Problems in the Design and Administration of the 2018 NAPLAN

Les Perelman, PhD
in consultation with Walter Haney, Ed.D.

Executive Summary

- **Simultaneously administering Computer Adaptive Testing (CAT) to one part of a national population and Pen-and-Paper (P&P) tests to the other appears to be an unprecedented strategy for evaluating the transition of a large-scale national test to CAT.** There are no studies I know of that report successful use of the two testing modes on a regular single national assessment.
- **The fatal error of calibrating CAT item choices for Grammar and Punctuation on performance on the Reading portion invalidates individual student marks on the CAT for at least the Grammar and Punctuation sections and, possibly, for the entire test.** Having students begin a test section with overly difficult test items produces frustration that negatively affects student performance. However, the level of frustration experienced and the effect of that frustration on performance are two variables that will vary dramatically among individual students (Birenbaum 1986; Gershon 1992). Consequently, even if it is possible to determine some approximation of equating marks for the aggregate population (which is by no means certain), it is impossible to determine accurate marks for individual students.
- **There has been no publication of studies that are reported to equate the two test modes and to demonstrate that the marks on the two test modes are comparable** (Australian Curriculum Assessment and Reporting Authority 2017)**.** For a national test of approximately one million students to transition from one testing mode to another, there needs to be a public review of such studies and examination by independent experts.
- **The strategy employed by ACARA for cross-mode design invalidates comparisons of the 2018 NAPLAN, both the CAT and P&P Versions, with prior years' NAPLAN tests.** Items and item types on the P&P test that were not comparable to items and item types on the CAT were eliminated. Consequently, both modes of the 2018 NAPLAN omitted item types that appeared in previous years, making them different from past tests.
- **Online writing tasks are inherently incomparable with P&P Tests**
    - Typed essays receive lower scores, especially for the bottom quartile of students (Powers et al. 1992). The attention markers spend on deciphering handwriting somewhat reduces their ability to notice common errors in grammar, punctuation, spelling, and even logic.
    - Students with substantial previous word processing experience received substantially higher marks in the 2011 National Assessment of Educational Progress (USA) (Tate, Warschauer, and Abedi 2016), and other studies confirm

that computer experience correlates with performance on computer-based writing assessments (Wolfe et al. 1996; Bridgeman and Cooper 1998; Russell and Haney 1999).

- Because the marking sessions for the online test were separate from the P&P sessions, the marking procedure for online essays may have differed in personnel and in marking guidelines or training.

- **The design and execution of the 2018 NAPLAN make it so flawed that its results are of very limited use to students, teachers, parents, and schools**. Because the CAT and P&P tests are not comparable, posting results in the *My School* database would be inappropriate and misleading. In addition, comparison of 2018 results with those of prior years is, for the most part, a futile exercise. The CAT and P&P Tests are of unknown comparability, and because the composition of the P&P test items has been modified, any comparison of the test with prior years' results has been compromised. The only part of the 2018 NAPLAN suitable for longitudinal comparison is the P&P writing task, which, itself, has many defects (Perelman 2018). In sum, the 2018 NAPLAN results should be discarded.

Problems in the Design and Administration of the 2018 NAPLAN

Les Perelman, Ph.D.
in consultation with Walter Haney, Ed.D.

20 August 2018

Introduction

This report, an analysis of possible problems in the 2018 in equating student performance on the new online NAPLAN with student performance on the traditional pen-and-paper test, has been commissioned by the New South Wales Teacher Federation. This year approximately 200,000 students sat the online NAPLAN while the remainder of the approximately one million students completed the paper-and-pen test. However, it has been reported that the Reading and Grammar and Punctuation sections of the online NAPLAN were treated as a single section by the adaptive testing algorithm, causing students who performed well on the Reading section to encounter difficult questions when starting the Grammar and Punctuation sections (Baker and Cook 2018). Moreover, there are the additional issues of whether marks on sections of the adaptive test are equivalent and comparable to marks on the traditional paper-and-pen test as well as issues of possible inequities in the online and paper-and-pen administrations of the writing task.

Although my charge is to identify issues in the introduction of the online NAPLAN, it should be made clear that rather than being an opponent of employing online tools in the teaching and assessment of writing, I have long been a practitioner and innovator of these practices. In the 1980's, I taught writing at the Massachusetts Institute of Technology in one of the world's first networked online writing classroom environments (Barrett and Paradis 1988). From 2001-2006, I was Principal Investigator on a grant from Microsoft to develop jointly with other universities and colleges in the United States an application for online writing assessment (Perelman 2006). The application and its adaptations are currently used in a number of universities and colleges in the United States (Peckham 2009).

