NAPLAN TECHNICAL SUMMARY

I. INTRODUCTION

The following is a high level and technical summary of the key processes and stages associated with the development, administration, scoring, analysis and reporting of the NAPLAN tests. It is intended to provide a more detailed specification of the NAPLAN processes than is included in the main submission, but is not itself a technical manual. The information included in this summary has drawn upon information from the following documents:

1. *2008 and 2009 Draft NAPLAN Central Analysis Technical Reports*
2. *Audit of the NAPLAN Item and Test Development Processes 2009 and 2009 (Report)*
3. *NAPLAN Writing 2009 (Report)*
4. *2010 National Assessment Program – Literacy and Numeracy (NAPLAN) – Development of Testing and Trial Invitation to Offer EDUC-100453*
5. *2010 National Assessment Program – Literacy and Numeracy (NAPLAN) – central Data Analysis and Reporting Invitation to Offer EDUC-100534*
6. *(Draft) OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes – Australian Country Background Report March 2010*

II. NAPLAN QUALITY, VALIDITY AND RELIABILITY

An examination of the specifications, processes, quality controls and audits at each stage of the NAPLAN program confirms the quality, validity and reliability of the tests and the test results data.
There are five main stages of NAPLAN testing including:

i. test development;
ii. test administration;
iii. marking and data capture;
iv. data analysis and
v. reporting.

Each stage is extensively quality assured through highly specified and rigorously applied processes, continuous quality control and auditing. Wherever appropriate, these processes draw on expert advice and are informed by national and international best practice, such as processes associated with the Programme for International Student Assessment (PISA).

A summary of each of these stages is provided below, together with relevant information on quality management and control.

III. NAPLAN TEST DEVELOPMENT

The test development process for NAPLAN is comprehensive, rigorous and draws on the best available expertise within Australia and national and international best practice.

The development model includes central management of the project by ACARA working with expert organisations providing services under contract and supported by expert review and recommendations from officials from all states and territories.

The development cycle for the NAPLAN tests is approximately 12 months and proceeds in phases.

Phase One - Test Design

NAPLAN has four tests that are administered at each of Years 3, 5, 7 and 9. The four tests are Reading, Writing, Language Conventions (Spelling, and Grammar and Punctuation) and Numeracy. The test results are reported in five domains: Reading, Writing, Spelling, Grammar and Punctuation, and Numeracy.

In Reading, Language Conventions and Numeracy, there is a mix of multiple-choice items and short-answer items. The multiple-choice items (MC) are of standard multiple-choice format with a number of possible answers (usually four) from which students are required to select the best answer. The short-answer items are closed-constructed response items (CR) that generally require a numeric answer, a word or a short phrase. All MC and CR items are dichotomously scored (correct or incorrect).

Each of the Year 7 and Year 9 Numeracy tests consist of two test forms. The use of calculators is permitted in the first form (calculator form) and is not permitted in the second form (non-calculator form). Each form typically consists of 32 items. The Year 7 and Year 9 numeracy tests each consist of 64 items with the two forms combined. Year 3 and Year 5 Numeracy tests are non-calculator forms.

For Writing, all students in Years 3, 5, 7 and 9 are required to write a response to the same writing prompt or writing task. There is a range of genres from which to select the task, including narrative and persuasive writing. The scripts are rated (scored) based on the same ten criteria (Criteria 1 to 10) across all four year levels.

Detailed and nationally endorsed specifications govern the development of items and the final tests forms.

Phase Two – Item Development and Test Construction

This phase involves the engagement of specialist writers to develop test questions (items) that meet the nationally agreed specifications. These include specifications for the types of test items that can be used, test item content (curriculum content), length of tests and the spread of difficulty across items. The distribution of item difficulty in each of the learning areas should be 20%, 30%, 30% 20% across the four equal logit quarters scale of students' proficiency. Item development contractors must adhere to these specifications and also comply with very detailed formatting and layout requirements as specified in the NAPLAN Style Guide.

