing process.

However, the results of the present study cast doubt on whether computer-based writing is sufficiently comparable to paper-based writing, given the differences in the skillsets being assessed through the writing tests administered in different modes. This is particularly an issue for the NAPLAN online transition period during which some schools will conduct assessments online, and others will still be using pen and paper tests. Teachers and schools would always benefit from analysing individual students' writing results, whether computer-based or paper-based. However, the comparability of aggregate reporting (for example, jurisdictional average results and percentages in bands), during the transition years as well as trends over time (that is, comparing data from pre and post transition years), is questionable, particularly in the earlier years. Given the mode effect appears to impact different groups of students differently, it is also questionable whether the current equating process adopted by the national testing authority for the NAPLAN tests is sufficient for the purpose of equating online test results to the previous paper test results. Unless these issues can be addressed, caution is urged when making comparisons of performance across schools and jurisdictions in the transition years when the tests are run in dual modes or when comparing trend data from pre transition years to post transition years.

This study also presented results on primary students' typing abilities, about which little is known in the Australian context. At the median typing speed of 9 words per minute, which the literature suggests is lower than the handwriting speed for that age group, many Year 3 students would struggle to produce online texts comparable to handwritten texts in a timed condition. Given the recommendation in the literature that typing instruction is best commenced in the upper primary years and evidence of cognitive and educational benefits of teaching handwriting to students in early learning years, it is recommended Year 3 students continue to participate in NAPLAN writing tests on pen-and-paper.

Findings from this study also have implications for the work schools and teachers need to do in order for students to perform to their ability level in online assessments of writing.

School technical readiness was found to vary across the participating schools in different areas and of different SES levels. More disadvantaged schools appeared to struggle to a greater extent with the administration of the online tests, which could have resulted in a more negative test experience for students. To ensure the online tests are accessible to all students, it is recommended that schools access the resources available from ACARA for the online tests (for example, practice tests on the public demonstration site). Note that following this 2016 research, the department has initiated a cross-sectoral online transition project to support schools to become technically ready for online assessment.

Consistent with other studies, typing skills have been identified as a contributor to students' performance in online writing tasks, after holding constant other factors including students' literacy ability levels. Examination of the typing data indicates a great variability in the typing speed from school to school and from student to student in the same classroom in the same school. Schools should consider their local contexts and identify an effective method for developing students' typing fluency and to monitor the development of their typing proficiency over time.

This study raises important questions that relate to the impact of new technologies on the teaching of writing. Findings from the teacher interviews suggest that technology is only being used in limited ways in the classroom in the teaching and learning of writing. However, many students have considerable expertise in using technology outside of school, using personal computers and devices, for example, to access social media, send text messages and play video games. In schools, there is a need to develop effective pedagogy with a stronger focus on teaching students how to better use technology to develop their writing.

In this regard, further research is planned in NSW schools to explore how technology can best be employed to combine traditional writing skills with innovations in digital design and delivery, as well as the explicit teaching of keyboard and typing skills, to prepare students to become fully literate in contemporary society.

References

Alstad, Z., Sanders, E., Abbott, R. D., Barnett, A. L., Henderson, S. E., Connelly, V., & Berninger, V. W. (2015). Modes of alphabet letter production during middle childhood and adolescence: Interrelationships with each other and other writing skills. *Journal of Writing Research, 6*(3), 199 -231. doi:10.17239/jowr-2015.06.03.1

Alves, R. A., Castro, S. L., & Olive, T. (2008). Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International Journal of Psychology, 43*(6), 969-979. doi:10.1080/00207590701398951

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi:10.1007/BF02293814

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*(2), 241-259. doi:10.1177/0265532213509810

Berninger, V. W., Abbott, R. D., Augsburger, A., & Garcia, N. (2009). Comparison of pen and keyboard transcription modes in children with and without learning disabilities. *Learning Disability Quarterly, 32*(3), 123-141. doi:10.2307/27740364

Bisschop, E., Morales, C., Gil, V., & Jiménez-Suárez, E. (2016). Fluency and Accuracy in Alphabet Writing by Keyboarding: A Cross-Sectional Study in Spanish-Speaking Children With and Without Learning Disabilities. *Journal of Learning Disabilities. 50*(5). 534-542. doi:10.1177/0022219416633865.

