

OFFICIAL

NAPLAN 2022

Technical Report

June 2023

Acknowledgement of Country

ACARA acknowledges the Traditional Owners and Custodians of Country and Place throughout Australia and their continuing connection to land, waters, sky and community. We pay our respects to them and their cultures, and Elders past and present.

Copyright

© Australian Curriculum, Assessment and Reporting Authority (ACARA) 2023, unless otherwise indicated. Subject to the exceptions listed below, copyright in this document is licensed under a Creative Commons Attribution 4.0 International (CC BY) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that you can use these materials for any purpose, including commercial use, provided that you attribute ACARA as the source of the copyright material.



Exceptions

The Creative Commons licence does not apply to:

1. logos, including (without limitation) the ACARA logo, the NAP logo, the Australian Curriculum logo, the My School logo, the Australian Government logo and the Education Services Australia Limited logo;
2. other trade mark protected material;
3. photographs; and
4. material owned by third parties that has been reproduced with their permission. Permission will need to be obtained from third parties to re-use their material.

Attribution

ACARA requests attribution as: “© Australian Curriculum, Assessment and Reporting Authority (ACARA) 2023, unless otherwise indicated. This material was downloaded from [insert website address] (accessed [insert date]) and [was][was not] modified. The material is licensed under CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>). ACARA does not endorse any product that uses ACARA’s material or make any representations as to the quality of such products. Any product that uses ACARA’s material should not be taken to be affiliated with ACARA or have the sponsorship or approval of ACARA. It is up to each person to make their own assessment of the product”.

Contact details

Australian Curriculum, Assessment and Reporting Authority Level 13, Tower B, Centennial Plaza, 280 Elizabeth Street Sydney NSW 2000 T 1300 895 563 F 1800 982 118
www.acara.edu.au

The appropriate citation for this report is: Australian Curriculum, Assessment and Reporting Authority 2023, NAPLAN Technical Report for 2022, ACARA, Sydney.

Contents

Acknowledgement of Country	2
Copyright	2
List of tables	5
List of figures	7
Chapter 1: Introduction	10
Chapter 2: Item development	11
Numeracy item development	11
Reading item development.....	12
Conventions of language item development	14
Writing task development.....	14
Chapter 3: Item trial	16
Item trial test design.....	16
Test composition	20
Sampling	21
Approach.....	21
Sample size.....	21
Exclusions.....	22
Stratification	23
Test administration.....	24
Participants.....	24
Marking of writing.....	25
Psychometric analysis of item trial data	26
Item selection for the 2022 NAPLAN tests	28
Chapter 4: Test construction	29
Multistage, tailored test design.....	29
Construction of NAPLAN online tests.....	31
Test length	31
Difficulty of testlets.....	32
Item types for online tests.....	34
Curriculum coverage	35
Paper test design	42
Writing test design.....	43
Writing marking training and quality assurance	45
Example items in reporting bands	48
Setting branching rules	63
Branching rules for numeracy, reading, and grammar and punctuation tests	64
Branching rules for spelling.....	67
Pathway utilisation.....	69

Chapter 5: Data collection and preparation	71
Data collection and delivery.....	71
Paper tests.....	72
Online tests.....	72
Data cleaning validation process.....	73
Data preparation.....	73
Distribution of not reached items.....	74
Not reached items in online tests.....	74
Final student participation rates.....	77
Chapter 6: Scaling methodology and outcomes	79
Scaling model.....	79
Software used for analyses.....	79
Item calibration.....	79
Review of test and item characteristics.....	80
Test reliability.....	81
Test targeting and item spread.....	81
Item fit.....	86
Differential Item Functioning (DIF) analyses.....	88
Estimation of student ability and generation of PVs.....	95
Chapter 7: Equating procedures	98
Equating of numeracy, reading, spelling, and grammar and punctuation results.....	98
Horizontal equating shifts of the online tests.....	99
Equating paper tests.....	106
Scaling factors.....	107
Equating of writing results.....	107
Pairwise study results.....	109
Summary of equating parameter estimates for NAPLAN 2022.....	111
Estimating equating errors.....	113
Estimating long-term trend errors.....	117
Chapter 8: NAPLAN proficiency bands	118
Illustrations.....	124
Chapter 9: Reporting of national results	125
Calculation of statistics using plausible values.....	125
Computation of standard errors.....	125
Sampling error.....	125
Measurement error.....	126
Testing for differences.....	127
Effect sizes.....	127
Effect size for comparing means.....	128
Effect size for differences in percentages.....	129
Effect size for long-term trends.....	129

References	130
Appendices	131
Appendix A.....	131
Appendix B.....	131
Appendix C.....	131
Appendix D.....	131
Appendix E.....	131
Appendix F.....	131
Appendix G.....	131
Appendix H.....	131
Appendix I.....	131
Appendix J.....	131
Appendix K.....	131
Appendix L.....	132
Appendix M.....	132
Appendix N.....	132
Appendix O.....	132
Appendix P.....	132

List of tables

Table 1: Number of items developed for numeracy.....	12
Table 2: Development of spelling items.....	14
Table 3: Development of grammar and punctuation items.....	14
Table 4: Numeracy test design for primary students for the item trial held in 2021.....	17
Table 5: Numeracy test design for secondary students for the item trial held in 2021.....	17
Table 6: Reading test node structure for the item trial held in 2021.....	18
Table 7: Reading test design for the item trial held in 2021.....	18
Table 8: Conventions of language test design for the item trial held in 2021.....	19
Table 9: Writing test design for the NAPLAN item trial held in 2021.....	19
Table 10: Composition of the trial numeracy item pool including horizontal and vertical links.....	20
Table 11: Composition of the trial reading item pool including horizontal and vertical links.....	20
Table 12: Composition of the trial grammar and punctuation item pool including horizontal and vertical links.....	20
Table 13: Composition of the trial spelling item pool including horizontal and vertical links.....	20
Table 14: Primary schools sample.....	22
Table 15: Secondary schools sample.....	22
Table 16: Explicit stratification primary sample.....	23
Table 17: Explicit stratification secondary sample.....	23
Table 18: Number of students participating in the online item trial sample, by domain and year level.....	25
Table 19: Number of responses for writing by genre, mode, task and year level.....	25
Table 20: NAPLAN online numeracy test: number of items and time available.....	32

Table 21: NAPLAN online reading test: number of items and time available.....	32
Table 22: NAPLAN online conventions of language test: number of items and time available.....	32
Table 23: NAPLAN online numeracy: predefined difficulty parameters for each testlet	33
Table 24: NAPLAN online reading: predefined difficulty parameters for each testlet	33
Table 25: NAPLAN online grammar and punctuation: predefined difficulty parameters for each testlet	33
Table 26: NAPLAN online spelling: predefined difficulty parameters for each testlet.....	33
Table 27: NAPLAN online numeracy: item types in the item pool by year level.....	34
Table 28: NAPLAN online reading: item types in the item pool by year level	34
Table 29: NAPLAN online conventions of language: item types in the item pool by year level	34
Table 30: NAPLAN numeracy Year 3 curriculum coverage by mode and pathway	35
Table 31: NAPLAN numeracy Year 5 curriculum coverage by mode and pathway	35
Table 32: NAPLAN numeracy Year 7 curriculum coverage by mode and pathway	36
Table 33: NAPLAN numeracy Year 9 curriculum coverage by mode and pathway	36
Table 34: NAPLAN reading Year 3 curriculum coverage by mode and pathway	37
Table 35: NAPLAN reading Year 5 curriculum coverage by mode and pathway	37
Table 36: NAPLAN reading Year 7 curriculum coverage by mode and pathway	38
Table 37: NAPLAN reading Year 9 curriculum coverage by mode and pathway	38
Table 38: NAPLAN conventions of language Year 3 curriculum coverage by mode and pathway	39
Table 39: NAPLAN conventions of language Year 5 curriculum coverage by mode and pathway	40
Table 40: NAPLAN conventions of language Year 7 curriculum coverage by mode and pathway	41
Table 41: NAPLAN conventions of language Year 9 curriculum coverage by mode and pathway	42
Table 42: NAPLAN numeracy paper test number of items and time available	43
Table 43: NAPLAN reading paper test number of items and time available	43
Table 44: NAPLAN conventions of language paper test number of items and time available	43
Table 45: NAPLAN writing prompt designation schedule according to test day.....	44
Table 46: Recommended allocation of time for the writing test.....	44
Table 47: NAPLAN narrative marking criteria and skill focus descriptions.....	45
Table 48: NAPLAN narrative marking criteria and skill focus descriptions.....	45
Table 49: Writing scripts marked for each jurisdiction.....	46
Table 50: Approximate number of NAPLAN writing markers per day by jurisdiction.....	46
Table 51: NAPLAN 2022 marking centre operational periods and duration by jurisdiction.....	46
Table 52: The number of Training, Practice and Control scripts developed for each prompt	47
Table 53: National marking protocols.....	48
Table 54: Numeracy example items in reporting bands.....	48
Table 55: Reading example items in reporting bands.....	52
Table 56: Grammar and punctuation example items in reporting bands.....	57
Table 57: Spelling items in bands.....	60
Table 58: Example writing prompt.....	63
Table 59: Stage 1 cut scores (testlet A to C B D)	65

Table 60: Stage 2 cut scores (testlet AB to C E F).....	66
Table 61: Stage 2 cut scores (testlet AD–C E F)	67
Table 62: Stage 1, testlet SA–SB SD cut scores.....	68
Table 63: Stage 2, testlets SA–SB to PB PD cut scores	69
Table 64: Stage 2, testlet SA–SD to PB PD cut scores	69
Table 65: Rules for data coding	73
Table 66: Pathway assignment rules to incomplete online tests.....	74
Table 67: Student participation rate	77
Table 68: Reliability (EAP/PV, WLE) for NAPLAN 2022 tests	81
Table 69: Summary of item statistics in NAPLAN 2022 online tests	86
Table 70: Number of items showing gender DIF by domain by year level	89
Table 71: Number of items showing LBOTE DIF by domain by year level	90
Table 72: Number of items showing Indigenous DIF by domain by year level.....	91
Table 73: Number of items showing state/territory DIF by domain by year level.....	93
Table 74: Number of students by device	94
Table 75: Number of items showing device DIF by domain and year level.....	95
Table 76: Equating design for online tests	98
Table 77: Horizontal link review summary for online tests.....	106
Table 78: Horizontal equating shifts (Shift22to21) between 2022 item locations and 2021 item locations by year level by domain for online tests	106
Table 79: Parameters for locating 2022 paper test scales on the 2021 scales by year level and domain	107
Table 81: Summary of parameters for transforming the 2022 online logit scores to the NAPLAN reporting scales.....	112
Table 82: Summary of parameters for transforming the 2022 paper logit scores to the NAPLAN reporting scales.....	113
Table 83: Standard errors of equating	116
Table 84: Lower bounds of proficiency bands in scale scores and in logits.....	118
Table 85: Described scale for numeracy	119
Table 86: Described scale for reading.....	120
Table 87: Described scale for writing.....	121
Table 88: Described scale for conventions of language	122
Table 89: Intercept and slope of growth regression by domain.....	128

List of figures

Figure 1: A sample ICC for a well-performing item	27
Figure 2: A sample item category characteristic curve for a well-performing MC item	27
Figure 3: The multistage tailored test design for numeracy, reading and grammar and punctuation ..	30
Figure 4: Online test design for conventions of language	31
Figure 5: Test information functions: curves for testlets C, B and D.....	64
Figure 6: Stage 1 testlet A–C B D cut scores.....	65