NAPLAN is developed to ensure curriculum coverage appropriate to the relevant year levels. Currently this is achieved by referencing the *National Statements of Learning* in English and Mathematics, and state and territory curriculum and learning frameworks. In future years,

NAPLAN will be specified against an Assessment Framework based on the new national curriculum once that curriculum is implemented.

Development of innovative assessment items that permit students to demonstrate the variety of skills and understandings described in the *National Statements of Learning* is required.

Items must show a relevance to classroom teaching practice and school assessment of the outcomes, as well as having the capacity to fulfil diagnostic and measurement requirements.

Consideration is given to possible dependencies between items to address data distribution issues that may impact on reporting.

The developers are required to interact with students when developing items to improve the quality of the items and the functioning of the distractors proposed for trialling. Multiple-choice distractors must provide diagnostic information on students' misconceptions.

Developers are further expected to utilise cognitive laboratory interviews where a small number of carefully selected students are interviewed while they are taking the trial test or shortly after they have completed a set of items. The interviews will provide an in-depth understanding of how the items are working and are especially useful when information is needed about particular item formats and the functioning of distracters. While not all items need to be workshopped in this way, the development contractor will be expected to provide a rationale for the items selected for student interviews and indicate the anticipated proportion chosen from the total.

For all assessment items that are developed, the following information must be provided:

i. links to the *Statements of Learning*;

ii. the level of difficulty of each in terms of quartiles of the anticipated student distribution and NAPLAN bands; and

iii. the marking keys (correct responses).

Full marking keys and guides are developed and, where appropriate, exemplified.

All assessments should give valid and reliable measurements of student ability at the individual student, state and national levels.

Phase Three – Item Review

Panels of experts review proposed items from a range of perspectives. Initially the test development contractor will undertake in-house panelling to determine the suitability and quality of items. The items are then presented to panels from all state and territory test authorities for review by experts in curriculum and measurement, practising teachers and specialist officers in areas such as Indigenous education, English as a Second Language, and students with special needs.

Item and distractor descriptors must be provided for all items submitted for the panels to be able to assess the appropriateness of the item. The distractor descriptors indicate the misconception or errors leading to students selecting distractors for all multiple choice items. The test development contractor is required to conduct research into students thought processes and reasoning to assist in the development of accurate descriptors to be considered for inclusion in final forms.

Only those items that meet the stringent criteria of the panels proceed to trialling.

Phase Four – Item Trialling, Data Analysis and Final Selection

Once the test questions are agreed, they are constructed into 'trial test forms' that are then sat by a scientifically chosen sample of students within Australia, to obtain critical item performance data. The performance of each question, including for example how well it is able to discriminate high-performing and low-performing students, or whether there is any bias, is determined by psychometric analysis of the data, conducted after the trial.

The development contractor designs the trialling regime using expert advice and experience. The design must take into account the curriculum imperatives of each learning area, psychometric requirements and equating needs (see below).

Trials are fully invigilated to ensure the secure return of all test materials.

The trial tests are returned after the trial testing period and marking and data capture of student responses is undertaken. There are detailed quality assurance processes in place to ensure the accuracy and consistency of these data. The data are collected, managed and stored in accordance with strict protocols in relation to cleaning, auditing, transmission, storage, access and usage of data.

The final selection of items for inclusion in tests is based on a set of quality assurances including: (i) the psychometric data collected through trialling (ii) professional judgments from educational measurement, test construction and curriculum experts from all jurisdictions and (iii) the requirement to have the final test forms comply with the detailed NAPLAN test specifications. Relevant data to inform this decision making includes the following.

i. Summary statistics (overall fit, person separation, etc.)
ii. Category frequencies/proportions
iii. Item difficulties
iv. Item fit statistics
v. Item Characteristic Curves (ICCs)
vi. Differential Item functioning information (DIF)
vii. Multiple Choice Distractor Analysis
viii. Category Probability Curves and Threshold Probability Curves for polytomous items
ix. Statistics attesting to the quality and performance of link items
x. Person-item maps

Phase Five – Construction of Final Test Forms and the NAPLAN Common Scale

After analysis of the trial data, the project manager and review panels (comprising nominated jurisdictional staff with expertise in data interpretation and assessment task evaluation) determine items that will be flagged for selection for the final test forms. Prior to meeting with the panels the development contractor will make available to jurisdictional nominees data from the trial and the proposed final forms.