Boyd, S., & McDowall, S. (2001). *Techno-magic: Whizz or fizz? The relationship between writing mode, editing process, and writing product.* Wellington: New Zealand Council for Educational Research.

Broad, B. (2006). More work for teacher? Possible futures of teaching writing in the age of computerized assessment. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp. 221–233). Logan, UT: Utah State University Press.

Carr, N. (2000). A comparison of the effects of analytic and holistic composition in the context of composition tests. *Issues in Applied Linguistics, 11*(2), 207-241. Retrieved from: http://escholarship.org/

Chwirka, B., Gurney, B., & Burtner, P. (2002). Keyboarding and visual-motor skills in elementary students: A pilot study. *Occupational Therapy In Health Care, 16*(2-3), 39-51. doi:10.1080/J003v16n02_03

Christensen, C. A. (2004). Relationship between orthographic-motor integration and computer use for the production of creative and well-structured written text. *British Journal of Educational Psychology, 74*(4), 551-564. doi:10.1348/0007099042376373

Cohen, R. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Connelly, V., Gee, D., & Walsh, E. (2007). A comparison of keyboarded and handwritten compositions and the relationship with transcription speed. *British Journal of Educational Psychology, 77*(2), 479-492. doi:10.1348/000709906X116768

Dinehart, L., & Manfra, L. (2013). Associations between low-income children's fine motor skills in preschool and academic performance in second grade. *Early Education and Development, 24*(2), 138-161. doi:10.1080/10409289.2011.636729

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Edwards-Groves, C. J. (2011). The multimodal writing process: Changing practices in contemporary classrooms. *Language and Education, 25*(1), 49-64. doi:10.1080/09500782.2010.523468

Erthal, M.(1998). *Who should teach keyboarding and when should it be taught?* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.585.4719

Eyre, J., (2015). Dual modes for written assessments: Examining validity and reliability. *Assessment Matters*, 9, New Zealand Council for Educational Research, Wellington.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College composition and communication. 32*(4), 365-387. Retrieved from https://www.jstor.org/

Fluck, A., Pullen, D. & Harper, C. (2009). Case study of a computer based examination system. *Australasian Journal of Educational Technology*, 25(4), 509-523. Retrieved from: https://www.ebsco.com/

Foy, P. (1995). Calculation of sampling weights. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report Volume II: Implementation and Analysis (Primary and Middle School Years)* (pp. 71-79). Chestnut Hill, MA: Boston College.

Freeman, A. R., Mackinnon, J. R., & Miller, L. T. (2005). Keyboarding for students with handwriting problems. *Physical & Occupational Therapy In Pediatrics, 25*(1-2), 119-147. doi:10.1080/J006v25n01_08

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *The Journal of Technology, Learning, and Assessment, 2*(1). Retrieved from https://ejournals.bc.edu/ojs/index.php/

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal Of Educational Psychology, 89*(1), 170-182. doi:10.1037/0022-0663.89.1.170

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools*. Washington, DC: Alliance for Excellence in Education.

Herrington, A., & Moran, C. (2006). Writeplacer Plus in place. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp. 115–127). Logan, UT: Utah State University Press.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 5*(2). Retrieved from https://ejournals.bc.edu/ojs/index.php/

Horne, J., Ferrier, J., Singleton, C., & Read, C. (2011). Computerised assessment of handwriting and typing speed. *Educational & Child Psychology, 28*(2), 52-66. Retrieved from https://www.ebsco.com/

Hughes, J. & Akbar, S. (2010, August). *The influence of presentation upon examination marks*. Paper presented at 11th Annual Conference of the Subject Centre for Information and Computer Sciences, Durham University. Retrieved from http://oro.open.ac.uk/47178/

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263. Retrieved from https://www.jstor.org/

Jackowski-Bartol, T. R. (2001). The impact of word processing on middle school students. Unpublished master's thesis, Chestnut Hill College, Philadelphia.