Computer Adaptive Testing (CAT)

CAT works by adapting the difficulty of questions presented based on previous student performance. Let us hypothesize a test in which question items are ranked in difficulty on a scale of 1 to 100, with 50 being level of difficulty attained by the mean, that is, an exactly

average student. Usually, a test begins with a fairly easy question item to build student confidence and then increases the level of difficulty until a student is unable to correctly answer a question. As a simplified illustration, Student A is given two questions at level 30 on a 1-100 scale, she answers them correctly. She is then given two questions at level 40, which she answers correctly, which she also does for question items at levels 50, 60, 70, and 80. However, she fails to answer the two question items at level 90 correctly. The testing algorithm then offers her a series of question items in the range of 81-89, finally identifying that she consistently answers question correctly up to level 86 but misses questions beginning at level 87. Consequently, she attains a mark of 86. Student B, on the other hand, incorrectly answers the first two questions at level 30, but correctly answers two question items at level 20. A process similar to one experienced by Student A, identifies Student B achieving a mark of 24.

Advantages of CAT

Computer adaptive testing offers several advantages over pen-and-paper (P&P) testing. First most large-scale tests are designed to differentiate performance of the majority of the population but are extremely ineffective at differentiating at the two ends of the scales, which is why the scales of many common tests represent three standard deviations below the mean to three standard deviations above. For unusual populations clustering at one of the tails of the bell curve, mass standard tests are often ineffective in differentiation. CAT, by having the potential to offer easier and more difficult test items, has the potential to widen the scale to differentiate more precisely the abilities of those on the top and the bottom of the traditional scale. By focusing more test items around the limits of participants ability, CAT can produce a more accurate and precise score. Finally, by not having to include test items at all ability levels, CAT has the potential to significantly shorten testing time.

Disadvantages of CAT

CAT, however, has several disadvantages. First, CAT tests are black boxes. Although the various general statistical techniques are well known, some of the specific algorithms determining the sequence and difficulty of items are proprietary. Furthermore, because of the continuous reuse of test items, they are never released, and students lose the opportunity of reviewing the test afterwards and, potentially, learning from their mistakes. Moreover, because of the potential opportunity of changing answers to game the system, students are usually prohibited from returning to previous questions. Finally, some questions used on P&P tests, such as open-ended questions, are not appropriate for CAT, making equating the two types of tests difficult, if not impossible (Linacre 2000).

Equating CAT and P&P Test Items and Ensuring Comparability

The various methods for sampling and equating test items at specific levels of difficulty are complex and widely discussed processes (von Davier 2013; Livingston, Dorans, and Wright 1990; Kolen 1990; Schmitt et al. 1990; Dorans 1990; Rudner 2007; Wyse and Hao 2012; Dorans, Moses, and Eignor 2010; Ware, Bjorner, and Kosinski 2000; Green et al. 1984; Weiss and Gage Kingsbury 1984; Angoff 1984). One of the greatest problems in transitioning from P&P tests to CAT is equating the two tests to make marks in the two types of tests comparable. One of the

principal advantages of CAT is to extend the scale and its accuracy at both ends, differentiating both high scoring and low scoring students. How the wider scale is mapped onto the narrower scale is not well explained. ACARA needs to publicly report in detail its previous comparability studies and discuss its specific methods of sampling and equating test items.

Transitioning from P&P Test to CAT.

The procedure for transitioning from P&P testing to CAT is well documented and well mapped. Two relevant examples are the Graduate Management Admissions Test (GMAT) and the Smarter Balanced Assessment Consortium (SBAC), both from the USA. In the year before the first administration of the GMAT CAT version, two large scale-studies were conducted to ensure the comparability of CAT versions (Rudner 2007). Over 4,000 students were recruited to take the GMAT twice, once on a CAT and once on a traditional P&P test. The inducement was that the tests would be free and that only the highest of the two marks would be reported. The first study reported that the two modes were not comparable. Although performance at the middle of the scale was close, there were large differences in performance in the two modes at the high and low ends of the scale. A second study was conducted, but with a low participation rate. The results of the two trials were merged, and reviewed by an outside consultant who critiqued the design. Beginning the next year, all students took the CAT version of the test. Although the new scale was similar to the old P&P scale, it was not equivalent. Moreover, there was a significant increase in the mean scores in the quantitative section. Finally, the CAT version failed to provide better differentiation among high-end scores.

