acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

National Assessment and Surveys Online Program

Tailored test design study 2013: Summary research report

National Assessment and Surveys Online Program

Tailored test design study 2013: Summary research report

Project management

ACARA's online research program is directed by Dr Goran Lazendic. The National Assessment and Surveys Online Program is managed by Julie-Anne Justus, John Skehan, Dominik Swierk, Justine Tyrrell and Rosemary Vo. Item and test development is managed by Anna Cohen (reading) and Stephen Phillip (numeracy). Dr Kelvin D. Gregory, Stephen Phillip and Anna Cohen contributed to the research presented in this report.

This research is funded by the Australian Government Department of Education.

Report author(s)

The *Tailored test design study 2013: Summary report* was written by Dr Goran Lazendic, with input from Julie-Anne Justus and Dr Stanley Rabinowitz.

National Assessment and Surveys Online Program

The National Assessment and Surveys Online Program is designed to deliver national assessments and surveys online. ACARA is responsible for planning and implementing a clearly defined assessment and reporting research agenda that will allow reporting to Education Council on issues and options for delivering NAPLAN online. A key aspect of the program is ACARA's expanded assessment and reporting research agenda, incorporating a comprehensive investigation into assessment instruments and programs using online technology.

Acknowledgements

The appropriate citation for this report is:
Australian Curriculum, Assessment and Reporting Authority (ACARA), September 2014, *Tailored test design study 2013: Summary research report,* ACARA, Sydney.

Background

In June 2012 the Australian Government Department of Education funded the Australian Curriculum, Assessment and Reporting Authority (ACARA) to conduct research to inform decisions on the transition of the National Assessment Program (NAP), including the National Assessment Program – Literacy and Numeracy (NAPLAN), from paper-based to computer (online) assessments. ACARA developed a comprehensive research program to examine and respond to a range of transition issues, including test design and impacts on student performance, measurement and reporting. In December 2012 Australian education ministers accepted ACARA's research program as the base for further detailed work on transitioning NAP to a computer-based assessment and the enhancement that such a transition could bring to future NAP assessments.

The initial phase of research, undertaken from September 2012 through to October 2012, consisted of a pilot study to examine how the test delivery mode (paper or computer) affected student performance across year levels, including the impact of using a keyboard to complete a writing assessment. This phase also involved interviewing students to explore their level of cognitive and behavioural engagement with the computer-based assessment. Results from this study support the claim that students at all year levels are capable of engaging with the current NAPLAN tests delivered online and that the transition of items to the online delivery mode would not substantially change the assessment properties of the NAPLAN tests.

This paper describes follow-up research conducted in 2013. In particular, ACARA examined the feasibility of a type of computerised adaptive testing designed to enhance the targeting of NAPLAN tests to the individual student's ability and learning needs. This study also investigated how students interacted and engaged with these computerised, adaptive tests.

Computer adaptive testing and the ‘tailored test design’

In computer adaptive testing, a computer algorithm adjusts the difficulty of the tests to match the ability (that is, achievement level) of each student. In a standard computer adaptive test, the difficulty of the test is adjusted after response to each item presented to a student. Consequently, the final test represents a unique combination of items that targets student ability with the highest possible measurement precision using a relatively smaller number of items to achieve such an outcome.

An alternative to fully computer adaptive testing is multistage adaptive testing, where test difficulty is adjusted after a student provides responses to a set of items (see Lord, 1971). Consequently, a student progresses through a series of stages containing item sets of varying difficulties to complete the test. In educational assessment these item sets are called testlets, with each testlet being a self-contained set of items that reflects the composition of a complete test in terms of test content and specification coverage. Multistage tests are typically longer than fully computer adaptive tests. They provide a considerable increase in measurement precision relative to that achieved in static paper tests, although typically not as great as can be achieved in fully adaptive tests.

Advantages of item-level computer adaptive testing over paper and multistage testing are well-documented in educational research (for example, Kingsbury, McCall & Hauser, 2009). However, a computer adaptive test requires an item bank containing a relatively large number of items to cover the whole range of assessed content and to prevent overexposure of items targeting the most frequent ability levels of students. In addition, if the assessment has to cover different strands of the main assessment domain, then a separate item bank needs to be constructed and administered for each of these content strands. These requirements pose considerable logistical and cost challenges for development, implementation and long-term maintenance of tests in large-scale educational assessments, limiting the degree of adaptability that can be achieved in CAT (computer adaptive testing).