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practive, 23*(4), 495-518. doi:10.1080/0969594X.2016.1147420

Kervin, L., Verenikina, I., Jones, P., & Beath, O. (2013). Investigating synergies between literacy, technology and classroom practice. *Australian Journal of Language and Literacy, 36*(3), 135-147. Retrieved from https://www.ebsco.com/

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418. doi:10.1111/j.1745-3984.2011.00152.x

Lim, G. (2009). *Prompt and rater effects in second language writing performance assessment*. (Unpublished Doctoral dissertation). The University of Michigan, Ann Arbor, MI. Retrieved from http://hdl.handle.net/2027.42/64665

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. (Doctoral dissertation). Chicago, IL: MESA Press. Retrieved from http://www.proquest.com/

Linacre, J. M. (2008). *Facets Rasch model computer program* [Software program manual]. Chicago, IL: Winsteps. Retrieved from http://www.winsteps.com/facets.htm

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 484-509. Retrieved from http://jampress.org

Longcamp, M., Zerbato-Poudou, M. T., & Velay, J. L. (2005). The influence of writing practice on letter recognition in preschool children: A comparison between handwriting and typing. *Acta psychologica, 119*(1), 67-79. doi:10.1016/j.actpsy.2004.10.019

MacArthur, C. A. (1999). Overcoming barriers to writing: Computer support for basic writing skills. *Reading and Writing Quarterly, 15*(2), 169-192. doi:10.1080/105735699278251

MacArthur, C. A. (2009). Reflections on research on writing and technology for struggling writers. *Learning Disabilities Research & Practice, 24*(2), 93-103. doi:10.1111/j.1540-5826.2009.00283.x

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. doi:10.3102/0013189X018002005

Mills, K. A., & Exley, B. (2014). Time, space, and text in the elementary school digital writing classroom. *Written Communication, 31*(4), 434-469. doi:10.1177/0741088314542757

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: Letting the students choose. *Research in Learning Technology, 18*(1), 29-47. doi:10.3402/rlt.v18i1.10675

Morphy, P., & Graham, S. (2012). Word processing programs and weaker writers/readers: A meta-analysis of research findings. *Reading and Writing, 25*(3), 641-678. doi:10.1007/s11145-010-9292-5

Oviatt, S., Arthur, A., Cohen, J. (2006). Quiet interfaces that help students think. In UIST 2006 Proceedings of the 19th Annual ACM Symposium on User Interface Software Technology (pp. 191-200). New York: ACM Press. doi:10.1145/1166253.1166284

Oviatt, S., L., & Cohen, A., O. (2010). Toward high-performance communications interfaces for science problem solving. *Journal of Science Education and Technology, 19*(6), 515-531. doi:10.1007/S10956-010-9218-7

Poole, D. M. & Preciado, M. K. (2016). Touch typing instruction: Elementary teachers' beliefs and practices. *Computers & Education, 102*, 1-14. doi:10.1016/j.compedu.2016.06.008

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3), 1-20. doi:10.14507/epaa.v5n3.1997

Russell, M., & Plati, T. (2000). *Mode of administration effects on MCAS composition performance for grades four, eight, and ten*. Retrieved from http://www.bc.edu/

Russell, M., & Plati, T. (2001). *Effects of computer versus paper administration of a state-mandated writing assessment*. Chestnut Hill, MA: Technology and Assessment Study Collaborative. Retrieved from http://www.bc.edu/.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modelling. *Language Testing, 22*(1), 1-30. doi:10.1191/0265532205lt295oa

Sweller J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312. doi:10.1016/0959-4752(94)90003-5

Sweller, J., van Merrienboer, J. & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296. doi:10.1023/A:1022193728205

Tao, W., & Russell, M. (2004). Effects of handwriting and computer print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research & Evaluation, 9*(1), 1-9. Retrieved from https://doaj.org

Walsh, M. (2010). Multimodal literacy: What does it mean for classroom practice? *Australian Journal of Language & Literacy, 33*(3), 211-239. Retrieved from https://www.ebsco.com/

Walsh, M., & Simpson, A. (2013). Touching, tapping… thinking? Examining the dynamic materiality of touch pad devices for literacy learning. *Australian Journal of Language and Literacy, 36*(3), 148-157. Retrieved from https://www.ebsco.com/