Figure 7: Stage 2 testlet AB–C E F cut scores.....	66
Figure 8: Stage 2 testlet AD–C E F cut scores	67
Figure 9: Stage 1 testlet SA–SB SD cut scores.....	68
Figure 10: Stage 2 testlet SA–SB to PB PD cut scores.....	68
Figure 11: Stage 2 testlets SA–SD to PB PD cut scores.....	69
Figure 12: Percentage of students assigned to each pathway in Year 3 numeracy	70
Figure 13: Ability distribution by pathway for Year 3 numeracy.....	70
Figure 14: NAPLAN 2022 stage 1 data flow	72
Figure 15: NAPLAN 2022 stage 2 data flow	72
Figure 16: Trailing missing percentage in numeracy.....	75
Figure 17: Trailing missing percentage in reading	75
Figure 18: Trailing missing percentage in spelling	76
Figure 19: Trailing missing percentage in grammar and punctuation	76
Figure 20: NAPLAN 2022 participation categories.....	77
Figure 21: Wright map for Year 3 numeracy online test (an example).....	83
Figure 22: Wright map for writing test (a polytomous example).....	84
Figure 23: Thurstonian thresholds for writing test.....	85
Figure 24: Item characteristic curves for an item with $\text{infit} = 1.00$	87
Figure 25: Item characteristic curves for an item with $\text{infit} = 1.26$	88
Figure 26: Example of item characteristic curves displaying gender DIF†	89
Figure 27: Example of item characteristic curves displaying LBOTE DIF†	90
Figure 28: Example of item characteristic curves displaying Indigenous DIF†	91
Figure 29: Example of item characteristic curves displaying jurisdictional DIF	92
Figure 30: Conditioning variables for the multidimensional item response model with latent regression model	97
Figure 31: Scatterplot of numeracy, horizontal equating items between 2022 and 2021 for Year 3 online students	100
Figure 32: Scatterplot of numeracy, horizontal equating items between 2022 and 2021 for Year 5 online students	100
Figure 33: Scatterplot of numeracy, horizontal equating items between 2022 and 2021 for Year 7 online students	101
Figure 34: Scatterplot of numeracy, horizontal equating items between 2022 and 2021 for Year 9 online students	101
Figure 35: Scatterplot of reading, horizontal equating items between 2022 and 2021 for Year 3 online students	101
Figure 36: Scatterplot of reading, horizontal equating items between 2022 and 2021 for Year 5 online students	102
Figure 37: Scatterplot of reading, horizontal equating items between 2022 and 2021 for Year 7 online students	102
Figure 38: Scatterplot of reading, horizontal equating items between 2022 and 2021 for Year 9 online students	102
Figure 39: Scatterplot of spelling, horizontal equating items between 2022 and 2021 for Year 3 online students	103
Figure 40: Scatterplot of spelling, horizontal equating items between 2022 and 2021 for Year 5 online students	103
Figure 41: Scatterplot of spelling, horizontal equating items between 2022 and 2021 for Year 7 online students	103

Figure 42: Scatterplot of spelling, horizontal equating items between 2022 and 2021 for Year 9 online students	104
Figure 43: Scatterplot of grammar and punctuation, horizontal equating items between 2022 and 2021 for Year 3 online students	104
Figure 44: Scatterplot of grammar and punctuation, horizontal equating items between 2022 and 2021 for Year 5 online students	104
Figure 45: Scatterplot of grammar and punctuation, horizontal equating items between 2022 and 2021 for Year 7 online students	105
Figure 46: Scatterplot of grammar and punctuation, horizontal equating items between 2022 and 2021 for Year 9 online students	105
Figure 47: Scatterplot for writing criteria between 2022 and 2021 online and paper tests	108
Figure 48: Pairwise location estimates from the 2021 project plotted against the estimates from the 2022 project for the 2021 scripts	110
Figure 49: Rubric location estimates plotted against the pairwise location estimates from the 2022 project for the 2021 and 2022 scripts	110
Figure 50: Rubric location estimates plotted against the pairwise location estimates from the 2022 project for the 2021 and 2022 year 3 paper scripts	111
Figure 51: A schematic of the equating errors accumulated across NAPLAN administrations	114
Figure 52: A schematic of the equating errors accumulated across NAPLAN administrations	114
Figure 53: Schematic picture of proficiency bands by year levels	124
Figure 54: Examples in SPSS and SAS for estimating sampling variance	126
Figure 55: Logarithmic regression function for numeracy	128

Chapter 1: Introduction

The first National Assessment Program – Literacy and Numeracy (NAPLAN) tests took place in 2008. They were conducted by the then Ministerial Council for Education, Employment, Training and Youth Affairs (MCEETYA), now Education Ministers Meeting. This was the first time all students in Australia in Years 3, 5, 7 and 9 were assessed in literacy and numeracy using year level specific tests. The national tests, which replaced a raft of tests administered by Australian states and territories, improved the comparability of students' results across states and territories.

NAPLAN data provide federal and jurisdictional governments, schools and parents/carers with information about whether young Australians are reaching important educational goals.

NAPLAN tests are the only Australian assessments that provide nationally comparable data on the performance of students in the vital areas of literacy and numeracy. This gives NAPLAN a unique role in providing robust data to inform and support improvements to teaching and learning practices in Australian schools.

The NAPLAN 2022 tests were administered nationally in May. As in previous cycles of NAPLAN, students at each year level were assessed in the domains of reading, writing, conventions of language (spelling, grammar and punctuation) and numeracy.

The Australian Council for Educational Research (ACER) was appointed by the Australian Curriculum, Assessment and Reporting Authority (ACARA) to undertake the central analysis of test data from the NAPLAN 2022 administration.

The central analysis of NAPLAN data essentially involves placing each domain test in the current year onto the relevant NAPLAN historic domain scale through test calibration and a series of horizontal and vertical equating exercises. The equating process enables the reporting of student performance on the NAPLAN historic scale for each of the NAPLAN domains and for comparisons across year levels and over assessment cycles.

NAPLAN results are reported using 5 national achievement scales, one for each of the assessed aspects of literacy – reading, writing, spelling, grammar and punctuation – and one for numeracy. Each NAPLAN achievement scale spans Years 3, 5, 7 and 9 with scores that range from approximately 0 to 1,000. There are also 10 proficiency bands that span Years 3, 5, 7 and 9. Each year level is reported against 6 of these bands.

Over one million students in Years 3, 5, 7 and 9 in all states and territories of Australia participated in NAPLAN 2022. From 2008 to 2017, NAPLAN delivered only paper-based tests. From 2018, NAPLAN delivered both paper-based tests and online multistage adaptive tailored tests. The online tailored tests in reading, spelling, grammar and punctuation, and numeracy were delivered to students in participating schools. In 2022, approximately 95% of students took the NAPLAN test online (50% in 2021, 30% in 2019 and 15% in 2018). NAPLAN was cancelled in 2020 due to the COVID-19 pandemic.

Four outcome reports were produced for NAPLAN 2022. The first report was the Student and School Summary report (SSSR). This interactive report was produced for online schools and provided an opportunity for schools to take a first glance at the achievement of their students. The second report type was the Individual Student Report (ISR), providing information to parents/carers about their children's performance on the NAPLAN tests. The third report was the official NAPLAN 2022 National Report that was based on full census data. The National Report for 2022 and all previous NAPLAN assessments are available on the ACARA website. The final cut of the census data was used for the school-level online My School reports, which are beyond the scope of this technical report.

The aim of this technical report is to describe in detail the methodology used for NAPLAN 2022. Chapter 2 of this report describes the NAPLAN 2022 item development. Chapter 3 of this report describes the NAPLAN 2022 item trial. Chapter 4 describes the test design. Chapter 5 describes the data preparation process. Chapter 6 describes psychometric scaling methodology and outcomes. Chapter 7 describes the test equating processes to place the NAPLAN 2022 tests on the NAPLAN historic scales. Chapter 8 describes the proficiency bands on the NAPLAN scales. Chapter 9 describes the methodology used for reporting of NAPLAN 2022 performance.

Technical details that are not included in this report are available upon request from ACARA.

Chapter 2: Item development

The aim of this chapter is to describe the item development activities that took place in preparation for the NAPLAN 2022 test.

Commercial contractors developed new items in all the assessment domains with the exception of spelling, where items were developed by the conventions of language test development team. Item developers complied with the following documents:

- NAPLAN Assessment framework and Item development guidelines
- ACARA accessibility guidelines
- Assessment and Delivery System (ADS) user guides
- Web Content Accessibility Guidelines (WGAG2.0 AA).

Items for the item trial conducted in 2021 were developed in batches across the 2 project periods because of interruptions to the assessment program caused by COVID. The 2 development cycles spanned from September 2019 until May 2020 and then again from September 2020 until May 2021.

Items in each batch were reviewed by ACARA, the National Testing Working Group (NTWG) and independent domain experts. Feedback was synthesised by ACARA and the items requiring modification were returned to the contractors for revisions. All modified items were reviewed by ACARA before final delivery in May 2020 and May 2021.

Contractors submitted compliance tables showing how the items met the specifications outlined in the contracts. Source files for all graphics were supplied and copyright licenses for all third-party material centrally stored in ACARA's intellectual property management platform.

Where appropriate, graphics were converted to scaled vector graphics (SVGs) by the ACARA graphic designers to better accommodate universal graphic design and enable graphics to be magnified without losing clarity.

Items that contained table shading were copied, modified and added as Disability Adjustment Code (DAC) alternative items for students who require items in black and white, or use a coloured background adjustment (lilac, blue, yellow and green).

Audio was recorded for all numeracy, audio dictation (spelling) items and writing prompts prior to trialling. This entailed scripting of each item (including DAC alternative items), recording, editing, attaching audio and checking of all recordings.

Numeracy item development

Items for the NAPLAN 2022 numeracy tests were procured from 2 separate contractors. The main contractor, the National Foundation for Educational Research (NFER), provided ACARA with items from the Number and Algebra, Measurement and Geometry, and Statistics and Probability strands for all test years. This included a small number of innovative item types previously unused in NAPLAN.

The second contractor, the University of Melbourne (UoM), provided high and low facility items from the Number and Algebra, Measurement and Geometry, and Statistics and Probability strands for all test years.

Approximately 10% of the delivered items required accessibility substitute items. These were prepared by ACARA.

The numbers of items developed for each Australian Curriculum strand are shown in Table 1.

Table 1: Number of items developed for numeracy

	NFER 2019–2020	NFER 2020–2021	UoM 2020–2021
Number and Algebra	326	237	123
Measurement and Geometry	144	132	71
Statistics and Probability	74	67	31
Total	544	436	225

Items were developed across the full range of item difficulties needed for the main study test design. Items were assigned proficiency standards that cover a range of cognitive demands: fluency, understanding, problem-solving and reasoning.

Items were supplied to cover 3 broad item types: 55% multiple-choice(s), 30% text entry and 15% technology-enhanced items.

Reading item development

From August 2019 to May 2020:

- ACARA contracted University of New South Wales Global Assessments (UNSWG) to produce 36 reading units predominantly targeting the lower and upper end of the performance scale for Years 3 and 5. UNSWG's final delivery included 36 stimulus texts and 291 items.
- ACARA contracted NFER to produce 36 reading units predominantly targeting the lower and upper end of the performance scale for Years 7 and 9. NFER's final delivery included 36 stimulus texts and 289 items.
- ACARA contracted UNSWG and NFER to each provide 45 items to supplement pre-existing reading units, most of which had been trialled but not yet used in a main study. These additional items were required to ensure the pre-existing units could readily fit testlet boundaries. Each contractor's final delivery included 45 items.
- ACARA contracted NFER to produce 58 innovative standalone items (13 at each of Years 3 and 5, and 16 at each of Years 7 and 9). Standalone items are items targeting specific skills that can be used on their own or with a very short stimulus text. These items were designed to target the lowest end of the performance scale, with a focus on the types of texts encountered in everyday contexts (for example, applications, menus, shelves, instructions). NFER's final delivery included 62 items.
- The ACARA Reading Test Development Team hosted 2 author workshops. Eight Australian authors renowned for their writing for children and young adults worked with the Reading Test Development Team over a 2-day period to produce new imaginative and persuasive stimulus texts targeting the lowest and highest ends of the performance scale. In total, 85 stimulus texts were developed.

From August 2020 to May 2021:

- ACARA contracted Educational Assessments (Janison) to produce 18 reading units predominantly targeting the upper end of the performance scale for Years 3 and 5. The contractor's final delivery included 18 stimulus texts and 144 items.
- ACARA contracted NFER to produce 36 reading units: 18 units predominantly targeting the lower end of the performance scale for Years 3 and 5, and 18 units predominantly targeting the upper end of the performance scale for Years 7 and 9. NFER's final delivery included 36 stimulus texts and 288 items.

- ACARA contracted UoM and NFER to each provide 45 items to supplement pre-existing reading units, most of which had been trialled but not yet used in a main study. These additional items were required to ensure the pre-existing units could readily fit testlet boundaries. Each contractor's final delivery included 45 items.
- ACARA contracted UoM and NFER to each produce 9 items sets for pre-existing reading stimulus texts developed during the author workshops. Each contractor's final delivery included 72 items.
- ACARA contracted NFER to produce 58 innovative items that could act as standalone items or small units (2 to 4 items on a short stimulus text). These items were designed to target the highest end of the performance scale.
- The ACARA Reading Test Development Team repaired 37 units: writing items for units that had insufficient items within a testlet range to enable inclusion in the NAPLAN test. The team produced 142 items.

Stage 1 of the reading item development cycle began with the submission and review of a matrix outlining the units to be developed for each year group. Required metadata included genre and text type, topic and a summary, word length, text complexity, targeted testlet, and source. This iterative matrix was submitted and revised throughout the item development cycle.

The difficulty of items, to a large extent, was dependent on the complexity of the stimulus texts. A common concern for NAPLAN reading items was appropriate targeting for early childhood and entry-level texts for all years. Entry-level texts target students working at a skill level one to 3 years below their school year level, using subject matter that is still engaging and age appropriate for these students. All Year 3 texts and entry-level Year 5 texts were reviewed by experienced pre-primary and/or primary teachers. Entry-level Year 7 and Year 9 texts were also reviewed by teachers who have extensive experience with students of lower reading ability.