Multistage tests, while not providing a full student-level adaptation, have a number of advantages over computer adaptive tests in the context of large-scale assessment programs

such as NAPLAN. These advantages include better control over the administration and structure of the final tests (including item content and cognitive demands, as well as alignment to the test blueprint), better control over the exposure of items and tests, and the capacity for students to preview and review items and to change their answers (see Hendrickson, 2007; Van der Linden & Glas, 2010). In addition, the number of items required to implement and maintain multistage tests is typically smaller than that required to maintain standard CAT tests.

These conceptual advantages and logistical benefits have led ACARA to propose that future NAPLAN online tests implement a multistage adaptive test design – ACARA's 'tailored test design'. The proposed tailored test design (TTD) for future NAPLAN online tests represents a solution that will both provide adequate tailoring of tests to students' learning needs and be feasible in terms of logistical and cost requirements for the development, implementation and long-term maintenance of NAPLAN as a computer-based assessment program.

The proposed tailored test design consists of three stages and thus has two branching points, as illustrated in Figure 1.

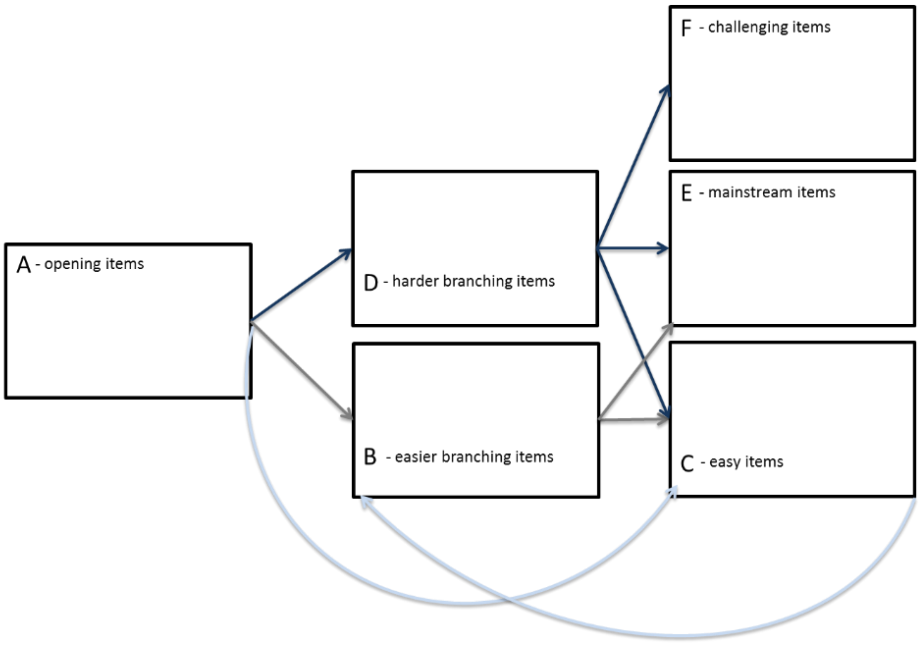

Figure 1: Tailored test design. This figure illustrates testlets and six test pathways available in TTD

All proposed NAPLAN test pathways are shown by the arrows in Figure 1. To complete a test, each student goes through three testlets. Each testlet contains approximately one-third of the total number of items in a test. In addition, each testlet is representative of the whole test in terms of knowledge and content coverage. Consequently, each student is assessed at the same level of domain breadth regardless of the test pathway taken.

Unlike the standard multistage test design, the tailored test design offers a pathway to assist students struggling with items in testlet A to engage with the rest of the test. This test pathway takes students who correctly answer few, if any, items in testlet A directly to testlet C. For some students, this provides an early opportunity to engage with the easiest set of test items. Once these students respond to testlet C, they are re-routed back to testlet B so that they have the opportunity to demonstrate the extent of their knowledge.