Items in each domain for all year groups will be calibrated on the same scale through the use of link items embedded in the trial tests.

NAPLAN results are able to be reported using five scales, one for each of the following domains: Reading, Writing, Spelling, Grammar and Punctuation and Numeracy. Each of the

NAPLAN reporting scales describe the development of student achievement in a domain from Year 3 through to Year 9 along a ten-band scale. Students in year 3 are reported against bands 1 - 6, in year 5 bands 3 – 8, in year 7 bands 4 – 9 and in year 9 bands 5-10. At each year level the second achievement or 'proficiency' band is designated as the national minimum standard for that year level.

The use of a common scale provides significant information about the performance of, and growth in, individual student achievement which can be monitored over time and add a longitudinal dimension to the data. Through the use of these common scales, it is possible to gauge the achievement of the most able group of students and, at the same time, to pay attention to the group of students who have yet to reach the agreed national minimum standard.

An Expert Advisory Group (EAG) consisting of five pre-eminent educational measurement experts provides advice on all relevant aspects relating to technical methodology and specification, equating of tests and quality assurance. The final test specifications are reviewed by the EAG to ensure there is an acceptable level of compliance between the target NAPLAN test specifications and the achieved specifications in the tests.

IV. TEST ADMINISTRATION

State and territory Test Administration Authorities (TAAs) are responsible for the implementation and administration of the NAPLAN tests in their jurisdictions. These authorities manage the printing and distribution of test materials, coordinate the testing program within their jurisdictions and administer special provisions to assist eligible students with particular needs to participate in testing

The NAPLAN tests are conducted at schools and administered by classroom teachers, school deputies or the principal.

The test administration authorities for NAPLAN are:

- ACT - Department of Education and Training
- NSW - Department of Education and Training
- NT - Department of Education and Training
- QLD - Queensland Studies Authority
- SA - Department of Education and Children's Services
- TAS - Department of Education
- VIC - Victorian Curriculum and Assessment Authority
- WA - Department of Education

ACARA has nationally agreed protocols for the administration of NAPLAN testing that are used by all test administration authorities. The *National Protocols for Test Administration* forms the basis for the principals' handbook and test administration manuals to ensure the integrity and consistency of the testing process. The National Protocols also include detailed requirements for the management of test security.

The security of the NAPLAN tests is also strengthened through contractual obligations on commercial service providers together with strict instructions for the handling of materials in schools. Contractors responsible for the printing, packing and delivery of NAPLAN test

materials must comply with strict quality assurance and security requirements. For example, there are specific requirements for the secure packaging of materials, highly restricted access for staff to areas where test materials are produced and secured and agreed protocols for the delivery of materials to schools.

Once students have sat the tests, those tests are collected and the TAA in each state and territory manages the marking of the tests and the capture of answer data through an electronic scanning process. Tests for Reading, Language Conventions (Spelling, Grammar and Punctuation) and Numeracy are scored using optical mark recognition software for multiple-choice items.

V. MARKING WRITING

Writing tasks are professionally marked using quality assured procedures for maintaining marker accuracy and consistency. There are agreed marking standards and quality assurance processes for achieving consistency and reliability within and between marking centres. TAAs are responsible for the marking of student scripts, including adherence to agreed marking standards and processes. All NAPLAN writing tasks are marked online, onscreen by markers in approximately 15 marking centres nationally.

There is consistent training for all marking staff and common training materials. The Chief Assessor for marking nationally is responsible for training Centre Leaders who then train all markers in their respective marking centres. There is common terminology used to name various types of writing scripts used for training, monitoring and support of markers.

Control scripts (common pre-marked scripts) are used to monitor marker accuracy. The national requirement for the use of control scripts is to have control scripts delivered electronically every day that marking is conducted and completed by every person involved in marking student scripts. The national Marking Quality Team has refined the common guidelines for national consistency in marking centres.