ACARA's internal graphic designer and the contractors' desktop publishing teams were tasked with designing and illustrating stimulus texts that were engaging and that provided appropriate support for students reading the texts. Special attention was paid to ensuring:

- online readability, particularly in font selection, and text layouts aimed at reducing the need for scrolling
- accessibility for visually impaired students, taking into account ACARA's guidelines for colour, contrast and font selection
- resource file size being kept at a maximum of 120 kb per text.

The stimulus texts in each cycle were reviewed in 2 batches by panels of assessment and curriculum experts convened by each jurisdiction. Following the review and subsequent modification stages, stimulus texts were accepted for item development.

During stage 2 of the cycle, multiple levels of review were undertaken by the contractors prior to items being submitted to ACARA. These included reviews by item writers, subject and language specialists, reviewers from First Nations Australian backgrounds, item development managers and editors. ACARA also requested follow-up cultural reviews for some texts and these were provided. For all informative texts, a fact check was carried out by a team member other than the text writer and again by ACARA during the item review process. All texts were reviewed for intellectual property and moral rights.

ACARA facilitated 5 reading reviews of the reading stimuli and items over the period of each item development cycle. Feedback was sought from the NTWG and ACARA's student diversity specialist. ACARA synthesised the feedback, and items were returned to contractors classified as "accepted", "needing modification as specified" or "needing replacement".

Conventions of language item development

Conventions of language tests consist of a spelling section, and a grammar and punctuation section.

Spelling items were developed by the ACARA writing/conventions of language team. Target words were sourced from different sources including errors in past NAPLAN writing trial scripts. The team identified the words students commonly misspell as well as likely error patterns. The words were used in simple, concise, age-appropriate context sentences that provided enough support for the misspelt words to be readily understood. Items were allocated to audio dictation, mistake-identified or mistake-not-identified (proofreading) sections of the spelling test and assigned targeted testlets according to year level, predicted difficulty, skill focus and item type. Each audio dictation item was paired with an accessibility alternative (AA) mistake-identified item for hearing-impaired students that was identical in content but had a single element of the target word spelled incorrectly.

Table 2: Development of spelling items

Spelling items	Sept 2019 – May 2020	Sept 2020 – May 2021	TOTAL
Audio dictation	58 per year level	58 per year level	116 per year level
Mistake-identified	25 per year level	25 per year level	50 per year level
Mistake-not-identified	25 per year level	25 per year level	50 per year level
TOTAL	108 per year level	108 per year level	

Grammar and punctuation items were developed by the National Foundation for Educational Research (NFER) (Years 7 and 9) and Janison (Years 3 and 5). These contractors delivered 4 batches of items, totalling approximately 351 grammar and 94 punctuation items: 6 testlets for each of Years 3, 5, 7 and 9. ACARA facilitated 5 reviews of the grammar and punctuation items over a 6-month period. Additional feedback on accessibility alternative items was sought from NTWG and ACARA's student diversity specialist. All modifications to items were made by ACARA.

Table 3: Development of grammar and punctuation items

Grammar and punctuation items	Sept 2019 – May 2020	Sept 2020 – May 2021	TOTAL
Grammar	154 per year level	154 per year level	308 per year level
Punctuation	76 per year level	76 per year level	152 per year level
TOTAL	230 per year level	230 per year level	

Items were developed across the full range of item skills and difficulties. Both contractors were provided with a skill index that required them to target particular skills at a range of difficulty levels. All items were cross-referenced in a compliance grid that indicated the breadth of skills covered and the scope of difficulty. Each item was assigned a facility estimate and an estimated testlet (for grammar and punctuation: C/E/F).

Writing task development

Prompts for the NAPLAN writing trial held in 2021 were developed according to the following process:

1. Education experts from all jurisdictions contributed to the development of a large pool of potential writing tasks, intended for students in Years 3 and 5, and/or Years 7 and 9. Each jurisdiction convened panels of experts with significant experience in the assessment of writing in the development of their contributions to the pool.

Expert panels in each jurisdiction undertook a 4-stage review of all writing tasks in the pool to ensure that the topics progressed for further refinement into prompts were accessible to students from a wide range of backgrounds and abilities. Panels considered what students might write about and whether the task would be fair for all students. In the early stages of the review, the panels prioritised the writing topics, providing feedback where necessary. In later stages of the review, they reduced the pool down to the most suitable tasks and suggested changes to the wording and images for the prompts in readiness for trial. Educators representing First Nations Australian students and students with disability also reviewed and provided advice on the writing tasks before they were trialled.

Chapter 3: Item trial

The aim of this chapter is to describe the item trialling and psychometric analysis for the NAPLAN 2022 tests. The first part of this chapter describes the item trial sampling and administration, and the second part focuses on the psychometric analysis.

As part of the NAPLAN item trial process, items were presented to a sample of students in the relevant year level to obtain critical item performance data to guide construction of the final NAPLAN tests and develop each domain's item bank. Trialling allowed additional quantitative and qualitative feedback on the tests to be gathered, including time on task, engagement with test content and identification of online display issues. Individual items and suites of test items (based on common stimulus texts) were administered to samples of students within Australia. Psychometric analysis of the data, conducted after the trial, was used to evaluate the performance of each individual item.

The Australian Council for Educational Research (ACER) was engaged to analyse items that were included in tests according to the trial design developed by ACARA for each of the test domains.

Item trial test design

The trial test included items from the previous main study so that the trial results could be equated to the historical NAPLAN scale.

As items presented at the end of a test have the potential to perform differently from those presented at the beginning (due to accumulated cognitive load or time pressure), the trial tests were designed so that testlets were presented at differing positions within the tests. To illustrate, Year 3 reading had the following rotational design:

- twenty-four testlets plus one testlet of stand-alone items¹
- four nodes: node 1 had one testlet with approximately 8 stand-alone items; nodes 2, 3 and 4 had 8 testlets each
- students started by answering a single stand-alone item from node 1, then **one** of the following 3 options:
 - one testlet from node 2 followed by one testlet from node 3 and then one testlet from node 4
 - one testlet from node 3 followed by one testlet from node 4 and then one testlet from node 2
 - one testlet from node 4 followed by one testlet from node 2 and then one testlet from node 3.

As such, items were trialled in 3 different positions, with one third of students seeing an item in each of the first, middle and final stage of the test.

Each student sat 2 assessment events, composed of either 2 non-writing domains, one non-writing domain and one writing domain, or 2 writing domains. Trial test designs for the each of the domains are presented in Table 4 to Table 9.

In both primary and secondary numeracy tests, testlets in stage 1 were randomly assigned to each student; in secondary tests only, one of 3 testlets was randomly assigned in stage 2.

Table 4: Numeracy test design for primary students for the item trial held in 2021

Primary school		
Stage 1	Stage 2	Stage 3
Node 1	Node 2	Node 3
T01	T02	T04
T02	T03	T05
T03	T04	T06
T04	T05	H01
T05	T06	T07
T06	H01	T08
H01	T07	T09
T07	T08	T10
T08	T09	T11
T09	T10	T12
T10	T11	H02
T11	T12	T13
T12	H02	T14
H02	T13	T15
T13	T14	T16
T14	T15	T17
T15	T16	T18
T16	T17	H03
T17	T18	T01
T18	H03	T02
H03	T01	T03

Table 5: Numeracy test design for secondary students for the item trial held in 2021

Secondary school			
Stage 1	Stage 2	Stage 3	Stage 4
Node 1	Node 2	Node 3	Node 4
HN01	T08	T09	T11
HN01	T18	T19	T21
HN01	T22	T01	T03
NC01	T01	T02	T04
NC01	T10	T11	H01
NC01	T16	T17	T19
NC02	T02	T03	T05
NC02	T12	H01	T14
NC02	T17	T18	T20
NC03	H01	T13	T15

NC03	T03	T04	T06
NC03	T19	T20	T22
NC04	T04	T05	T07
NC04	T13	T14	H02
NC04	T20	T21	T01
NC05	T05	T06	T08
NC05	T09	T10	T12
NC05	T15	H02	T17
NC06	H02	T16	T18
NC06	T06	T07	T09
NC06	T11	T12	T13
NC07	T07	T08	T10
NC07	T14	T15	T16
NC07	T21	T22	T02

In each reading test, one item from testlet 1 was randomly assigned to each student in stage 1; in stage 2, one node was randomly assigned to each student. Testlets in sets 1 and 2 consisted of 2 reading units. Testlets in set 3 consisted of 2 units for primary school students and 3 units for secondary students.

Table 6: Reading test node structure for the item trial held in 2021

Stage 1 (one item)	Stage 2 (early)	Stage 3 (middle)	Stage 4 (late)
	Node 2:	Node 5:	Node 8:
Node 1:	Set 1	Set 2	Set 3
One random item from Testlet 1	Node 3:	Node 6:	Node 9:
	Set 2	Set 3	Set 1
	Node 4:	Node 7:	Node 10:
	Set 3	Set 1	Set 2

Table 7: Reading test design for the item trial held in 2021

Set 1	Set 2	Set 3
Testlet 2	Testlet 10	Testlet 18
Testlet 3	Testlet 11	Testlet 19
Testlet 4	Testlet 12	Testlet 20
Testlet 5	Testlet 13	Testlet 21
Testlet 6	Testlet 14	Testlet 22
Testlet 7	Testlet 15	Testlet 23
Testlet 8	Testlet 16	Testlet 24
Testlet 9	Testlet 17	Testlet 25

In each of the convention of language tests, one testlet was randomly assigned to each student in stage 1 (grammar and punctuation) and stage 3 (spelling).

Table 8: Conventions of language test design for the item trial held in 2021

Stage 1	Stage 2	Stage 3	Stage 4
Node 1	Node 2	Node 3	Node 4
G1	G2	AD2	PR12
G2	G3	AD3	PR13
G3	G4	AD4	PR14
G4	G5	AD5	PR15
G5	G6	AD6	PR16
G6	G7	AD7	AD1
G7	G8	AD8	AD2
G8	G9	AD9	AD3
G9	G10	AD10	AD4
G10	G11	PR11	AD5
G11	G12	PR12	AD6
G12	G13	PR13	AD7
G13	G14	PR14	AD8
G14	G15	PR15	AD9
G15	G16	PR16	AD10
G16	G1	AD1	PR11

In the writing test, students were assigned one of 10 writing prompts: 8 of the 10 prompts were from the preferred genre for selection for NAPLAN 2022; the remaining 2 prompts were from the non-preferred genre.

Table 9: Writing test design for the NAPLAN item trial held in 2021

Stage 1	
Node 1	Genre
W1	Preferred
W2	Preferred
W3	Preferred
W4	Preferred
W5	Preferred
W6	Preferred
W7	Preferred
W8	Preferred
W9	Non-preferred
W10	Non-preferred

A number of items were included in adjacent NAPLAN year levels (for example, Year 3 and Year 5). This enabled reviewing the psychometric properties of the items for several year levels. Depending on these properties, items could be used for the main study in only one year level or in both year levels.

Test composition

Table 10 to Table 13 show the composition of the trial pools by domain and by item format: technology enhanced items (which includes text entry), multiple-choice (MC) and multiple-choices (MCs).

Table 10: Composition of the trial numeracy item pool including horizontal and vertical links

Numeracy			
	TEI ¹	MC/S	Total
Year 3	114	159	273
Year 5	127	188	315
Year 7	179	237	416
Year 9	176	240	416

Table 11: Composition of the trial reading item pool including horizontal and vertical links

Reading			
	TEI	MC/S	Total
Year 3	62	274	336
Year 5	48	288	336
Year 7	34	366	400
Year 9	48	360	408

Table 12: Composition of the trial grammar and punctuation item pool including horizontal and vertical links

Grammar and punctuation			
	TEI	MC/S	Total
Year 3	111	97	208
Year 5	118	90	208
Year 7	128	80	208
Year 9	140	68	208

Table 13: Composition of the trial spelling item pool including horizontal and vertical links

Spelling			
	TEI	MC/S	Total
Year 3	207	0	207
Year 5	205	0	205
Year 7	206	0	206
Year 9	206	0	206

¹ TEI includes technology enhanced items and text entry item

For the writing domain, a shortlist of 8 narrative topics and 2 persuasive topics was selected for trial.

A short survey was included at the start of the trial tests. This survey collected information about

- gender
- device used
- general device usage
- where computer skills were learnt
- whether students were used to typing stories or essays at school.