The goals of the proposed test design are to ensure:

- comprehensive measurement of the main student cohort, by extending the curriculum and content coverage
- better measurement of the proficiency of each student, including
    - better targeting of higher performing students, by providing more challenging items and test content
    - better targeting of underperforming students by tailoring tests to their ability and thus increasing the opportunity to collect information about factors that prevent these students from reaching their full potential
- assessment of a broader set of skills and a wider range of content, without increasing the number of items in the test taken by each student. The better-targeted test design can assess student ability with fewer items, leaving scope to expand on the content of the final testlets
- the ability to identify gaps that students might be carrying from previous years of schooling. This will be achieved by the inclusion in the relevant testlet of a proportion of items targeting earlier year levels.

The proposed tailored test design may enhance the stability of future NAPLAN online assessment scales by increasing the number of available link items between tests covering the adjacent NAPLAN year levels. The tailored test design may also provide an opportunity to expand on methods used to conduct longitudinal equating between test cycles. This has the potential to significantly strengthen the stability of the longitudinal equating and thus the long-term stability of NAPLAN assessment scales.

Feasibility of the tailored test design

In order to investigate the feasibility of the proposed test design, a series of pilot simulation studies was conducted (Adams and Lazendic, 2013; Lazendic and Adams, 2014). Outcomes in the tailored test design were simulated using the actual distribution of student ability and the hypothesised ideal distribution of testlets and item difficulties. These simulations show that the proposed tailored test design is a feasible multistage test design and that it is likely that such a design will considerably improve the precision of student achievement measurement.

ACARA investigated the feasibility of the proposed test design through a field trial delivered in schools from all states and territories. The primary purpose of 'the tailored test design study' was to collect empirical evidence about the performance of the proposed multistage branching test design for NAPLAN reading and numeracy tests. To that end, two test conditions were constructed. In the first test condition, the branching condition, students completed a multistage NAPLAN test. In the second test condition, the fixed condition, students took a fixed linear test corresponding to one of the possible test pathways of the multistage design. Within a school students were randomly assigned to the each of the test conditions. The performance of items and students in the fixed linear test condition served as the baseline to determine the effectiveness of the proposed multistage test design regarding the performance of the branching mechanism and the increase in measurement precision of person–ability estimates.

More than two hundred and fifty schools participated voluntarily in the tailored test design study. The sample contained schools from all Australian school sectors, across all states and territories, including some remote and very remote schools. Although being self-selected and

convenient, the resultant sample demonstrated a satisfactory diversity of students. Students in the sample completed either numeracy or reading tests in one of the two testing conditions. Over 2500 students in Years 3 and 5, and 1500 students in Years 7 and 9 participated in each of the two tests.

Testlets used in this study were created from existing NAPLAN test items that had been rendered to suit the delivery mode. Minor adjustments to the items, including the repositioning and resizing of the item stems and/or item illustrations where applicable, were made. The position of the multiple-choice options or text response box was also altered for some items. For the reading tests, the computer screen was vertically divided into two halves. The stimulus reading material was shown in the left-hand pane, while the right-hand pane displayed the item and response choices. Students were able to extend the reading pane horizontally where necessary and were also able to scroll up or down to read multipage prompts.

ACARA's content and assessment experts adjusted the items and allocated them to the relevant testlets based on the item location on the NAPLAN scale and the item facility rate. For reading, there were insufficient existing items to construct testlets C and F and so testlets containing newly developed items were used instead. Given that these testlets were not used in the branching algorithm, the inclusion of new items in the test did not affect the ability to collect information about the efficacy of the tailored test design.

The distribution of the initial item difficulties used to construct testlets for the Year 3 numeracy test against the distribution of the Year 3 students is provided in Figure 2. In addition to item difficulty, each testlet covered all assessed skills and content strands in its respective test domain.