All marking centres also follow agreed protocols for monitoring and remediating markers, with the majority implementing additional measures such as: exceeding the agreed parameters for re-marking scripts where there was evidence of discrepant marking by a marker, group of markers or whole centre; check marking of all markers above the recommended 10% rate; and re-training and supporting markers, particularly those who are less experienced.

The provision daily of de-identified marker data by TAAs enables the Chief Assessor to produce jurisdiction and national summaries of control script data by total score by day, control script data by criterion by day and consolidated control script data over time. An analysis and commentary on the day's marking performance is provided as feedback to marking centres on a 24-hour turnaround basis.

The analysis of both the 2008 and 2009 marking of writing confirmed markers were marking consistently and within acceptable levels of variation. Anomalous marking will invariably be detected through the quality assurance processes.

The marking 'rubric' (marking scheme) is the basis for marker judgement and decision making. It comprises ten criteria. The analysis of the performance of the marking rubric in both 2008 and 2009 showed that results for each year level, based on the national calibration sample, indicated that the rubric was applied consistently across year levels. Analysis of the results by jurisdiction, also based on the national calibration sample,

concluded that the rubric was applied consistently across the jurisdictions, in that students with similar ability estimates achieved a similar score on each criterion.

VI. DATA ANALYSIS

TAAs submit de-identified student data from all tests to a contractor appointed to undertake the analysis of the test data on behalf of ACARA and the states and territories. The national contractor performs a range of analyses across the data for purposes of individual, school, jurisdiction and national reporting.

A summary of the key features of the central analysis of NAPLAN data follows.

National Calibration (or Scientific) Sample

Over 1 million students participate in the NAPLAN testing program, producing over 4 million tests for scoring. In order to be able to commence the analysis of the data as soon as possible, a sample population of students from within the total test population is drawn and used for the initial stages of analysis. The 'calibration' sample is scientifically designed to ensure it is representative of the total population ('the school sampling frame'). The calibration sample size is approximately 75,000 students.

This sample is used to estimate item parameters, perform common-item vertical equating, evaluate the psychometric characteristics of the tests, and establish score-equivalence tables after equating.

The school sampling frame used for 2008, 2009 and 2010 was the Australian Council for Educational Research (ACER) Sampling Frame, a comprehensive list of all schools in Australia, developed by ACER by coordinating information from multiple sources, including the Australian Bureau of Statistics and the Commonwealth, State and Territory education departments.

The sample design developed for the project is a stratified cluster sample. Prior to sampling, the schools are explicitly stratified by jurisdiction and sector. That is, the sampling frame is divided into twenty four separate parts representing each jurisdiction by sector combination. Within each of these strata, the frame is sorted by geographic location (the MCEECDYA geolocation code), a school-postcode-based measure of socio-economic status (the ABS SEIFA index of Education and Occupation), and school size.

Data Analysis Approaches and Methodology

The findings of the central analysis of NAPLAN data inform student, school, jurisdiction and national reporting. This data is also central to the information provided about school performance on the *My School* website. The accuracy and validity of the analysis of data is therefore critical. This section of the submission summarises key features of the processes and methodologies used in the central analysis of NAPLAN data.

The psychometrics and scaling methods used are proven methods that have been widely utilised in other large scale assessment programs and in survey research.

Data collection is undertaken by the TAAs in jurisdictions and there are a total of three deliveries of data. With each round of data delivery, the datasets are cleaned and recoded in preparation for analysis. Each TAA is required to prepare its jurisdiction data according to the common codebook provided. All data files are checked for invalid codes and inconsistencies. Data is cleaned and recoded by the contractor and any concerns about data are communicated to the TAA and rectified as necessary. Recoded data files are generated and verified in preparation for data analysis.

The Reading, Spelling, Grammar and Punctuation, and Numeracy tests are calibrated separately by domain and year level, resulting in 16 separate calibrations.

Test calibration and scaling is performed based on the Rasch model.