Sampling

Approach

To support the placement of items on the NAPLAN scale, the test was administered to a sample of schools and students reflecting a range of educational contexts across a number of strata; for example, sector, Socio-Economic Indexes for Areas (SEIFA), geolocation, school size and previous NAPLAN performance. Samples of primary and secondary schools were drawn with the intention of capturing sufficient responses to attain stable item parameter estimates to inform item selection for NAPLAN 2022 – approximately 400 responses per item for each non-writing domain (numeracy, reading and language conventions) and 12 writing prompts. Students from each selected school completed 2 of the 4 domains to be tested, as opposed to the NAPLAN main study, in which students complete the test on all 4 domains.

In line with the practice of previous cycles, the proposed sample design for the NAPLAN 2022 item trial accommodated the following:

- **Population definition:** Two independent populations were surveyed – primary and secondary students. Students from Years 3, 5, 7 and 9 in Australian schools comprised the overall population of interest. For all states and territories except South Australia, primary students were those from Years 3 and 5, while secondary students were those from Years 7 and 9. For South Australia, Year 7 was classified as a primary year level at the time of the item trial¹.
- **Representativeness of sample:** The item trial sample is a sample of convenience across all states and territories. Trial schools were selected “to reflect the range of educational contexts around the nation and included schools from government, Catholic and independent sectors; low and high socio-economic areas; metropolitan and regional locations; large and small schools; and students from a variety of language backgrounds” (ACARA 2022, p 22).
- **Historical participation rate and provisions for participation due to COVID-19:** In 2019, overall student participation rates were around 81.3% of the target sample. To allow for the possibility of further participation losses through COVID-19 related issues, an overall participation rate of 75% was assumed.

Sample size

Two samples were drawn, one for primary and one for secondary schools. A maximum of 250 schools were sampled from each cohort. Within the selected schools, a full class was selected from each of the target grades in the cohort. To take account of the expected participation rate, some schools were instructed to select an additional class of students to perform the test.

Assuming an average of 25 students per session in each year level the school can provide, the total expected student yield was 6250 students for each year level. Up to 2 matched substitutes were identified for each sampled school.

¹ From 2022, the majority of South Australian students commenced secondary school in Year 7.

Table 14 and Table 15 show the school allocations at primary and secondary levels. The allocations are broadly proportional by population size. A minimum of 6 schools were targeted for the smallest jurisdictions of Australian Capital Territory, Northern Territory and Tasmania.

Table 14: Primary schools sample

Jurisdiction	Per cent of student population	Number of schools	Schools to sample
ACT	1.06%	42	6
NSW	34.70%	1410	81
NT	0.45%	26	6
Qld	21.65%	716	52
SA	4.15%	202	14
Tas	1.08%	65	6
Vic	26.80%	1107	62
WA	10.09%	453	23
Grand Total	100.00%	4021	250

Table 15: Secondary schools sample

Jurisdiction	Per cent of student population	Number of schools	Schools to sample
ACT	0.62%	9	6
NSW	36.18%	626	87
NT	0.44%	7	6
Qld	21.70%	321	49
SA	4.43%	95	10
Tas	0.75%	25	6
Vic	27.25%	446	65
WA	8.63%	141	21
Grand Total	100.00%	1670	250

Exclusions

School level exclusions¹:

- remote and very remote schools
- schools with fewer than 20 students in targeted years
- schools participating in NAP–ICTL field trial or main study
- schools participating in international studies (PISA field trial, and PIRLS main study and field trial)

¹ In other assessment years, participation in the previous year's equating and trial samples was considered in exclusion variables. This was not applicable for sampling undertaken in 2021 as no schools were sampled in 2020.

- distance education schools
- Montessori, Steiner and Waldorf schools
- special schools
- schools without NAPLAN performance data.

Stratification

Explicit stratification

Schools were stratified by state and sector for most jurisdictions. However, due to the smaller number of eligible schools in some of strata, some schools in smaller jurisdictions were merged into one stratum to be sampled (e.g. some Catholic and independent schools). In such cases, schools are merged into one non-government (NG) school stratum. Table 16 and Table 17 show the strata and definition for each sample:

Table 16: Explicit stratification primary sample

Stratum	State	Sector
01	ACT	Catholic
02	ACT	Government
03	ACT	Independent
04	NSW	Catholic
05	NSW	Government
06	NSW	Independent
07	NT	Government
08	NT	Catholic and independent
09	Qld	Catholic
10	Qld	Government
11	Qld	Independent
12	SA	Catholic
13	SA	Government
14	SA	Independent
15	Tas	Government
16	Tas	Catholic and independent
17	Vic	Catholic
18	Vic	Government
19	Vic	Independent
20	WA	Catholic
21	WA	Government
22	WA	Independent

Table 17: Explicit stratification secondary sample

Stratum	State	Sector
01	ACT	Government
02	ACT	Catholic and independent
03	NSW	Catholic
04	NSW	Government
05	NSW	Independent
06	NT	Government
07	NT	Catholic and independent
08	Qld	Catholic

09	Qld	Government
10	Qld	Independent
11	SA	Catholic
12	SA	Government
13	SA	Independent
14	TAS	Government
15	TAS	Catholic and independent
16	Vic	Catholic
17	Vic	Government
18	Vic	Independent
19	WA	Catholic
20	WA	Government
21	WA	Independent

Implicit stratification

Within each explicit stratum, schools were implicitly stratified by the following variables:

- school sector (Catholic/government/independent) for strata with merged sectors in small jurisdictions
- school size (Small <50, Large ≥50)
- NAPLAN performance quintiles
- state SEIFA IEO deciles
- ASGS Remoteness Area Classification (0 = Major cities of Australia / 1 = Inner Regional / 2=Outer Regional).

Test administration

The Educational Services Australia (ESA) test delivery platform was used to administer the trial tests in a sample of schools in Australia for all domains of the NAPLAN program. Schools from all states and territories participated in the trial from 26 July to 24 September 2021¹.

A trained invigilator was sent to each trial school to administer the trial tests. At the completion of each assessment session, the invigilator completed a session report to provide feedback about aspects of the trial administration. This feedback, in conjunction with feedback from a range of other sources, informed the selection and refinement of items for the final pool of assessment items and the design of the 2022 NAPLAN tests.

Participants

Due to the impact of COVID on school closures and accessibility, only 227 of the 493 sampled schools participated. While schools across all states and territories were sampled, no schools in New South Wales and Victoria participated due to COVID-related restrictions. As a result, modifications were made to the allocation of domains to schools to ensure maximal responses to items across domains and year levels. The number of participating students for each non-writing domain and year level is presented in Table 18. Students completed tests from 2 domains, with the majority of students completing 2 different domains. Despite attempts to sample enough students to achieve stable item parameter estimates during scaling, a considerable deficiency in the number of secondary schools completing the reading and numeracy tests was observed, thereby reducing the pool of items viable for selection for the main study.

¹ The testing window was extended from 13 August to 24 September to accommodate additional testing of schools due to COVID-related impacts on test administration.

Table 18: Number of students participating in the online item trial sample, by domain and year level

Domain	Year 3	Year 5	Year 7	Year 9	Total
Reading	1353	1361	623	553	3890
CoL	2082	2012	2013	1761	7868
Numeracy	1458	1564	768	701	4491

For the writing domain, approximately 5000 students each responded to 2 of the tasks under test conditions. Students were required to write a narrative response to one of 8 prompts (writing tasks), and a persuasive response to one of 2 prompts. Students in Years 5, 7 and 9, and the majority of students in Year 3, completed tasks online. In Year 3, 2 narrative prompts were administered on paper as well as online so that mode effect could be examined.

Table 19: Number of responses for writing by genre, mode, task and year level

Genre	Mode	Task	Year level				Total
			Y3	Y5	Y7	Y9	
Narrative	online	1	253	285	306	306	1,150
	paper	1	61	0	0	0	61
	online	2	257	286	306	312	1,161
	online	3	261	287	310	325	1,183
	online	4	245	277	301	302	1,125
	online	5	289	267	298	229	1,083
	paper	5	54	0	0	0	54
	online	6	288	270	300	226	1,084
	online	7	294	270	310	243	1,117
	online	8	284	255	299	231	1,069
Persuasive	online	9	216	237	258	280	991
	online	10	303	286	306	258	1,153
Total			2,805	2,720	2,994	2,712	11,231

Marking of writing

Pearson was contracted to develop marking materials and manage marking operations for the NAPLAN 2022 trial of writing tasks. Marking materials for training markers were developed by the contractor in collaboration with ACARA and a subgroup of the Marking Quality Team. A team of experienced NAPLAN markers was engaged by Pearson to mark the writing scripts remotely due to COVID restrictions in Victoria. ACARA's writing test manager supported Pearson's training of the markers and monitored the project carefully throughout the duration of marking.

The students' writing scripts from the trial were marked by these trained expert markers. The same quality assurance measures as those used in main study NAPLAN writing marking operations were implemented. Prompts that led to higher than usual discrepancies or difficulties were noted and data were analysed for abnormal patterns at the individual writing assessment criterion level.

After the marking of each prompt was completed, a debriefing session was held with markers. Qualitative feedback on the marking of each prompt was gathered to be used alongside the quantitative data when selecting prompts for the main study. This feedback included how successfully they perceived students had engaged with each task, marker fatigue concerns and any other difficulties encountered during marking.

Psychometric analysis of item trial data

The trial data were extracted from the assessment platform and then sent to the Australian Council for Educational Research (ACER) for analysis. Writing data was marked by another contractor and the marked data were sent to ACER for analysis.

Prior to data analysis, item response data was checked to confirm that the structure of the final data files was consistent with what was expected against the codebook and the trial test design. Records with all missing testlets were removed for non-writing domains, and records with raw score of zero were removed from the item calibration model for writing.

Item calibration and scaling was performed based on the Rasch model (Rasch 1960) using the software *ACER ConQuest 5* (Adams et al. 2022). The mathematical form of the model is provided in Chapter 6. For item calibration, embedded omits and the first of each sequence of trailing omits were treated as not-administered when estimating item difficulties to obtain an appropriate estimate of the item difficulty. However, these omits were treated as incorrect when estimating student abilities.

Numeracy, reading, spelling, and grammar and punctuation tests were calibrated separately by domain and year level, resulting in 16 separate calibrations in total. For each of the 4 non-writing online tests, items from all testlets within a domain and year level were calibrated in a concurrent analysis.

The writing test data from Years 3, 5, 7 and 9 were calibrated concurrently as some scores did not occur for some year levels. Due to the differences in marking rubric between persuasive tasks and narrative tasks, writing test data were calibrated separately by genre but concurrently for all the tasks in the same genre. The Rasch partial credit model (Rasch 1980, Masters 1982) was used for the calibration of writing. Three additional item response theory (IRT) models were used to review the properties of writing tasks in more detail: (1) task and year level effect model, (2) gender effect model and (3) Year 3 mode effect model.

After the calibration, trial items were reviewed in terms of their difficulty, discrimination and fit, and item characteristic curves (ICCs) showing score functioning were also examined. For the simple multiple-choice (MC) items, item category characteristic curves showing distractor functioning were examined. A sample item characteristic curve (ICC) for a well-performing item is presented in Figure 1. In this plot, student abilities are on the horizontal axis and the probabilities of correct responses (Proportion) are on the vertical axis. A sample category characteristic curves for a well-performing MC item is presented in Figure 2. In this plot, student abilities are on the horizontal axis and the probabilities of endorsing each response category (Proportion) are on the vertical axis.