In the Spring of 2014, SBAC conducted a large-scale Field Test involving more than 4.2 million students in 16,549 participating schools in 13 states (Doorey 2014). The Field Test had many purposes including logistical and pedagogical. It was intended to prove that the implementation of online testing would work along with an assessment of the new test items in terms of rigor and alignment to school curricula. There was little attempt to make the new assessment comparable to previous SBAC assessments. Although some specific issues, mainly technical and administrative, were identified, the Field Test was a major success, allowing the launch of the new assessment the following year.

These two examples represent two different approaches to transitioning to CAT. Because the GMAT is an admissions test, the compelling issue of fairness necessitates that comparability was and is an important issue. SBAC, like NAPLAN, measures student achievement. To produce an assessment that closely aligned with the common curriculum and that made full use of the new online technologies, SBAC made the design trade-off of sacrificing comparability of marks with prior years' results.

Administering CAT and P&P Tests Simultaneously to Segments of the Same Population

The 2018 NAPLAN had two forms: 1) the traditional P&P test including the writing task administered to approximately 80% of students; and 2) the CAT Test including an online writing task to the remaining 20%. Essential and inescapable differences between P&P and online

writing will be discussed below. The focus here is on the other parts of the test. I know of no other instance when such a large-scale national test was offered in these two modes.

Moreover, ACARA has not completely explained how the two tests were equated. There are reports that ACARA removed items from the P&P test that were not comparable to items on the CAT and that the CAT contained an overrepresented sample of items from the P&P test. Both of these strategies are highly problematic. By removing categories of the test items from the P&P test that had been on previous NAPLANs, ACARA changed the composition of the test, making longitudinal comparisons with previous years extremely difficult if not impossible. Moreover, populating the CAT with P&P test items contracts the range of difficulty on the test, diminishing CAT's crucial feature of being able to quickly and precisely differentiate student performance.

The Fatal Design Error in the 2018 NAPLAN Test

By far the most serious problem in the 2018 NAPLAN was CAT's treating the Reading section and the Grammar & Punctuation section as a single cognitive domain rather than two separate domains (Baker and Cook 2018). The result was that instead of resetting difficulty levels for the Grammar and Punctuation section, the CAT algorithm served students test items with a level-of-difficulty determined by their performance on the Reading section.

The best way to illustrate the damage caused by this error is to return to the narrative of Student A. Student A was calibrated by the Reading Section CAT as receiving an 86 mark. Because the system did not reset for the new section, the CAT algorithm offers her two question items on Grammar & Punctuation slightly below her reading mark, for illustration, at level 80. Because Student A is not as proficient in these topics as she was in reading, she does not know the correct answer to either question. The algorithm then presents Student A two questions at level 70, which again are too difficult for her, as are questions at level 60. Student A becomes frustrated, and her frustration causes her to incorrectly answer questions that she would normally answer correctly. The CAT algorithm descends down questions between 30 and 40 marks before Student A answers correctly. She receives a final mark of 34, even though she would have received a much higher mark had she taken the P&P test.

Having students begin a test section with overly difficult test items produces frustration that negatively affects student performance (Birenbaum 1986; Gershon 1992; Ling et al. 2017). It would be extremely difficult, if not impossible, to devise a method to correct for this effect for an aggregate population such as an Australian state or territory. It is unmanageable that a technique could be developed to correct for individual student marks. The level of frustration experienced by each student will vary dramatically as will the effect of that frustration.

If ACARA attempts to rectify this error through statistical techniques, such efforts should be completely transparent and subject to external review. Moreover, such results should be reported with the estimated probability of the actual (true) score occurring within a specified

error range. Finally, as noted above, such estimations are not feasible for individual student marks on Grammar and Punctuation.

Comparability of Online and P&P Writing Tasks

Although it may seem counter-intuitive, students typing their writing tasks online are, with one exception, at a disadvantage compared to students writing with pen-and-paper. Typed essays receive lower scores, especially for the bottom quartile of students (Powers et al. 1992). As someone who has had thirty-eight years' experience conducting sessions evaluating student essays, the cause is obvious. The attention readers spend on deciphering handwriting somewhat reduces their ability to notice common errors in grammar, punctuation, spelling, and even logic.

There is one notable exception. Students with substantial previous word processing experience perform very well. A recent study of over 24,100 Year 8 students taking the 2011 National Assessment of Educational Progress (USA) online writing task revealed that substantial experience writing on computers, especially in classroom settings, corresponded to high performance on the writing test (Tate, Warschauer, and Abedi 2016). Other studies confirm that computer experience correlates with performance on computer-based writing assessments (Wolfe et al. 1996; Bridgeman and Cooper 1998; Russell and Haney 1999). Since word processing experience may correlate with socio-economic status, there is the possibility that online writing tests, especially timed online writing tests, further disadvantage already disadvantaged groups.