For data cleaning and data analysis for reporting, the statistical software packages SPSS and SAS are used. For the Rasch scaling analysis, the software ACER ConQuest (Wu, Adams and Wilson, 1997) is used. This software provides tools for the estimation of a variety of different item response models and regression models.

Analyses are undertaken in the following order:

- Item and test analyses based on cleaned and recoded calibration sample data (treating 'not reached' items as 'not administered' to obtain appropriate estimates of item difficulty). Senate weights are used for case weights
- Checking of item and test characteristics, distractor analysis, and DIF analysis
- Vertical equating based on common items in tests of adjacent year levels (Year 3 and Year 5, Year 5 and Year 7, Year 7 and Year 9)
- Horizontal equating using the off-shore and on-shore equating data
- Combining results of the horizontal and vertical equating to construct the NAPLAN domain scales
- Generation of student weighted likelihood estimates (WLE) to obtain score equivalence tables
- Generation of plausible values for the calibration sample. Student weights were used in the calculations of preliminary statistics.
- Transformation of logit scores into NAPLAN scale scores
- Analysis to obtain preliminary results based on the sample data
- Calculation of equating errors

The statistical information regarding item characteristics provided includes:

- Item facility, expressed as percentage correct for each relevant year level
- The item location on individual year level scales
- Test targeting and item spread
- Information about the fit of the item to the Rasch model specified
- Plots for assessing potential differential item functioning (DIF)
- Information on the consistency of the functioning of link items between adjacent year levels

Rasch estimates are used for reporting proficiencies in the five domains. The weighted likelihood estimates (WLEs) are used for reporting to individual students and to schools.

The plausible value methodology is used with the background variables being gender, LBOTE status, ATSI status, school geolocation, and school Reading WLE average score. In addition, parental education and occupation have been added to the set of conditioning variables as the quality of data and response rates of these variables has improved significantly.

The five sets of plausible values are used to calculate means, standard deviations, percentiles and percentages of students within proficiency bands, for each domain and each year level at the jurisdictional and national level.

Estimates of sampling and measurement errors are combined to obtain final standard errors for the performance statistics reported for the census data. The standard errors are used to determine statistical significance in mean differences and percentage differences in NAPLAN performance in the Reports. Equating errors are also taken into consideration, in addition to sampling and measurement errors, in the estimation of standard errors for the determination of statistical significance in the comparisons of means and percentages between years.

Equating Tests

In order to be able to compare the performance of students on different tests conducted across different years, an 'equating' process is completed to determine any variation in the difficulty of the tests so that the difficulty of one set of tests can be aligned to the level of difficulty of the second set of tests. This process enables tests to be located on a common scale and valid comparisons made between the performances of students on different tests.

In the case of NAPLAN, it is important to be able to equate tests in subsequent years so that comparisons can be made between student performances, comparisons that are valid because they are not affected by the variations in the relative difficulty of the tests.

In 2009, equating tests were developed so that future NAPLAN tests could be located on the same scale (the NAPLAN scale). The 2009 equating process used both 'on-shore' and 'off-shore' testing. Students in New Zealand participated in the testing and sat the 2008 NAPLAN tests and the equating tests. In Australia another sample population of students sat the 2009 NAPLAN tests and also the equating tests. Using a combination of equating methods, the 2008 and 2009 tests were able to be placed on the same scale through the process of the common equating tests.

From 2010 a sample of students from each year, drawing from all states and territories and school sectors, will sit the secure equating tests as well as the current year's NAPLAN tests. The equating tests will be administered by specially trained independent test administrators. This ensures that the security of the equating tests can be preserved.

The equating process for NAPLAN was informed by advice from the EAG. Care is taken to provide a high level of assurance as to the reliability of comparisons between years. The equating process provides confidence that any test difference has been taken into account before reporting any differences in student performance between years.