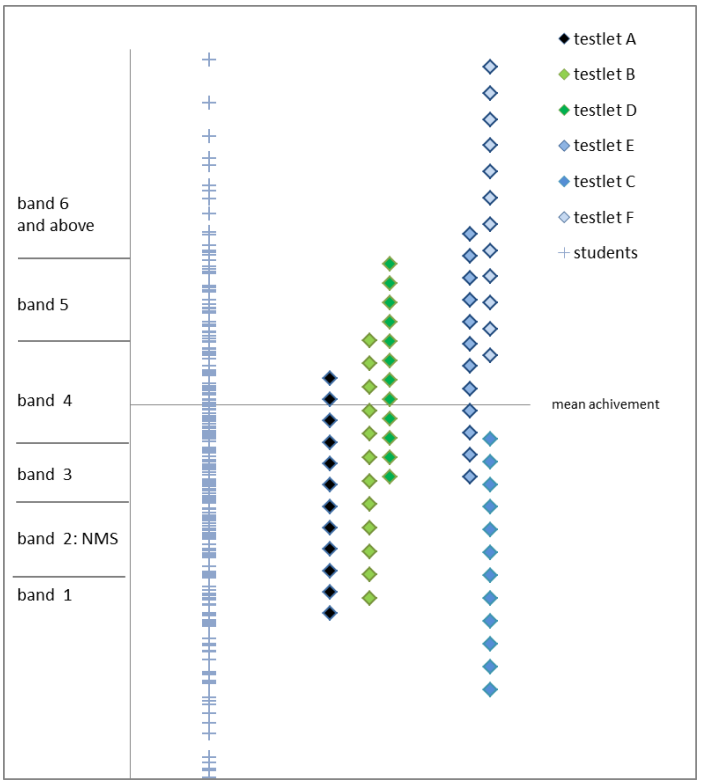

Figure 2: Spread of Year 3 numeracy items and testlet difficulty against the distribution of Year 3 student ability against the current NAPLAN scale

The blue crosses in the left column of Figure 2 represent the actual distribution of student abilities drawn for the nationally representative sample of Year 3 students who completed 2011 NAPLAN numeracy test. The ability is shown on the established NAPLAN proficiency scale and bands for numeracy, including the position of national minimum standard (NMS). The item difficulty based on the existing NAPLAN item location is represented by a diamond for each item. Per our study blueprint, there is significant overlap in testlet difficulty between different testlets both within and across the three stages of the tailored test design.

The positions of NAPLAN performance bands for numeracy tests are provided on the vertical axis in Figure 2. As can be seen, the boundaries of testlets are set in such a way so that students can demonstrate knowledge across several performance bands, irrespective of test pathway. For example, testlet C covers band 1, band 2 (set as the national minimum standard

for Year 3), band 3 and even the lower edge of band 4. Similarly, the top performance band (band 6 for Year 3) is available to students who are assigned to either testlet E or testlet F.

In order to create a base condition to examine the increase in the precision of measurement offered by the proposed test design, a set of four linear tests was created for use in the fixed test condition for numeracy and reading tests. The four linear tests corresponded to the main test pathways available to students in the branching test condition (A-D-F, A-D-E, A-B-E and A-B-C). The test pathways A-D-C and A-C-B were not translated into linear tests because they had a corrective purpose (that is, to correct for errors in the branching) in the multistage design, which is not focus of the study presented in this paper.

Students within a school were randomly assigned to either the branching or the fixed test condition. For the fixed test condition, students received one of four linear tests. Each linear test was numbered and allocated to students using a rotating random design, which ensured that the same number of students took each of the four available tests.

Students were able to navigate freely through all items within a testlet and change their responses to earlier items within the same testlet. However, once routed to the next testlet, students were not able to access the items from the previous testlet(s). Students were given a warning before being routed to the next testlet and had an opportunity to check their responses before continuing the test.

Results of the tailored test design study

Full psychometric results of this study are available in the report prepared by the Australian Council for Educational Research.[1]

[1] Australian Council for Educational Research (ACER) December 2013, *Analytical Report: Psychometric Analysis for the Trial of the Tailored Test Design,* ACER, Melbourne.

This paper provides outcomes of the investigation into Year 5 reading tests (the other year levels studied provided similar findings).

The results of the implementation of the fully branching test in the tailored test design are provided in Figure 3.