Finally, because the marking sessions for the online test were separate from the sessions marking the P&P writing tasks, the marking procedure for online writing task may have differed in personnel and in marking guidelines or training. Given that typescript makes surface errors more apparent, the monitoring for the evaluation of online writing could have diverged substantially making comparability impossible.

Conclusion

The design and execution of the 2018 NAPLAN make it so flawed that its results are **of** very limited use **t**o students, teachers, parents, and schools. Because the CAT and P&P tests are not comparable, posting results in the *My School* database would be inappropriate and misleading. In addition, comparison of 2018 results with those of prior years is, for the most part, a futile exercise. The CAT and P&P Tests are of unknown comparability, and because the composition of the P&P test items has been modified, any comparison of the test with prior years' results has been compromised. The only part of the 2018 NAPLAN suitable for longitudinal comparison is the P&P writing task, which, itself, has many defects (Perelman 2018). In sum, the 2018 NAPLAN results should be discarded.

Moreover, ACARA must take responsibility for what can only be described as educational and statistical incompetence. Some key questions need to be answered:

- What exactly was the error? Did someone conceptualize Reading and Grammar & Punctuation as a single cognitive domain or was it an error of implementation or programming?
- Who was responsible for the error? ACARA? If so, exactly who in the organization? If it is an outside vendor, the vendor needs to be identified.
- ACARA needs to provide a clear narrative of how the machine algorithm treated students who scored high in reading when they sat the Grammar & Punctuation section.
- There is a claim that prior research by the ACARA / National Assessment and Surveys Online Program (NASOP) Research Team indicated that tests and test items functioned the same or similar in CAT and P&P Modes. Could that report be made public? In addition, this team also authored the report on Automated Essay Scoring (ACARA NASOP Research Team 2015). Who were the members of this team?
- What exactly is the algorithm for correcting student marks on the Grammar & Punctuation Section? What is the estimated standard error? Approximately what percentage of students will still be incorrectly marked?
- What was the original equating design? How has the equating design been modified to account for the error?

Finally, the failure of the 2018 NAPLAN provides an opportunity for the development of a new NAPLAN, jointly developed by all stakeholders, government, teachers, parents, and students, and much more closely aligned with the national curriculum.

Works Cited

ACARA NASOP Research Team. 2015. “An Evaluation of Automated Scoring of NAPLAN Persuasive Writing.” http://nap.edu.au/_resources/20151130_ACARA_research_paper_on_online_automated_scoring.pdf.

Angoff, William H. 1984. *Scales, Norms, and Equivalent Scores*. Princeton NJ: Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/book/1984/ibmt.

Australian Curriculum Assessment and Reporting Authority. 2017. “FAQ Comparability of Paper and Online NAPLAN Tests.” https://www.nap.edu.au/docs/default-source/default-document-library/faq_comparability-of-paper-and-online-naplan-tests.pdf?sfvrsn=2.

Baker, Jordan, and Henrietta Cook. 2018. “NAPLAN: Grammar and Punctuation at Centre of Australia’s Education Controversy.” *Sydney Morning Herald*, August 10, 2018. https://www.smh.com.au/education/grammar-and-punctuation-at-centre-of-naplan-controversy-20180809-p4zwj5.html.

Barrett, Edward, and James Paradis. 1988. “Teaching Writing in an On-Line Classroom.” *Harvard Educational Review* 58 (2): 154–72. https://doi.org/10.17763/haer.58.2.4200232732613124.

Birenbaum, Menucha. 1986. “Effect of Dissimulation Motivation and Anxiety on Response Pattern Appropriateness Measures.” *Applied Psychological Measurement* 10 (2): 167–74. https://doi.org/10.1177/014662168601000208.

Bridgeman, B, and P Cooper. 1998. “Comparability of Scores on Word-Processed and Handwritten Essays on the Graduate Management Admissions Test.” 421528. https://eric.ed.gov/?id=ED421528.

Davier, Alina A. von. 2013. “Observed-Score Equating: An Overview.” *Psychometrika*. https://doi.org/10.1007/s11336-013-9319-3.

Doorey, Nancey. 2014. “Smarter Balanced ‘Tests of the Test’ Successful: Field Test Provides Clear Path Forward.” Los Angeles. https://portal.smarterbalanced.org/library/en/2014-field-test-report.pdf.

Dorans, Neil J. 1990. “Equating Methods and Sampling Designs.” *Applied Measurement in Education* 3 (1): 3–15.