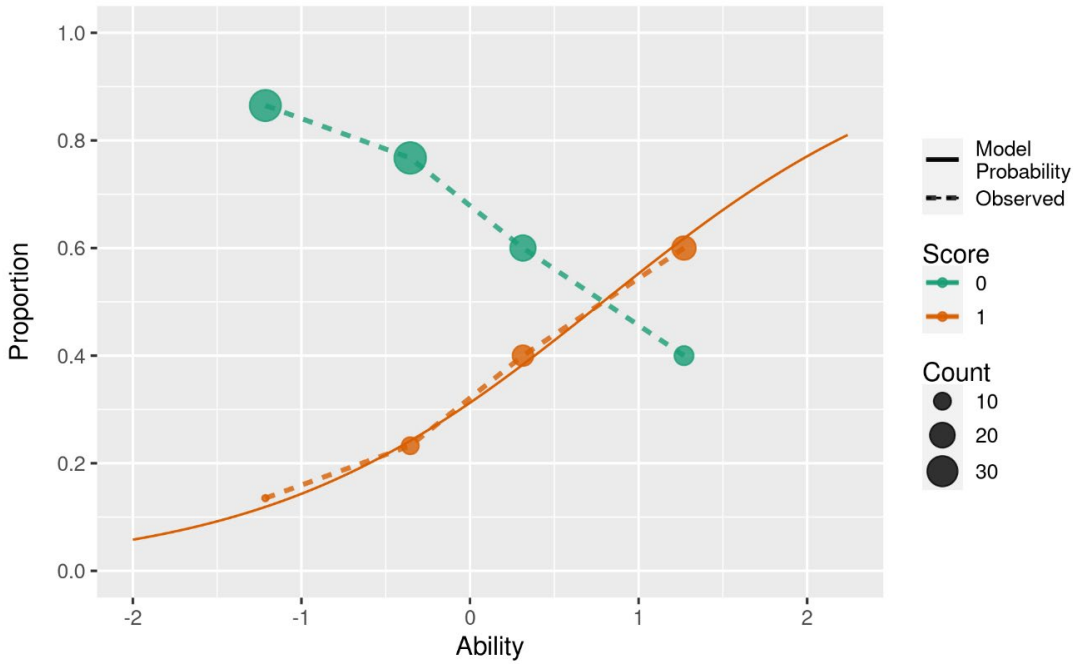


Figure 1: A sample ICC for a well-performing item

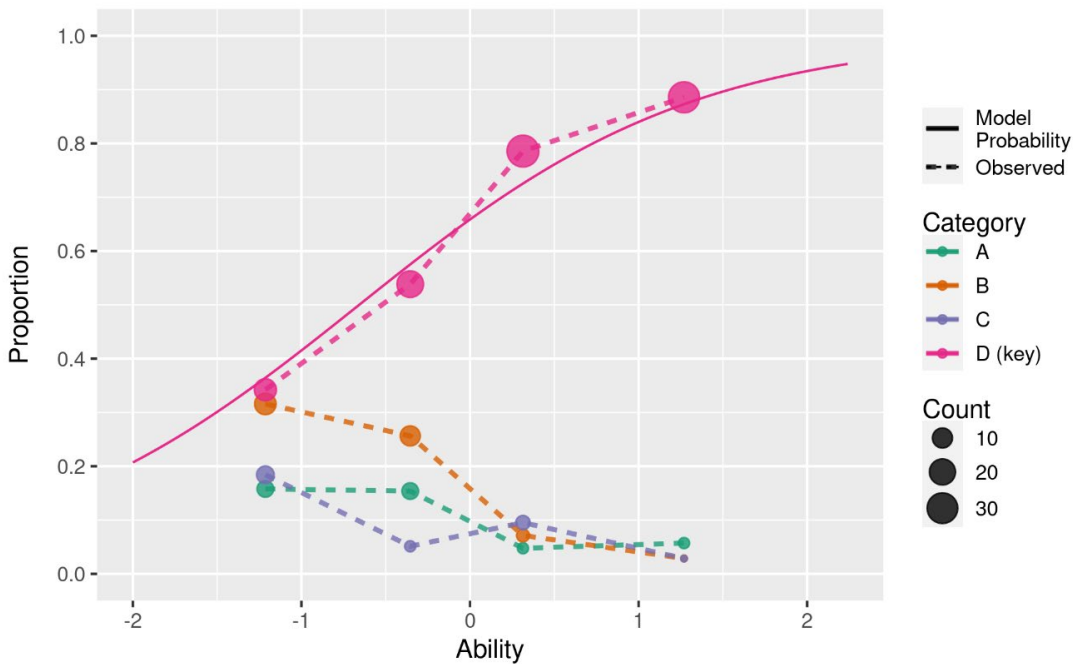


Figure 2: A sample item category characteristic curve for a well-performing MC item

In addition to the detailed item analyses listed above, a number of test-level metrics were summarised and examined. For each domain and year level, separate calibrations were carried out that can provide information on the targeting of test items to the ability distribution in each population. Reliability metrics (Cronbach's alpha and IRT-based reliabilities for different types of ability estimates) were also calculated for each test.

Differential item functioning (DIF) analyses on gender were performed on all trial items. Any item exhibiting a statistically significant difference in subgroup performance for students of the same

ability was flagged and subject to content analysis by test developers. More detailed descriptions of DIF are presented in Chapter 6.

To construct a common vertical scale for each non-writing domain, which included all items across all year levels, the year level trial tests were linked to each other by a set of *common items* between adjacent year levels. In addition, the trial tests were linked to the historical NAPLAN scale by a set of items used in previous NAPLAN main tests to align the scale with the NAPLAN historical scale. The quality and stability of the common items in terms of their functioning as equating links was systematically reviewed. More detail on the equating procedures is presented in Chapter 7.

Item selection for the 2022 NAPLAN tests

The results emerging from the psychometric analysis provided a pool of items for test managers to consider for inclusion in the final NAPLAN 2022 tests, alongside items from the existing NAPLAN item pool. Following evaluation of the psychometric properties of items from the item trial in 2021, statistics for reading and numeracy items trialled in Year 7 and Year 9 were deemed unreliable due to low response numbers. As a result, these items were subsequently excluded from the item pool. Furthermore, results obtained from DIF analysis enabled test managers to exclude those items that displayed bias against students of a particular gender. For the writing tests, the National Testing Working Group and the Marking Quality Team were provided with the relevant psychometric data on the trialled prompts and provided ACARA with advice on the final selection of prompts for each year level and the sequence in which they should be used for the 2022 tests.

Chapter 4: Test construction

The aim of this chapter is to describe the NAPLAN 2022 test construction and design. The first part of this chapter describes the test design for both online and paper tests. The branching methodology implemented in the NAPLAN multistage tailored test design is discussed in the second part.

Multistage, tailored test design

The NAPLAN online numeracy, reading and conventions of language assessments use a multistage tailored test design. A multistage tailored test is a type of Computerised Adaptive Test (CAT) with adaptivity taking place at the testlet level. A testlet is a small set of items that are administered together. Multistage tailored tests are considered a balanced compromise between non-adaptive paper-and-pencil and item-level adaptive tests (Hendrickson 2007).

Some benefits of tailored testing are:

- Tailored tests provide a more precise measurement of student performance. This allows for greater differentiation of students by using a wider range of questions at targeted difficulty, without adding to the length of the test for each individual student.
- Trials of the tailored test design show that students are more engaged with tests that adapt to their test performance. Students who experience difficulty early in the test are given some questions of lower complexity, more suited to their performance. These students are less likely to become discouraged as they progress through the tests. High-achieving students are given more challenging questions.
- The tailored test design has the potential to reduce anxiety in students who may find the historical paper-based format of NAPLAN too challenging.
- A wider range of aspects of the curriculum can be tested. While each student will answer the same number of questions as in the paper tests, the overall number of questions presented to students is larger.
- Tailored testing provides teachers and schools access to more targeted and detailed information on students' performance in online assessment.

The multistage tailored test design for numeracy, grammar and punctuation, and reading is illustrated in Figure 3. This figure shows a design with 6 nodes A, B, C, D, E and F. Each node comprises 3 testlets (e.g. A1, A2, A3), of which one is randomly allocated to the student. Each student completes 3 testlets in one of the following ordered combinations: ABC, ABE, ABF, ADC, ADE, ADF or ACB.

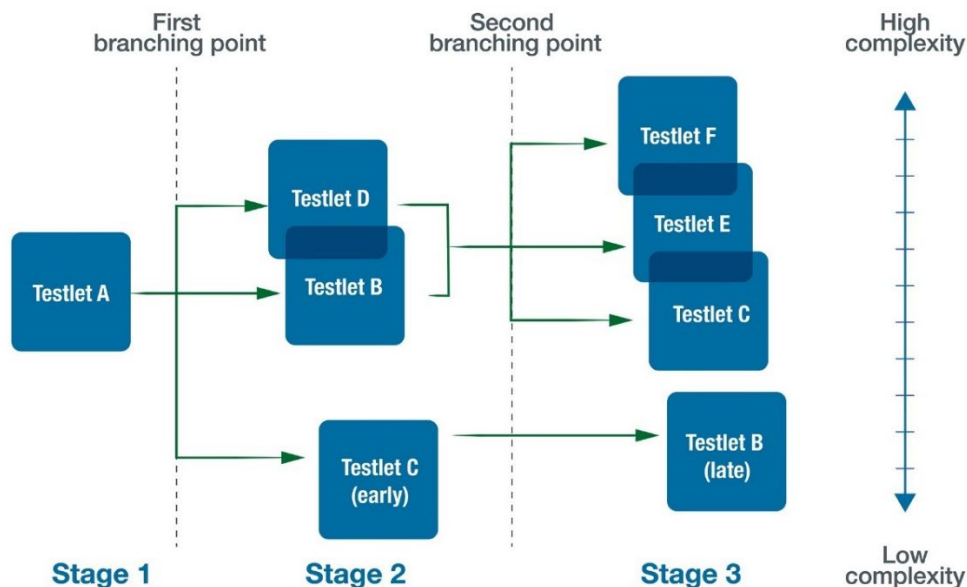


Figure 3: The multistage tailored test design for numeracy, reading and grammar and punctuation

Students at each year level start with testlet A. Each student's answers to testlet A determine the testlet they will be branched to and, as such, the questions they see. These may be less complex (B) or more complex (D). The student's answers in the first and second testlet determine branching to the final testlet: highest complexity (F), average complexity (E), lowest complexity (C). Students who receive a very low score for testlet A are branched directly to testlet C and then testlet B.

NAPLAN results for each student are based on both the number of the questions the student answers correctly and the average difficulty of the items that were assigned to the student. A student who completes a more complex set of questions is more likely to achieve a higher scale score (and a higher band placement), while a student who answers the same number of questions correctly, but follows a less complex pathway, will achieve a lower scale score.

The testlets within each node were designed with comparable item difficulties, curriculum coverage and skills assessed. This resulted in a minimum of 162 different test pathways that students could take, thus making it highly unlikely that 2 students sitting together in a classroom would be presented with the same items as each other.

The Year 7 and 9 numeracy test includes 2 sections in testlet A: non-calculator and calculator. An online calculator is available to students after they complete the non-calculator section of the test. Students were advised that they cannot return to the non-calculator section once they move to the calculator section.

The conventions of language (CoL) test includes a grammar and punctuation (G&P) section, and a spelling (Sp) section, each with 2 branching points. A message informs students that they cannot return to the G&P section once they move to spelling.

The grammar and punctuation section of the CoL test has the same multistage, multistream adaptive test design as numeracy and reading. The spelling test has a similar design but with only 2 testlets in the third stage (PD and PB). The graphical representation of the CoL test design is illustrated in Figure 4.

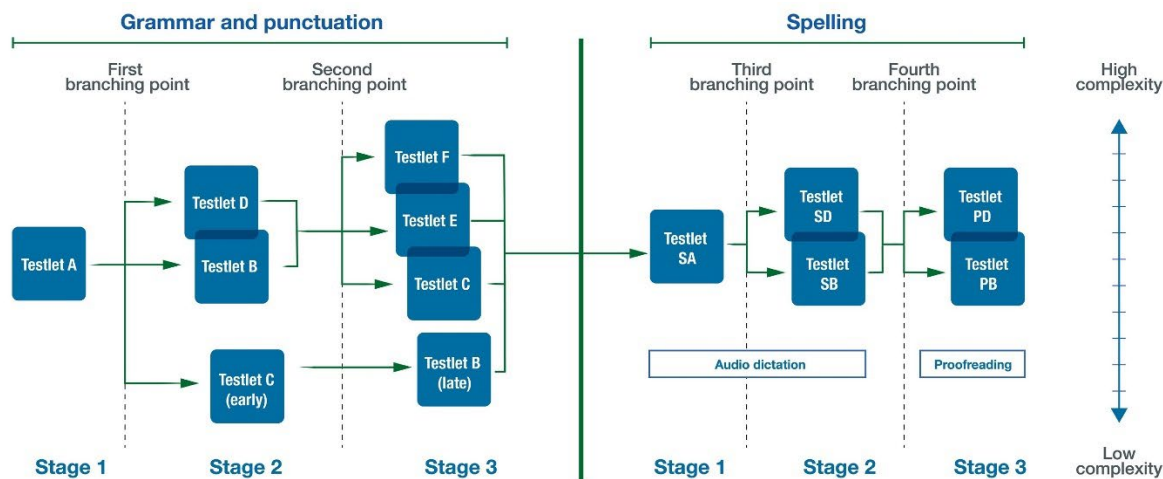


Figure 4: Online test design for conventions of language

As Figure 4 shows, the first 2 stages of the spelling section are focused on an audio component while the third stage is used to test proofreading. The spelling multistage design is discussed in more detail in the “Setting branching rules” section.

Construction of NAPLAN online tests

Data from the item trial and 2021 main study largely determined the placement of items within testlets. Skills, curriculum strands and proficiencies were balanced across nodes and testlets. When populating test designs, the choice and placement of link items were usually considered before other items, as they were vital to ensure comparability across vertical year levels and from calendar year to calendar year.