VII. REPORTING

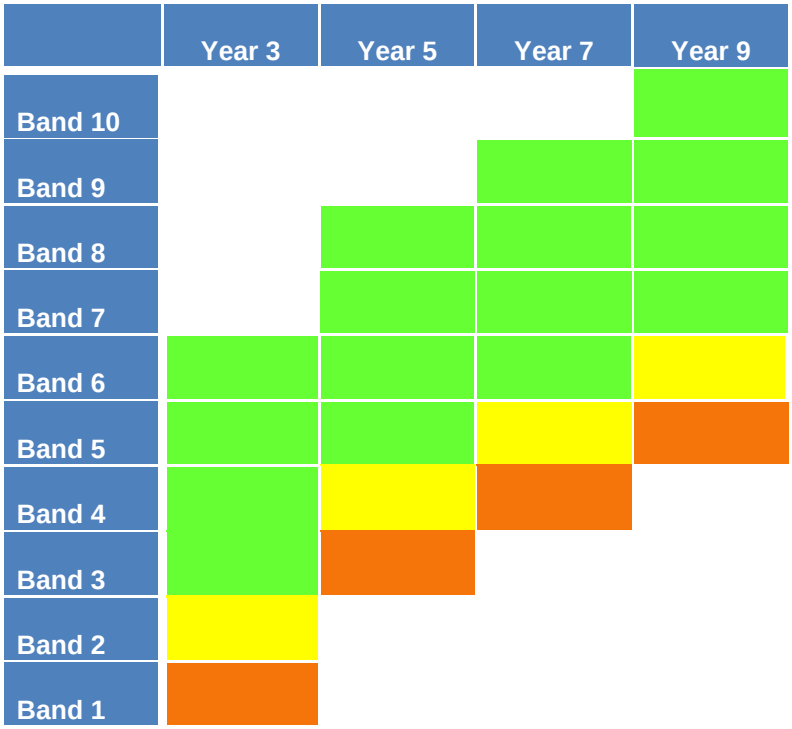
NAPLAN results are reported using five national achievement scales, one for each of the NAPLAN assessment domains of Reading, Writing, Spelling, Grammar and Punctuation, and Numeracy. Each scale consists of ten bands, which represent the increasing complexity

of the skills and understandings assessed by NAPLAN from Years 3 to 9. Six of these bands are used for reporting student performance in each year level.

The NAPLAN reporting scales are constructed so that any given scale score represents the same level of achievement over time. For example, a score of 700 in Reading in one year is equivalent to the same score in other testing years.

The ten proficiency bands on the NAPLAN reporting scales have the cut-points set at equal intervals apart. Year 3 results are reported against Band 1 to Band 6, Year 5 results reported against Band 3 to Band 8, Year 7 results reported against Band 4 to Band 9, and Year 9 results reported against Band 5 to Band 10.

Students were deemed to have performed above national minimum standard if their scores fell in the green bands at their respective year level and below national minimum standard if their scores fell in the orange band



NAPLAN results are reported nationally through the Summary (September release) and National Reports (December release) and at the school and student level in the form of reports to parents.

The Summary Report, released in September, provides national level data and a picture of how each State and Territory compares against a common assessment scale.

The National Report, released in December is a more detailed report which shows results at national and State and Territory levels by achievement levels and/or mean scores as well as by gender, Indigenous status, language background other than English status, parental occupation and parental education, and geographical location (metropolitan, provincial, remote and very remote) at each year level and for each domain of the test.

As part of NAPLAN, all schools are provided with a detailed report on their (individual) students' results. Detailed results for the school on the full range of NAPLAN achievement

are provided, including the number of students in each band at each year level. Principals and teachers can use this information to monitor student progress and identify students in need of additional support. The information can be used for diagnostic purposes and can assist them in their planning to cater for the individual needs of each student.

Parents of students taking the tests also receive a report showing their child's results along with common national key information about his or her performance, such as the national average. For example the parents of a Year 3 student will receive a report that shows the national average, the range for the middle 60 per cent of students, the national minimum standard and how their child is performing.

Some States and Territories will also provide the school average as well as the items the student successfully responded to and those they didn't. With this information parents can see if their child is performing at a satisfactory standard compared to other students in Australia, or if they need specialised intervention. As the child progresses through the years of schooling they can compare their child's position on the scale with previous years and monitor the improvement over time.