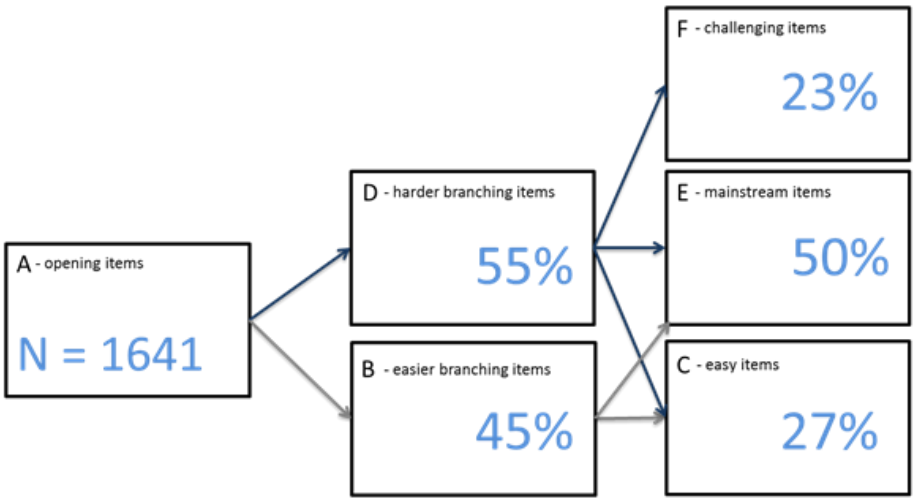

Figure 3: Outcome of branching for Year 5 TTD reading test

As Figure 3 shows, 50 per cent of students who participated in the study were routed to the mainstream set of items, 23 per cent of students reached the testlet containing the most challenging items and 27 per cent of the students finished with the testlet containing the easiest set of items.

The effectiveness of the proposed tailored test design in increasing the precision of measurement of student achievement was investigated in relation to the reduction of standard error of measurement (SEM) for the estimate of student ability using the Rasch model (Rasch, 1960). The analyses showed that SEM is significantly lower for ability estimates of students in the branching test condition than that for students in the fixed test condition.

This increase in measurement precision was observed across the whole range of the student ability estimates but it was more pronounced at the low and the high end of the ability distribution as illustrated in Figure 4.

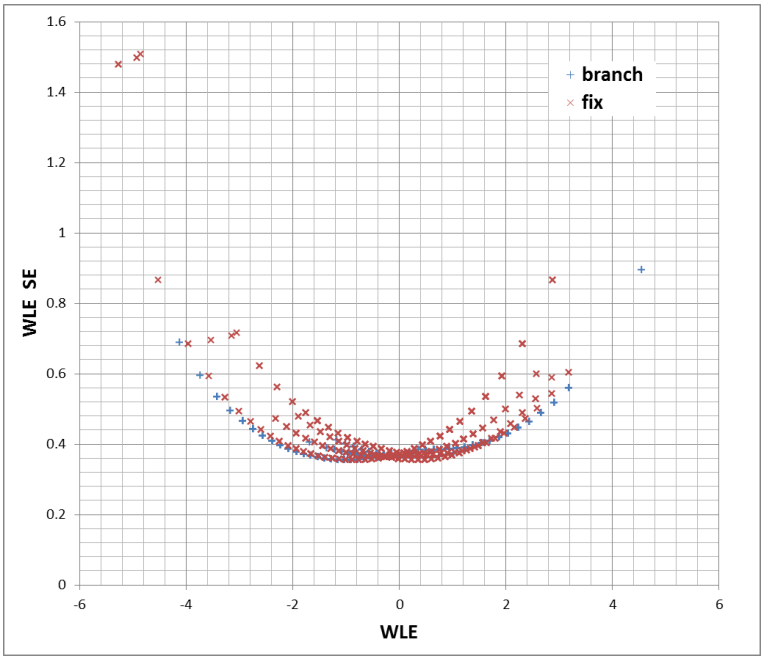

Figure 4: Distribution of estimates of student ability and corresponding standard error of measurement for Year 5 reading tests

Figure 4 shows the distribution of weighted mean likelihood estimates of student abilities (WLE) on the horizontal axis and the corresponding standard error of measurement (SE) on the vertical axis. The blue crosses show results for students in the fully branched condition and the red x-marks show results of students who were administered fixed tests. The observed increase in the measurement precision for students in the branching condition can be attributed solely to the better targeting of tests to the students' ability since the two conditions were administered to two equivalent groups of students.

Similar results are observed for all tests in the tailored test design study, which show not only that multistage tests are feasible for future NAPLAN tests but that they will bring tangible increase in measurement precision of student ability estimation.

The study also shows that implementation of the second branching point in the tailored test design provides an opportunity to correctly route students even if the initial branching might have misdirected them to a less optimal testlet. Figure 5 provides results for the branching of the Year 3 numeracy test. The majority of students were routed to testlet D after the first branching point in this test.