In considering link items, the guidelines shown below were followed:

- The weighted mean-square item fit must stay between 0.9 and 1.1.
- Items should not display the same gender DIF at 2 year levels.
- Item difficulty must be between -2 and 2 logits.
- The order of vertical links in both year levels should not change significantly, if at all.
- Horizontal links need to be placed as close as possible to the same position as in the 2019 main study (plus or minus 5).
- The items need to be representative of the balance of Australian Curriculum strands in the tests.

Test length

Table 20 to Table 22 outline the test lengths for each domain. The grammar and punctuation and spelling sections of the conventions of language tests are not delineated by year level as there were no differences in the specifications for each.

Table 20: NAPLAN online numeracy test: number of items and time available

Numeracy		Items per testlet	Total test items	Time available
Year 3		12	36	45 minutes
Year 5		14	42	50 minutes
Year 7	CA ¹	16 items x ½ testlet (8 items)	48	65 minutes
	NC ²	16 items x 2 ½ testlets (40 items)		
Year 9	CA	16 items x ½ testlet (8 items)	48	65 minutes
	NC	16 items x 2 ½ testlets (40 items)		

Calculators were not permitted in NAPLAN numeracy tests at Years 3 and 5. Calculators were also not permitted in the first half of testlet A in Years 7 and 9 but were permitted for the remainder of each of these tests.

Table 21: NAPLAN online reading test: number of items and time available

Reading	Items per testlet	Total test items	Time available
Year 3	13	39	45 minutes
Year 5	13	39	50 minutes
Year 7	16	48	65 minutes
Year 9	16	48	65 minutes

Table 22: NAPLAN online conventions of language test: number of items and time available

Conventions of language	Items per testlet	Item per section	Total test items	Time available
Grammar and punctuation	9	27	52	45 minutes
Spelling	7 items per testlet (audio dictation)	25		
	9 items per testlet (audio dictation)			
	9 items per testlet (proofreading)			

Difficulty of testlets

Items in each testlet were approximately uniformly distributed over the allowable logit range. For numeracy and conventions of language, items in each testlet were presented from least to most complex. For reading, in general, the unit³ with the lower average difficulty was presented first in each testlet and the unit with the higher average difficulty was presented last.

¹ CA – calculator-allowed

² NC – non-calculator

³ A reading unit comprises 1 stimulus text with 4-7 items related to that stimulus text.

Table 23 to Table 26 outline the predefined difficulty ranges in logits and average difficulty for the testlets in each test.

Table 23: NAPLAN online numeracy: predefined difficulty parameters for each testlet

Numeracy	Lower bound	Upper bound	Average
A	-3.0	1.0	-0.5
B	-2.0	0.5	-0.8
C	-3.5	-0.5	-2.0
D	-0.5	2.0	0.8
E	-1.5	1.5	0.0
F	0.5	3.5	1.4

Table 24: NAPLAN online reading: predefined difficulty parameters for each testlet

Reading	Lower bound	Upper bound	Average
A	-3.0	1.0	-1.0
B	-2.0	0.5	-0.8
C	-3.5	-0.5	-2.0
D	-0.5	2.0	0.8
E	-1.5	1.5	0.0
F	0.5	3.5	1.3

Table 25: NAPLAN online grammar and punctuation: predefined difficulty parameters for each testlet

Grammar and punctuation	Lower bound	Upper bound	Average
A	-3.0	1.0	-0.5
B	-2.0	0.5	-0.8
C	-3.5	-0.5	-2.0
D	-0.5	2.0	0.8
E	-1.5	1.5	0.0
F	0.5	3.0	1.25

Table 26: NAPLAN online spelling: predefined difficulty parameters for each testlet

Spelling	Lower bound	Upper bound	Average
SA	-4.0	2.0	-1.0
SB	-4.0	2.0	-0.8
SD	-3.0	3.0	0.8
PB	-5.0	2.0	-0.5
PD	0.0	5.0	1.0

Item types for online tests

The distribution of item types across the NAPLAN numeracy tests was nominally set at 50% multiple-choice(s) items, 20% text entry (constructed response) and 30% technology-enhanced items (TEI). The reading tests include multiple-choice(s) and technology-enhanced items only.

For the grammar and punctuation section of the conventions of language test, items were constructed either as multiple-choice(s) or TEI. In the spelling section, items were all text entry (constructed responses).

Table 27 to Table 29 show the final distribution of item types in the suite of items at each year level.

Table 27: NAPLAN online numeracy: item types in the item pool by year level

Numeracy	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in suite
Year 3	118	33	65	216
Year 5	130	47	75	252
Year 7	141	57	90	288
Year 9	168	47	73	288

Table 28: NAPLAN online reading: item types in the item pool by year level

Reading	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in suite
Year 3	192	-	42	234
Year 5	192	-	34	228
Year 7	254	-	34	288
Year 9	238	-	50	288

Table 29: NAPLAN online conventions of language: item types in the item pool by year level

Conventions of language	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in suite
Spelling Year 3	0	132	0	132
Spelling Year 5	0	132	0	132
Spelling Year 7	0	132	0	132
Spelling Year 9	0	132	0	132
G&P Year 3	100	0	116	216
G&P Year 5	115	0	101	216
G&P Year 7	88	0	128	216
G&P Year 9	88	0	128	216

Curriculum coverage

Items are written to cover the Australian Curriculum with a predefined balance of items from each strand across all year levels. This content coverage is the same for both the online and the paper tests.

For numeracy, the focus in Algebra is on pre-algebra concepts at Years 3, 5 and 7. At Year 9, after students have been introduced to variables in Year 7, the split between Algebra and Number is more pronounced.

For grammar and punctuation, the focus is predominantly on the sentence-level grammar, word-level grammar and punctuation sub-domains with a smaller focus on editing, text cohesion and vocabulary. Spelling items make up around half of a conventions of language test. Curriculum coverage is summarised in Table 30 to Table 41.

Table 30: NAPLAN numeracy Year 3 curriculum coverage by mode and pathway

Year 3	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	61%	54%	57%	58%	59%	57%
<i>Measurement and Geometry</i>	30%	28%	30%	28%	28%	28%	29%
<i>Statistics and Probability</i>	15%	11%	16%	15%	14%	13%	14%
Proficiencies							
<i>Fluency</i>	20%	17%	20%	22%	23%	21%	16%
<i>Understanding</i>	30%	31%	31%	43%	32%	25%	22%
<i>Problem-solving</i>	30%	33%	28%	20%	25%	32%	38%
<i>Reasoning</i>	20%	19%	21%	15%	19%	21%	25%
Item types							
<i>MC/MCs</i>	60%	72%	55%	57%	55%	50%	49%
<i>Text entry</i>	15%	28%	15%	11%	12%	17%	23%
<i>Interactive</i>	25%	-	30%	32%	33%	33%	28%

Table 31: NAPLAN numeracy Year 5 curriculum coverage by mode and pathway

Year 5	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	55%	55%	52%	54%	56%	56%
<i>Measurement and Geometry</i>	30%	29%	29%	30%	29%	30%	30%
<i>Statistics and Probability</i>	15%	17%	15%	17%	17%	14%	14%
Proficiencies							
<i>Fluency</i>	20%	19%	18%	19%	21%	20%	14%
<i>Understanding</i>	30%	29%	28%	33%	25%	25%	25%
<i>Problem-solving</i>	30%	29%	33%	31%	37%	37%	35%
<i>Reasoning</i>	20%	24%	20%	17%	17%	19%	25%

Year 5	Specified	Paper	Online	ABC	ABE	ADE	ADF
Item types							
<i>MC/MCs</i>	60%	71%	51%	54%	53%	51%	48%
<i>Text entry</i>	15%	29%	19%	16%	18%	20%	19%
<i>Interactive</i>	25%	-	30%	30%	29%	29%	33%

Table 32: NAPLAN numeracy Year 7 curriculum coverage by mode and pathway

Year 7	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	54%	56%	56%	56%	56%	56%
<i>Measurement and Geometry</i>	30%	29%	28%	30%	29%	29%	29%
<i>Statistics and Probability</i>	15%	17%	16%	15%	15%	15%	15%
Proficiencies							
<i>Fluency</i>	20%	21%	21%	22%	19%	21%	19%
<i>Understanding</i>	30%	29%	31%	33%	31%	26%	26%
<i>Problem-solving</i>	30%	29%	29%	29%	36%	35%	34%
<i>Reasoning</i>	20%	21%	19%	15%	15%	18%	22%
Item types							
<i>MC/MCs</i>	60%	68%	49%	51%	47%	48%	46%
<i>Text entry</i>	15%	31%	20%	22%	22%	19%	18%
<i>Interactive</i>	25%	-	31%	26%	31%	33%	36%

Table 33: NAPLAN numeracy Year 9 curriculum coverage by mode and pathway

Year 9	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	56%	51%	53%	53%	52%	52%
<i>Measurement and Geometry</i>	30%	29%	31%	31%	31%	31%	31%
<i>Statistics and Probability</i>	15%	15%	18%	17%	17%	17%	17%
Proficiencies							
<i>Fluency</i>	20%	19%	22%	24%	17%	23%	28%
<i>Understanding</i>	30%	33%	32%	38%	38%	31%	19%
<i>Problem-solving</i>	30%	29%	29%	23%	26%	28%	36%
<i>Reasoning</i>	20%	19%	17%	15%	19%	18%	17%
Item types							
<i>MC/MCs</i>	60%	73%	59%	57%	62%	59%	56%
<i>Text entry</i>	15%	28%	16%	21%	19%	26%	17%

Year 9	Specified	Paper	Online	ABC	ABE	ADE	ADF
<i>Interactive</i>	25%	-	25%	22%	19%	25%	27%

Table 34: NAPLAN reading Year 3 curriculum coverage by mode and pathway

Year 3	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Language</i>	5–15%	15%	16%	14%	15%	15%	15%
<i>Literature</i>	5–15%	10%	5%	2%	6%	9%	6%
<i>Literacy</i>	70–90%	74%	65%	85%	79%	77%	79%
Cognitive processes							
<i>Locating and identifying</i>	30–50%	41%	44%	61%	52%	39%	34%
<i>Integrating and interpreting</i>	30–50%	44%	47%	36%	42%	53%	50%
<i>Analysing and evaluating</i>	10–20%	15%	9%	3%	6%	8%	15%
Stimulus texts							
<i>Number of texts</i>		6	-	7	6	6	6
<i>Average word count</i>		178	155	90	145	174	195
Item types							
<i>MC</i>	90–100%	87%	76%	76%	81%	74%	72%
<i>MCs</i>	0–10%	5%	6%	4%	5%	9%	10%
<i>Other</i>	0–10%	8%	18%	20%	14%	17%	18%

Table 35: NAPLAN reading Year 5 curriculum coverage by mode and pathway

Year 5	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Language</i>	5–15%	10%	22%	23%	23%	26%	26%
<i>Literature</i>	5–15%	13%	11%	6%	9%	10%	12%
<i>Literacy</i>	70–90%	77%	70%	78%	75%	66%	65%
Cognitive processes							
<i>Locating and identifying</i>	30–50%	33%	24%	36%	31%	21%	17%
<i>Integrating and interpreting</i>	30–50%	46%	58%	58%	55%	54%	56%
<i>Analysing and evaluating</i>	10–20%	21%	19%	6%	15%	25%	27%
Stimulus texts							
<i>Number of texts</i>		6	-	6	6	6	6
<i>Average word count</i>		224	244	177	212	263	285
Item types							
<i>MC</i>	90–100%	90%	79%	74%	74%	80%	85%

Year 5		Specified	Paper	Online	ABC	ABE	ADE	ADF
	MCs	0–10%	8%	7%	9%	10%	9%	7%
	Other	0–10%	3%	15%	18%	15%	11%	9%

Table 36: NAPLAN reading Year 7 curriculum coverage by mode and pathway

Year 7		Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands								
	Language	10–20%	25%	22%	23%	23%	26%	26%
	Literature	10–20%	15%	11%	6%	9%	10%	12%
	Literacy	50–70%	60%	66%	71%	68%	64%	63%
Cognitive processes								
	Locating and identifying	20–40%	23%	24%	36%	31%	21%	17%
	Integrating and interpreting	40–60%	56%	58%	58%	55%	54%	56%
	Analysing and evaluating	20–40%	21%	19%	6%	15%	25%	27%
Stimulus texts								
	Number of texts		8	-	9	9	9	9
	Average word count		275	289	244	270	291	307
Item types								
	MC	90–100%	92%	84%	87%	86%	84%	82%
	MCs	0–10%	0%	4%	1%	2%	3%	6%
	Other	0–10%	8%	12%	12%	12%	13%	12%