Dorans, Neil J, Tim P Moses, and Daniel R Eignor. 2010. “Principles and Practices of Test Score Equating.” Princeton NJ. http://www.ets.org/research/contact.html.

Gershon, R. C. 1992. “Test Anxiety and Item Order: New Concerns for Item Response Theory.” In *Test Anxiety and Item Order: New Concerns for Item Response Theory. in Measurement: Theory into Practice. Vol. 1*, edited by M. Wilson. Norwood NJ: Abelex.

Green, Bert F, R Darrell Bock, Lloyd G Humphreys, and Robert L Linn. 1984. “Technical Guidelines for Assessing Computerized Adaptive Tests.” *JOURNAL OF EDUCATIONAL MEASUREMENT*. Vol. 21. https://about.jstor.org/terms.

Kolen, Michael J. 1990. “Does Matching in Equating Work: A Discussion.” *Applied Measurement*

in Education 3 (1): 97–104.
Linacre, John Michael. 2000. “ComputerAdaptive Testing: A Methodology Whose Time Has Come.” Chicago.
Ling, Guangming, Yigal Attali, Bridgid Finn, and Elizabeth A. Stone. 2017. “Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test?” *Applied Psychological Measurement* 41 (7): 495–511. https://doi.org/10.1177/0146621617707556.
Livingston, Samuel A., Neil J. Dorans, and Nancy K. Wright. 1990. “What Combination of Sampling and Equating Methods Works Best?:” *Applied Measurement in Education* 3 (1): 73–95.
Peckham, I. 2009. “Online Placement in First-Year Writing.” *College Composition and Communication*. http://www.jstor.org/stable/20457080.
Perelman, Les. 2006. “Assessment in Cyberspace.” 2006. http://www.mhhe.com/socscience/english/tc/perelman/perelman_module.html.
———. 2018. “Towards a New NAPLAN: Testing to the Teaching.” Surry Hills NSW Australia.
Powers, D. E., M. E. Fowles, M. Farnum, and P. Ramsey. 1992. “Will They Think Less of My Handwritten Essay If Others Word Process Theirs? Effects on Essay Scores of Intermingling Handwritten and Word-Processed Essays.” *Journal of Educational Measurement* 31: 220–33.
Rudner, Lawrence. 2007. “Implementing the Graduate Management Admission Test Computerized Adaptive Test.” In *Proceedings 0f the 2007 GMAC Conference on Computerized Adaptive Testing*, edited by D. J. Weiss. McLean, VA: Graduate Management Admission Council. https://doi.org/10.1007/978-0-387-85461-8_8.
Russell, Michael, and Walt Haney. 1999. “Testing On Computers.” *Education Policy Analysis Archives* 7 (0): 20. https://doi.org/10.14507/epaa.v7n20.1999.
Schmitt, Alicia P., Linda L. Cook, Neil J. Dorans, and Daniel R. Eignor. 1990. “Sensitivity of Equating Results to Different Sampling Strategies.” *Applied Measurement in Education* 3 (1): 53–71.
Tate, Tamara P., Mark Warschauer, and Jamal Abedi. 2016. “The Effects of Prior Computer Use on Computer-Based Writing: The 2011 NAEP Writing Assessment.” *Computers & Education* 101 (October): 115–31. https://doi.org/10.1016/J.COMPEDU.2016.06.001.
Ware, John E, Jakob B Bjorner, and Mark Kosinski. 2000. “Practical Implications of Item Response Theory and Computerized Adaptive Testing A Brief Summary of Ongoing Studies of Widely Used Headache Impact Scales.” *Medical Care* 38 (9): Supplement II 73-82. https://pdfs.semanticscholar.org/2e43/6893c89f845477f9b477bd52d3ed5737ec92.pdf.
Weiss, David J, and G Gage Kingsbury. 1984. “Application of Computerized Adaptive Testing to Educational Problems.” *Source: Journal of Educational Measurement*. Vol. 21. Winter. https://www-jstor-org.libproxy.mit.edu/stable/pdf/1434587.pdf?refreqid=excelsior%3A2e108b44ad6c0971611773dc74ce5236.
Wolfe, Edward W., Sandra Bolton, Brian Feltovich, and Donna M. Niday. 1996. “The Influence of Student Experience with Word Processors on the Quality of Essays Written for a Direct Writing Assessment.” *Assessing Writing* 3 (2): 123–47. https://doi.org/10.1016/S1075-2935(96)90010-0.
Wyse, Adam E., and Shiqi Hao. 2012. “An Evaluation of Item Response Theory Classification

Accuracy and Consistency Indices." *Applied Psychological Measurement* 36 (7): 602–24. https://doi.org/10.1177/0146621612451522.