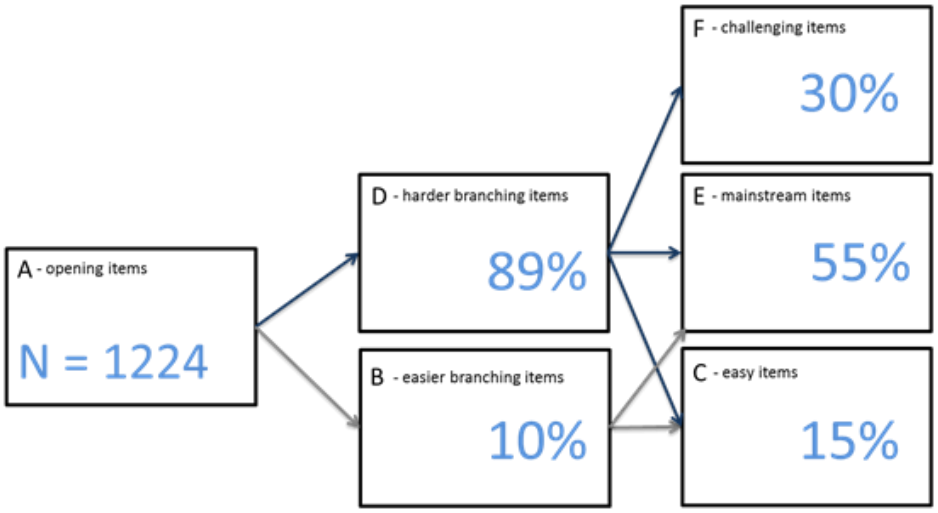

Figure 5: Outcome of branching for Year 3 TTD numeracy test

The less than optimal outcome after the first branching is the product of the unexpected change in testlet difficulty when it was presented online.. However, the distribution of students across the final three testlets is close to that observed in other years and tests, demonstrating the usefulness of the second branching point in the proposed tailored tests. It is anticipated that these less optimal branching outcomes will be eliminated in the future when information about item performance on computer-delivered tests becomes available and used in testlet construction and trialling.

Cognitive and behavioural engagement of students

Three additional follow-up studies were conducted to examine students' engagement with the multistage test and how they react to some of its key features. Two of these studies were conducted using cognitive laboratory methods, in which verbal data on students' engagement with test items and the testing event are obtained through structured observations and direct interviews. Cognitive laboratories have been widely recognised as appropriate for obtaining rich and comprehensive data (see Ericsson and Simon, 1999). Those include think-aloud methods in which students voice their mental process while engaging with the test item; and retrospective interviews in which students are asked to explain how they responded to a test item. A further advantage of these methods is that they do not require a large sample of students, with the typical sample sizes ranging from five to ten students.

One of these studies was conducted in parallel to the main tailored test design study with students drawn from 16 schools participating in it. Cognitive interviews were conducted with students across all year levels in both numeracy and reading with extensive experience in cognitive interviewing as well as an understanding of the test domains. One of the key questions investigated was how students reacted to the rising and falling pattern of item difficulty – a crucial feature of the tailored test design – given the common practice in fixed form testing where items are ordered to increase in difficulty.

This study showed that students were not distracted by the unusual progression of items in the test and that shifts in difficulty were largely ignored by students. Students did not report any adverse consequence on their engagement with items when moving through different stages of the test. Furthermore, when students were made to be aware of the branching, they regarded it as a positive feature of the test. The full method and findings of this study are available in the report prepared by Education Assessment Australia.[2]

[2] Educational Assessment Australia (EAA) November 2013, *NAPLAN Online Tailored Test Design August 2013 – Cognitive Interviews Numeracy and Reading,* UNSW Global Australia, University of New South Wales, Sydney.

The second cognitive interview study investigated whether the tailored test design could accommodate the assessment needs of students with socio-educational disadvantage. This study was conducted in collaboration with researchers from Charles Sturt University who are experts in the assessment and teaching of mathematics, and thus used only the numeracy tests. In addition to the focus on engagement with multistage testing, this study also investigated the adequacy of the test pathways containing testlets targeting the needs of educationally disadvantaged students.