Table 37: NAPLAN reading Year 9 curriculum coverage by mode and pathway

Year 9		Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands								
	Language	10–20%	23%	24%	22%	26%	25%	23%
	Literature	10–20%	15%	10%	3%	3%	10%	14%
	Literacy	50–70%	63%	66%	74%	70%	65%	63%
Cognitive processes								
	Locating and identifying	20–40%	29%	22%	28%	27%	22%	17%
	Integrating and interpreting	40–60%	50%	52%	56%	55%	53%	47%
	Analysing and evaluating	20–40%	21%	26%	15%	18%	24%	37%
Stimulus texts								
	Number of texts		8	-	9	9	9	9
	Average word count		310	308	259	294	308	338
Item types								

Year 9	Specified	Paper	Online	ABC	ABE	ADE	ADF
<i>MC</i>	90–100%	83%	76%	83%	81%	73%	69%
<i>MCs</i>	0–10%	4%	7%	5%	6%	8%	8%
<i>Other</i>	0–10%	13%	17%	13%	13%	19%	22%

Table 38: NAPLAN conventions of language Year 3 curriculum coverage by mode and pathway

Year 3	Spec.	Paper	Online	G&PA BC	G&P ABE	G&P ADE	G&P ADF	SASB PB	SASB PD	SASD PB	SASD PD
Australian Curriculum threads and test item formats											
<i>G&P grammar</i>	70%	72%	71%	75%	73%	69%	65%	-	-	-	-
<i>G&P punctuation</i>	30%	28%	29%	25%	27%	31%	35%	-	-	-	-
<i>Sp audio-dictation</i>	60%	0%	60%	-	-	-	-	61%	64%	61%	64%
<i>Sp mistake identified</i>	20%	48%	18%	-	-	-	-	20%	21%	20%	21%
<i>Sp mistake not identified</i>	20%	52%	22%	-	-	-	-	19%	15%	19%	15%
Australian Curriculum alignment to sub-domains											
<i>Editing</i>	-	-	2%	4%	5%	1%	-	-	-	-	-
<i>Punctuation</i>	-	14%	15%	25%	27%	31%	35%	-	-	-	-
<i>Sentence-level grammar</i>	-	10%	12%	25%	26%	27%	21%	-	-	-	-
<i>Text cohesion</i>	-	10%	8%	15%	7%	11%	17%	-	-	-	-
<i>Vocabulary</i>	-	-	4%	10%	10%	5%	5%	-	-	-	-
<i>Word-level grammar</i>	-	14%	12%	22%	25%	25%	22%	-	-	-	-
<i>Spelling</i>	50%	50%	48%	-	-	-	-	100%	100%	100%	100%
Item types											
<i>MC/MCs</i>	-	50%	24%	43%	46%	53%	57%	-	-	-	-
<i>Text entry</i>	-	50%	48%	-	-	-	-	100%	100%	100%	100%
<i>Interactive</i>	-	-	28%	57%	54%	47%	43%	-	-	-	-

Table 39: NAPLAN conventions of language Year 5 curriculum coverage by mode and pathway

Year 5	Spec.	Paper	Online	G&PA BC	G&P ABE	G&P ADE	G&P ADF	SASB PB	SASB PD	SASD PB	SASD PD
Australian Curriculum threads and test item formats											
<i>G&P grammar</i>	70%	68%	69%	67%	67%	65%	67%	-	-	-	-
<i>G&P punctuation</i>	30%	32%	31%	33%	33%	35%	33%	-	-	-	-
<i>Sp audio-dictation</i>	60%	-	55%	-	-	-	-	60%	60%	60%	60%
<i>Sp mistake identified</i>	20%	48%	20%	-	-	-	-	9%	23%	9%	23%
<i>Sp mistake not identified</i>	20%	52%	25%	-	-	-	-	31%	17%	31%	17%
Australian Curriculum alignment to sub-domains											
<i>Editing</i>	-	-	1%	4%	-	1%	1%	-	-	-	-
<i>Punctuation</i>	-	16%	16%	30%	30%	31%	31%	-	-	-	-
<i>Sentence-level grammar</i>	-	14%	16%	30%	35%	31%	36%	-	-	-	-
<i>Text cohesion</i>	-	8%	5%	9%	11%	11%	7%	-	-	-	-
<i>Vocabulary</i>	-	-	2%	9%	6%	4%	1%	-	-	-	-
<i>Word-level grammar</i>	-	12%	11%	20%	19%	22%	23%	-	-	-	-
<i>Spelling</i>	-	50%	49%	-	-	-	-	100%	100%	100%	100%
Item types											
<i>MC/MCs</i>	-	50%	27%	58%	56%	46%	45%	-	-	-	-
<i>Text entry</i>	-	50%	48%	-	-	-	-	100%	100%	100%	100%
<i>Interactive</i>	-	-	24%	42%	46%	54%	55%	-	-	-	-

Table 40: NAPLAN conventions of language Year 7 curriculum coverage by mode and pathway

Year 7	Spec.	Paper	Online	G&PA BC	G&P ABE	G&P ADE	G&P ADF	SASB PB	SASB PD	SASD PB	SASD PD
Australian Curriculum threads and test item formats											
<i>G&P grammar</i>	70%	68%	69%	70%	63%	64%	68%	-	-	-	-
<i>G&P punctuation</i>	30%	32%	31%	30%	37%	36%	32%	-	-	-	-
<i>Sp audio-dictation</i>	60%	-	55%	-	-	-	-	60%	60%	60%	60%
<i>Sp mistake identified</i>	20%	48%	18%	-	-	-	-	20%	12%	20%	12%
<i>Sp mistake not identified</i>	20%	52%	27%	-	-	-	-	20%	28%	20%	28%
Australian Curriculum alignment to sub-domains											
<i>Editing</i>	-	-	3%	-	-	5%	7%	-	-	-	-
<i>Punctuation</i>	-	18%	14%	30%	37%	33%	30%	-	-	-	-
<i>Sentence-level grammar</i>	-	16%	16%	31%	32%	31%	32%	-	-	-	-
<i>Text cohesion</i>	-	2%	5%	10%	9%	10%	9%	-	-	-	-
<i>Vocabulary</i>	-	-	1%	-	2%	4%	4%	-	-	-	-
<i>Word-level grammar</i>	-	14%	12%	30%	20%	17%	19%	-	-	-	-
<i>Spelling</i>	-	50%	49%	-	-	-	-	100%	100%	100%	100%
Item types											
<i>MC/MCs</i>	-	25%	21%	43%	41%	44%	40%	-	-	-	-
<i>Text entry</i>	-	25%	49%	-	-	-	-	100%	100%	100%	100%
<i>Interactive</i>	-	50%	30%	57%	59%	64%	60%	-	-	-	-

Table 41: NAPLAN conventions of language Year 9 curriculum coverage by mode and pathway

Year 9	Spec.	Paper	Online	G&PA BC	G&P ABE	G&P ADE	G&P ADF	SASB PB	SASB PD	SASD PB	SASD PD
Australian Curriculum threads and test item formats											
<i>G&P grammar</i>	70%	68%	72%	74%	72%	69%	67%	-	-	-	-
<i>G&P punctuation</i>	30%	32%	28%	26%	28%	31%	33%	-	-	-	-
<i>Sp audio-dictation</i>	60%	-	55%	-	-	-	-	60%	60%	60%	60%
<i>Sp mistake identified</i>	20%	48%	18%	-	-	-	-	24%	8%	24%	8%
<i>Sp mistake not identified</i>	20%	52%	27%	-	-	-	-	16%	32%	16%	32%
Australian Curriculum alignment to subdomains											
<i>Editing</i>	-	4%	1%	1%	8%	-	2%	-	-	-	-
<i>Punctuation</i>	-	18%	16%	27%	31%	31%	33%	-	-	-	-
<i>Sentence-level grammar</i>	-	8%	13%	22%	1%	28%	28%	-	-	-	-
<i>Text cohesion</i>	-	4%	5%	9%	28%	10%	11%	-	-	-	-
<i>Vocabulary</i>	-	4%	2%	2%	30%	5%	5%	-	-	-	-
<i>Word-level grammar</i>	-	12%	14%	33%	11%	26%	20%	-	-	-	-
<i>Spelling</i>	-	50%	49%	5%	2%	-	-	100%	100%	100%	100%
Item types											
<i>MC/MCs</i>	-	25%	21%	48%	42%	36%	40%	-	-	-	-
<i>Text entry</i>	-	25%	48%	-	-	-	-	100%	100%	100%	100%
<i>Interactive</i>	-	50%	31%	52%	54%	64%	60%	-	-	-	-

Paper test design

Four paper-based tests were administered at each of Years 3, 5, 7 and 9 as in previous cycles. The 4 tests were numeracy, reading, conventions of language (spelling, grammar and punctuation) and writing. All students who sat paper-based tests completed the same set of test items.

In numeracy, reading and conventions of language, there was a mix of multiple-choice (MC), multiple-choices (MCs) and constructed-response (CR) items. The MC and MCs items were presented in a standard format with a number of possible answers (usually between 4 and 6), from which students were required to select the best answer(s). The CR items generally required a numeric answer, a word or a short phrase. All items were dichotomously scored (correct or incorrect).

Items in all tests were distributed across the same difficulty range as the online tests. Specifically, the distribution of item difficulties in the paper test was approximately 20%, 30%, 30% and 20% across each quartile of the scale. Items were ordered approximately from easiest to hardest for numeracy, and within each section of the conventions of language tests. For reading, the average of each item set was used to arrange the units from easiest to hardest.

The use of calculators was not permitted in the numeracy tests in Year 3 and Year 5. For Year 7 and Year 9, calculator-allowed (CA) items preceded the non-calculator (NC) items.

Table 42: NAPLAN numeracy paper test number of items and time available

	Number of items		Time available	
Year 3	36		45 minutes	
Year 5	42		50 minutes	
Year 7 CA	8	48	10 minutes	65 minutes
Year 7 NC	40		55 minutes	
Year 9 CA	8	48	10 minutes	65 minutes
Year 9 NC	40		55 minutes	

Table 43: NAPLAN reading paper test number of items and time available

	Number of items		Time available	
Year 3	39		45 minutes	
Year 5	39		50 minutes	
Year 7	48		65 minutes	
Year 9	48		65 minutes	

Table 44: NAPLAN conventions of language paper test number of items and time available

	Number of items		Time available	
Year 3	25 spelling		45 minutes	
	25 grammar and punctuation			
Year 5	25 spelling		45 minutes	
	25 grammar and punctuation			
Year 7	25 spelling		45 minutes	
	25 grammar and punctuation			
Year 9	25 spelling		45 minutes	
	25 grammar and punctuation			

The numeracy, reading and conventions of language paper tests were created from a selected subset of online test items. Tables outlining test specifications encompassing average difficulty (logits), alignment to the Australian Curriculum and item types are included in Table 30 to Table 41.

Writing test design

The writing test covers the key writing aspects of the Australian Curriculum: English with a focus on accurate, fluent and purposeful writing of either a narrative or a persuasive text written in Standard Australian English.

Students are provided with a “writing stimulus” (sometimes called a prompt, task or topic) and instructed to write a response in a particular text type. To date, NAPLAN writing tests have required students to write in the narrative and persuasive genres. For NAPLAN 2022, all students were required to write a narrative text. Prior to the test, neither the students nor their teachers knew what the genre or topic would be. Students completed the writing test either on paper (handwritten) or online (typed). All Year 3 students completed their writing test on paper.

In 2022, 5 writing prompts were used across Years 3, 5, 7 and 9, and the paper and online modes. A further 3 prompts were kept in reserve in case of widespread technical issues or a security breach. No reserves were needed for 2022. Two of the 5 prompts were assigned to the Years 3 and 5 tests, and 3 to the Years 7 and 9 tests. The prompt that each student received depended on whether the test was taken on paper or online, and on which day of the writing test window the student sat the test (see Table 45). Each prompt has closely scripted scaffolding, or instructions. All prompts had been trialled and the prompts selected for the 2022 tests functioned similarly at the allocated year levels.

Table 45: NAPLAN writing prompt designation schedule according to test day

Writing prompt schedule					
	Day 1		Day 2	Day 3	Days 4–9
	Paper	Online	Online	Online	Online
Year 3	Prompt 1	N/A	N/A	N/A	N/A
Year 5	Prompt 1	Prompt 1	Prompt 3	Prompt 1 or 3 (rotational distribution)	Prompt 1 or 3 (rotational distribution)
Year 7	Prompt 2	N/A	Prompt 4	Prompt 5	Prompt 4 or 5 (rotational distribution)
Year 9	Prompt 2	N/A	Prompt 4	Prompt 5	Prompt 4 or 5 (rotational distribution)

All students were given 40 minutes to respond to the prompt. For the online tests, the timing commences before the students see or hear the prompt, whereas students doing the test on paper see the paper prompt and have it read to them immediately prior to the start of the test timer. Therefore, an additional 2 minutes is allocated to the online tests to allow students to read and/or listen to the audio recording of the prompt. It is recommended that students divide their time into 3 stages of writing: planning, writing and editing, although students can use their time as they choose.