Warschauer, M. (2006). *Laptops and literacy: Learning in the wireless classroom.* New York, NY: Teachers College Press

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

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.* Washington, DC: National Center for Education Statistics, Institute of Education Sciences, and U.S. Department of Education. Retrieved from https://nces.ed.gov

Wright, B. D., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range. Retrieved from https://www.rasch.org

Zou, X., & Chen, Y. (2016). Effects of test media on different EFL test-takers in writing scores and in the cognitive writing process. *Technology, Pedagogy and Education, 25*(1), 79-99. doi: 10.1080/1475939X.2014.954140

Appendix 1 –
Technical details for MFRM

Many Facet Rasch Measurement (MFRM) (details see Linacre, 1989; Linacre & Wright, 2002) is an expanded form of the Rasch Rating Scale model (Andrich, 1978). The below equation specifies the MFRM measurement model used in this study:

$$\log\left(P_{ijklm}/P_{ijkl(m\text{-}1)}\right) = \theta_i - C_j - D_k - E_l - \delta_m$$

where:

P_{ijklm} is the probability of person i achieving a writing score of m, for writing prompt j, produced via mode k given task order l;

$P_{ijkl(m\text{-}1)}$ is the probability of person i achieving a writing score of m–1, for writing prompt j, produced via mode k given task order l;

θ_i is the ability of person i;

C_j is the difficulty of prompt j;

D_k is the difficulty of mode k (that is, paper-based or computer-based);

E_l is the difficulty of task order l (that is, task done first or second)

δ_m is the threshold of being scored in score category m relative to category (m–1)

This measurement model calibrates the person ability, Rasch-Andrich category thresholds (see Andrich, 1978), and the difficulty of task related facets (that is, mode, prompt, task order) in one statistical and measurement framework. Model parameters were estimated using the Joint Maximum Likelihood Estimation procedure in FACETS (Linacre, 1989).

The model provides two types of fit statistics which are indicators of how well an observed response pattern fits the measurement expectations: INFIT and OUTFIT mean-squares (Wright & Stone, 1999). Both are chi-square ratios based on the standardised residuals. While the OUTFIT statistic is an unweighted statistic which is heavily influenced by outlying, off-target, unexpected responses, the INFIT is sensitive to irregular inlying patterns with relatively more impact being given to unexpected responses close to a person's or item's measure. Both mean-square statistics have an expected value of 1.0, and a range from 0 to positive infinity. Values less than 1.0 indicate over-fit; that is, data is too predictable with respect to model expectations, causing summary statistics such as reliability indices, to report inflated results. Values greater than 1.0 indicate under fit; that is, there is more un-modelled noise in the data than expected. High mean-squares are considered a much greater threat to the validity than low mean-square values, because they suggest a possible violation of the uni-dimensionality requirement. Fit statistics included in Table 4 show that the data fit the model sufficiently well.

Appendix 2 – **Technical details for mixed effects model for repeated measures**

Model 1 (referenced on page 24)

Mixed effects models can be used to analyse results from a repeated measures design in which the outcome is continuous and measured at fixed time points. In our analysis, the outcome is defined as the writing test score of student $\boldsymbol{i}$ in school $\boldsymbol{s}$ at time $\boldsymbol{t}$ (Y_{tis}) where each student has performed two writing tests at two different time points ($\boldsymbol{t}$ = 1, 2) and had a writing test score recorded at each time point. The writing test score is estimated as a linear function of independent variables, random student effects, random school effects and some random errors. The three-level mixed effects model for repeated measures is specified as:

$$Y_{tis} = \boldsymbol{XB} + \boldsymbol{WC} + \boldsymbol{ZG} + u_s + v_{is} + e_{tis} \qquad (1)$$

where $\boldsymbol{u_s}$ is the school random effect which captures any between-school differences in the writing test scores; $\boldsymbol{v_{is}}$ is the student random effect which captures any between-student differences within the same school; $\boldsymbol{e_{tis}}$ is the residual error which represents any time-specific variations within the same student. XB, WC and ZG are the linear functions of time-level variables, student-level variables and school-level variables respectively as specified below:

i. Level 1 (time level):

$$\boldsymbol{XB} = b_0 + b_1X_{1tis} + b_2X_{2\_1tis} + b_3X_{2\_2tis} + b_4X_{2\_3tis} + b_5time + b_6X_{1tis}\text{x}W_{1is} + b_7X_{1tis}\text{x}W_{2is} + b_8X_{1tis}\text{x}W_{3is} + b_9X_{1tis}\text{x}W_{4\_1is} + b_{10}X_{1tis}\text{x}W_{4\_2is} + b_{11}X_{1tis}\text{x}W_{4\_3is} + b_{12}time\text{x}W_{4\_1is} + b_{13}time\text{x}W_{4\_2is} + b_{14}time\text{x}W_{4\_3is}$$

ii. Level 2 (student level):

$$\boldsymbol{WC} = c_1W_{1is} + c_2W_{2is} + c_3W_{3is} + c_4W_{4\_1is} + c_5W_{4\_2is} + c_6W_{4\_3is}$$

iii. Level 3 (school level):

$$\boldsymbol{ZG} = g_1Z_{1s} + g_2Z_{2s}$$

The regression coefficients $b_0, \ldots, b_{14}, c_1, \ldots, c_6, g_1$ and g_2 are estimated through maximum likelihood method. Student weight is applied to the second level of the mixed effects model whereas school weight is applied at the third level. The set of independent variables are listed below:

X_{1tis} = 1 for computer-based writing test and 0 for paper-based writing test;

X_{2_1tis} = 1 for Narrative 2009 Box writing task and 0 otherwise;

X_{2_2tis} = 1 for Persuasive 2011 Toys writing task and 0 otherwise;

X_{2_3tis} = 1 for Persuasive 2012 Cook writing task and 0 otherwise;

time = 1 for test performed at time 1 and 2 for tests performed at time 2;

W_{1is} represents the students' general literacy ability standardised within each year level;

W_{2is} = 1 for male students and 0 for female students;

W_{3is} = 1 for Aboriginal students and 0 for non-Aboriginal students;

W_{4_1is} = 1 for Year 3 students and 0 otherwise;

W_{4_2is} = 1 for Year 4 students and 0 otherwise;

W_{4_3is} = 1 for Year 5 students and 0 otherwise;

Z_{1s} represents the ICSEA index for school s;

Z_{2s} = 1 for schools in non-metropolitan (provincial, remote and very remote) regions and 0 otherwise.

In the mixed effects model specified in Equation (1), the residual error (e_{tis}) is assumed to follow a normal distribution with zero mean and a constant variance σ_e^2. The random student effect (v_{is}) and random school effect (u_s) are also assumed to be of a normal distribution with zero mean and constant variances σ_v^2 and σ_u^2 respectively. The three components are assumed to be uncorrelated with each other and the random school and student effects are assumed to be uncorrelated with the independent variables.

For diagnostic checking, the residual error (e_{tis}) is plotted against the fixed linear prediction ($XB + WC + ZG$) to see if the residual error is randomly distributed with constant variance. The normality assumption is confirmed by outputting a Q-Q plot which plots the quantile of the residuals against the quantiles of a normal distribution. The Q-Q plot should show a straight line pattern if the residual errors are normally distributed. Diagnostic checks confirmed that the normality assumption of the residual errors was satisfied.

Model 2: with typing ability (referenced on page 29)

The specification is the same as the above model, but with two additional explanatory variables – typing accuracy rate and typing speed (words per minute), plus two interaction terms between task mode and typing variables. The level 1 and level 2 link functions are re-specified as follows:

i. Level 1 (time level):