The interview and observation data from this study showed that the tailored test design enabled educationally disadvantaged students to remain engaged with the full test administered to them, and most students exited the test feeling positive and with a sense of accomplishment. The data support the claim that the tailored test design is suitable for low-achieving students and has the capacity to provide better achievement proficiency estimates than a fixed design test. Further details of this study and its results are provided in the report prepared by the Research Institute for Professional Practice, Learning and Education, at Charles Sturt University.[3]

In a third study, ACARA collaborated with the Northern Territory Department of Education to collect information about the extent to which the proposed tailored test design provides a better testing experience for Indigenous students and students in remote communities. In this study, online tests were administered in eight Northern Territory schools, including two very remote schools. Students did numeracy and reading tests, both in the fixed format and in full branching format, and some students also participated in a writing study.

The reports from teachers and test invigilators involved in this study suggest that the proposed test design provided a more engaging testing situation for most students. The online tests also provided the opportunity for some students to showcase their knowledge more fully. For example, teachers reported that some students who struggled in NAPLAN writing tests found it

[3] Lowrie, T and Logan, L, November 2013, *NAPLAN Online – Trial of Tailored Test Design Numeracy Cognitive Interviews Final Report,* Research Institute for Professional Practice, Learning and Education, Charles Sturt University, Wagga Wagga.

easier to engage with the task online because typing reduced their anxiety in producing legible pieces of writing.

However, feedback was also received that some of the test targeting was outside the knowledge space of some students. Teachers reported that even some of the easier questions were still too difficult for the students, and some students reached a point where they were no longer engaged with the test. Such findings indicate that:

- further work is required to determine appropriate targeting of testlets
- to accommodate the needs of some students, testlet C may need to include content from a year level (or levels) below that of the tested year levels.

Summary of key findings

- Results of the tailored test design studies show that the delivery of multistage branching tests for NAPLAN online is sound and feasible, and that these tests offer more precise measurements of student performance, particularly for high- and low-achieving students. The results show that the current measurement model can be used to construct a NAPLAN online measurement scale.
- The psychometric analyses also show that further work is required to finalise the measurement aspects of the tailored test design; in particular, testlet boundaries require further refinement.
- The tailored test design and the proposed branching mechanism work effectively to adapt to the different ability groups. Consequently, well-targeted tests can be administered to different ability groups, thus increasing measurement precision.
- The investigation of cognitive and behavioural engagement of students with the tailored test design showed that multistage testing will provide an opportunity to all students to be assessed by tests catering more fully for their assessment and learning needs.

References

General

Adams, JR & Lazendic, G 2013, *Observations on the Feasibility of a Multistage Test Design for NAPLAN*. Unpublished technical report.

Ericsson, KA & Simon, HA 1999, *Protocol Analysis: Verbal Reports as Data,* Cambridge, MA: Massachusetts Institute of Technology.

Hendrickson, A 2007, An NCME instructional module on multistage testing, *Educational Measurement: Issues and Practice*; *26*(2): 44–52.

Kingsbury, GG, McCall, M & Hauser, C 2009, Tools for measuring academic growth, *Journal of Applied Measurement; 10*(1): 97-116.

Lazendic, G & Adams, JR, April 2014, *Multistage Test Design Incorporating Vertical Scale*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia.

Lord, FM 1971, A theoretical study of two-stage testing. *Psychometrika,* 36, 227-242.

Rasch, G 1960/1980, *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen, Danish Institute for Educational Research; Chicago: The University of Chicago Press.

Van der Linden, WJ & Glas, CAW (eds) 2010, *Elements of Adaptive Testing,* New York: Springer.

Study reports

Australian Council for Educational Research (ACER), December 2013, *Analytical Report: Psychometric Analysis for the Trial of the Tailored Test Design,* ACER, Melbourne.

Educational Assessment Australia (EAA), November 2013, *NAPLAN Online Tailored Test Design August 2013: Cognitive Interviews Numeracy and Reading,* UNSW Global Australia, University of New South Wales, Sydney.

Lowrie, T and Logan, T, November 2013, *NAPLAN Online: Trial of Tailored Test Design Numeracy Cognitive Interviews Final Report,* Research Institute for Professional Practice, Learning and Education, Charles Sturt University, Wagga Wagga.

acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

www.acara.edu.au