$$\boldsymbol{XB} = b_0 + b_1X_{1tis} + b_2X_{2\_1tis} + b_3X_{2\_2tis} + b_4X_{2\_3tis} + b_5time + b_6X_{1tis}\mathrm{x}W_{1is} + b_7X_{1tis}\mathrm{x}W_{2is}$$
$$+ b_8X_{1tis}\mathrm{x}W_{3is} + b_9X_{1tis}\mathrm{x}W_{4\_1is} + b_9X_{1tis}\mathrm{x}W_{4\_2is} + b_{10}X_{1tis}\mathrm{x}W_{4\_3is}$$
$$+ b_{11}time\mathrm{x}W_{4\_1is} + b_{12}time\mathrm{x}W_{4\_2is} + b_{13}time\mathrm{x}W_{4\_3is} + b_{14}X_{1tis}\mathrm{x}W_{5is}$$
$$+ b_{15}X_{1tis}\mathrm{x}W_{6is}$$

ii. Level 2 (student level):

$$\boldsymbol{WC} = c_1W_{1is} + c_2W_{2is} + c_3W_{3is} + c_4W_{4\_1is} + c_5W_{4\_2is} + c_6W_{4\_3is} + c_7W_{5is} + c_8W_{6is}$$

where W_{5is} represents the students' typing accuracy and W_{6is} represents the students' typing word count per minute.

The same model diagnostic checks were performed for Model 2. The random pattern in the residual plot and the straight line pattern in the Q-Q plot indicated the residuals were normally distributed with zero mean and constant variance. Furthermore, correlation analyses show that the correlations amongst pairs of student-level random effects, school-level random effects, residual errors and the linear predictions of the fixed portion of the model ($XB + WC$) are negligible.

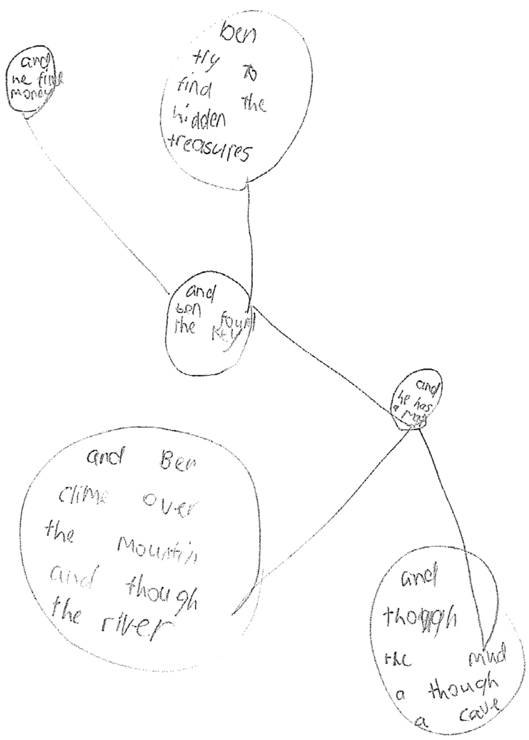
Appendix 3 – **Test instructions relating to planning and editing, and an example of planning sheets**

a. Instructions to students at the commencement of the writing test

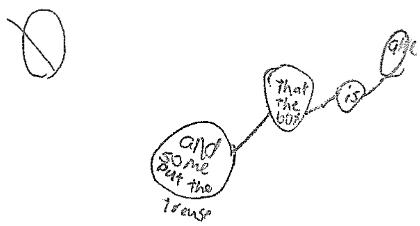
Teacher to read aloud:

> Before you start writing on your computer, laptop or tablet, you may want to do some planning.
>
> You could plan by writing your ideas using the planning booklet or by using your device.
>
> The planning booklet will not be marked, but it will be collected. Anything you type into the writing pane of your device will be marked.
>
> I recommend using the first 5 minutes of your test time to do your planning. I will let you know when 5 minutes is up so you can move from planning to writing.
>
> You should spend the last 5 minutes of your time editing your work. I will let you know when there are 5 minutes left.
>
> Once the test starts you should see a clock on the top of the screen telling you how much time you have left.
>
> You will have 40 minutes to do this test.

b. example of the difference in the planning sheets used by a Year 5 student for the paper and online tests



Paper-based test



Computer-based test

Centre for Education Statistics and Evaluation
GPO Box 33, Sydney NSW 2001, Australia

Visit our website to subscribe to the CESE newsletter

02 7814 1527
education.nsw.gov.au/cese
info@cese.nsw.gov.au
yammer.com/det.nsw.edu.au

226_221021GS_v2_AA2