Table 46: Recommended allocation of time for the writing test

Stage	Time available
Planning	5 minutes
Writing	30 minutes
Editing	5 minutes

The writing test targets the full range of student capabilities expected of students from Years 3 to 9. Year 3 and 5 students respond to the same prompts, and Year 7 and 9 students respond to the same prompts. The same marking guide is used from year to year to assess all students' writing, allowing for a national comparison of student writing capabilities across these year levels and over time.

The analytical, criterion-referenced marking guide consists of a rubric and exemplar scripts. The narrative rubric has 10 criteria and a total of 47 score points. In each criterion, each score category is cumulative and hierarchical. Each criterion is analysed as a polytomous item. The 10 criteria with the associated number of score categories are shown in Table 47 and Table 48.

Table 47: NAPLAN narrative marking criteria and skill focus descriptions

Criterion	Description of narrative writing marking criterion
Audience	The writer's capacity to orient, engage and affect the reader
Text structure	The organisation of narrative features including orientation, complication and resolution into an appropriate and effective text structure
Ideas	The creation, selection and crafting of ideas for a narrative
Character and setting	Character: The portrayal and development of character Setting: The development of a sense of place, time and atmosphere
Vocabulary	The range and precision of contextually appropriate language choices
Cohesion	The control of multiple threads and relationships across the text, achieved through the use of grammatical elements (referring words, text connectives, conjunctions) and lexical elements (substitutions, repetitions, word associations)
Paragraphing	The segmenting of text into paragraphs that assists the reader to negotiate the narrative
Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences
Punctuation	The use of correct and appropriate punctuation to aid the reading of the text
Spelling	The accuracy of spelling and the difficulty of the words used

Table 48: NAPLAN narrative marking criteria and skill focus descriptions

Item	Criterion	Score categories
1	Audience	0–6
2	Text structure	0–4
3	Ideas	0–5
4	Character and setting	0–4
5	Vocabulary	0–5
6	Cohesion	0–4
7	Paragraphing	0–2
8	Sentence structure	0–6
9	Punctuation	0–5
10	Spelling	0–6
Total raw score range		0–47

Writing marking training and quality assurance

Test administration authorities in each state and territory were responsible for marking student scripts from within their jurisdiction. Three jurisdictions – Qld, SA and WA – ran their own marking operations. ACT scripts were marked through the NSW marking operation, and Vic coordinated a marking operation for Vic, Tas and NT. In total there were over 1 million student scripts that needed to be marked nationally across the 5 marking operations. See Table 49 below.

Table 49: Writing scripts marked for each jurisdiction

	ACT	NSW	NT	Qld	SA	Tas	Vic	WA	Total
Y3	5430	93650	2770	62015	19675	6057	73858	33475	296930
Y5	5427	96509	2871	62894	20229	6226	74733	33360	302519
Y7	5516	93315	2587	61771	19739	6186	73416	32938	295467
Y9	4844	90401	1183	55719	91173	5971	70534	32271	281196
Total	21217	373875	9411	242399	150816	24440	292541	132044	1176112

Students' writing is marked by markers who are required to receive intensive training in the application of the 10 writing criteria. In 2022, 1883 markers were employed nationally (see Table 50). Most markers were practising or retired teachers. Markers were based in-centre or at home, depending on the operational needs of their local marking operation.

Table 50 shows the number of markers in each jurisdiction who participated in control script quality assurance processes during the marking, noting that the numbers of markers varied on any one day.

Table 50: Approximate number of NAPLAN writing markers per day by jurisdiction

	NSW, ACT	Qld	SA	Vic, NT, Tas	WA	Total
Number of markers	456	549	287	403	189	1883

To ensure national consistency across all marking operations, national protocols and comprehensive common training resources were delivered to each jurisdiction prior to marking, and quality assurance measures were implemented during the marking period. All markers across Australia used the same marking rubric, received the same training and were subject to comparable quality assurance measures.

Each marking operation ran for varying durations. The dates of commencement and conclusion were contingent on the number of scripts, the availability of the facilities for training and marking, the contractors' requirements and other factors. There was an overlap where all marking operations were running concurrently.

shows the commencement and conclusion dates of each primary operation and the total number of days each marking operation ran for, excluding "mop up" marking, which occurs in all operations.

Table 51: NAPLAN 2022 marking centre operational periods and duration by jurisdiction

	NSW, ACT	Qld	SA	Vic, NT, Tas	WA
Start of marking	16/05	23/05	23/05	16/05	19/05
Finish of marking	16/06	12/06	09/06	16/06	16/06
Days of marking	32	21	18	32	29

Nationally, all markers were trained with the same content and format to ensure continuity with previous years and consistency across jurisdictions. This was achieved through a number of different measures.

Intensive, detailed training was modelled to marking centre leaders and training staff in the form of a series of Centre Leader Training (CLT) workshops. These were conducted in the lead-up to the marking period and consisted of rigorous training in the writing criteria, effective marking methods and strategies for managing marking centres.

A comprehensive online Writing Marker Training course was also provided to test administration authorities (TAAs) for use in training new and experienced markers and leaders. The course was based on the face-to-face course used in previous years and delivered through a Learning Management System (LMS). Close to 1900 markers successfully completed the course nationally. Other resources provided for use in preparation for and during the marking period included slideshow presentations, exemplar training scripts and national marking protocols.

The core components of training and quality assurance materials were the pre-marked exemplar scripts with annotations called Training, Practice and Control (TPC) scripts. These scripts were originally selected from the pool of scripts from item trial, given individual marks by members of the Marking Quality Team¹ (MQT), then moderated to arrive at agreed consensus or “expert” scores for each criterion. Commentaries were then written for each script, explaining the category scores for each of the 10 criteria. Seventy-nine TPC scripts were developed in total, across the 5 prompts used for the 2022 tests. A subset of these scripts (Training and Practice) was used in the training of new and experienced markers and for “calibration” or “benchmarking” scripts to ensure comparability to the assigned expert score.

Table 52: The number of Training, Practice and Control scripts developed for each prompt

	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Total
Training	6	5	2	2	2	17
Practice	5	3	2	2	3	15
Control	13	12	3	3	1	32
Other	1	8	3	1	2	15
Total	25	28	10	8	8	79

Daily control scripts were used to monitor individual marker accuracy and collect data on the national consistency of marking. Each day of the marking period, control script data from each jurisdiction was provided directly to ACARA’s secure FTP site. This data was aggregated on a daily basis. A summary marking performance report for each control script was provided to each TAA so they could compare their jurisdiction’s marking accuracy for that control script with that of other jurisdictions. The first control script was issued when the first marking centre commences marking, and the last control was issued on the final day of the last marking centre. However, as each jurisdiction had a slightly different marking window, not all controls were completed by all centres.

In addition to control scripts, quality assurance through check-marking (sometimes referred to as double marking, spot checking or back-marking) was undertaken by marking centre leaders. Check-marking occurs for each marker and is done by a group leader, a centre leader or other experienced, expert marker appointed by the TAA responsible for the marking operation. Within each marking group or team, check-marking covered at least 10% of all scripts marked across the marking operation (although in some instances this was much higher than 20%).

¹ The MQT is made up of writing experts from each of the 10 jurisdictions, and is chaired by the manager of ACARA’s NAPLAN writing team.

Following administration of the national daily control scripts and implementation of local check-marking, jurisdictions used the data available to them in a range of reports. They used a variety of strategies to identify discrepant marking scores and marking patterns, and remediated scores as necessary. Centre leaders then had several courses of action that they could follow regarding the management of markers whose marking was discrepant, as required and informed by the national marking protocols (see Table 53 below).


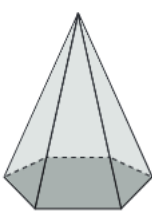
Table 53: National marking protocols

	Monitor	Discuss/ Re-train	Negotiate future marking
Total score	3–4 points discrepant	5–8 points discrepant	If 5 or more points discrepant on 3 occasions after retraining OR More than 8 points discrepant on 2 occasions
Criterion score	2 points discrepant	2 points discrepant on 3 or more occasions OR 3 or more points discrepant on 1 occasion	If 2 or more points discrepant on 3 occasions after retraining
General marking		Patterns in marking – repeated use of one score on any criterion OR Repeated score for many criteria	Unable to change poor marking after discussion/retraining

Example items in reporting bands

Table 54: Numeracy example items in reporting bands

Band	NAPLAN scale score	Item	Key / key string
1	270	<p>7 Kay has saved \$3247 for a holiday. She spends \$2000 on airfares. How much of her savings does Kay have left?</p> <p>\$5247 \$3227 \$3047 \$1247</p> <p><input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/></p>	D

Band	NAPLAN scale score	Item	Key / key string														
2	322	<p>6 Ning has this money in her money box.</p>  <p>In total, how much money does she have in her money box?</p> <p> <input type="radio"/> \$2.15 <input type="radio"/> \$6.10 <input type="radio"/> \$6.60 <input type="radio"/> \$7.10 </p>	D														
3	374	<p>14 The base of this pyramid is in the shape of a hexagon.</p>  <p>How many faces of the pyramid are triangles?</p> <p> <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7 </p>	D														
4	426	<p>19 This table shows the number of students who prefer different after-school activities.</p> <table border="1" data-bbox="534 1153 1165 1355"> <thead> <tr> <th rowspan="2">Activity</th> <th colspan="2">Number of students</th> </tr> <tr> <th>Girls</th> <th>Boys</th> </tr> </thead> <tbody> <tr> <td>Play computer games</td> <td>5</td> <td>3</td> </tr> <tr> <td>Play sport</td> <td>8</td> <td>10</td> </tr> <tr> <td>Read books</td> <td>4</td> <td>6</td> </tr> </tbody> </table> <p>How many more students prefer to read books than to play computer games?</p> <input type="text"/>	Activity	Number of students		Girls	Boys	Play computer games	5	3	Play sport	8	10	Read books	4	6	2
Activity	Number of students																
	Girls	Boys															
Play computer games	5	3															
Play sport	8	10															
Read books	4	6															
5	478	<p>14 Bindi takes the ferry from Darwin to Bathurst Island. She leaves Darwin at 11:15 in the morning and arrives at Bathurst Island at 1:45 in the afternoon.</p> <p>How long did Bindi take to get from Darwin to Bathurst Island?</p> <p> <input type="radio"/> 2 hours and 30 minutes <input type="radio"/> 2 hours and 45 minutes <input type="radio"/> 3 hours and 30 minutes <input type="radio"/> 3 hours and 45 minutes </p>	A														

Band	NAPLAN scale score	Item	Key / key string																																
6	530	<p>10 In Devonport, there are 30 604 people. Each day, the average person uses 173 litres of water. Which of these gives the best estimate for the total number of litres of water used in Devonport each day?</p> <p> <input type="radio"/> $30\,000 \times 200$ <input type="radio"/> $30\,000 \times 100$ <input type="radio"/> $30\,000 \div 200$ <input type="radio"/> $30\,000 \div 100$ </p>	A																																
7	582	<p>4 In 2017, workers at an office recorded the amount of paper they each recycled.</p> <ul style="list-style-type: none"> The office had 40 workers. Each worker recycled 50 kilograms of paper. Every 1000 kilograms of recycled paper saves 24 trees. <p>In total, how many trees did these workers save in 2017?</p> <input type="text"/>	48																																
8	634	<p>35 Students at a high school were surveyed to find whether they slept with a phone near their bed. The graph below shows the results.</p> <table border="1"> <caption>Data from the stacked bar chart</caption> <thead> <tr> <th>Age (years)</th> <th>No (%)</th> <th>Sometimes (%)</th> <th>Yes (%)</th> </tr> </thead> <tbody> <tr> <td>12</td> <td>50</td> <td>14</td> <td>36</td> </tr> <tr> <td>13</td> <td>45</td> <td>17</td> <td>38</td> </tr> <tr> <td>14</td> <td>52</td> <td>10</td> <td>38</td> </tr> <tr> <td>15</td> <td>30</td> <td>24</td> <td>46</td> </tr> <tr> <td>16</td> <td>30</td> <td>15</td> <td>55</td> </tr> <tr> <td>17</td> <td>24</td> <td>12</td> <td>64</td> </tr> <tr> <td>18</td> <td>17</td> <td>18</td> <td>65</td> </tr> </tbody> </table> <p>There were 150 12-year-old students at the high school. How many 12-year-old students responded 'No'?</p> <p> <input type="radio"/> 21 <input type="radio"/> 50 <input type="radio"/> 54 <input type="radio"/> 75 <input type="radio"/> 100 </p>	Age (years)	No (%)	Sometimes (%)	Yes (%)	12	50	14	36	13	45	17	38	14	52	10	38	15	30	24	46	16	30	15	55	17	24	12	64	18	17	18	65	D
Age (years)	No (%)	Sometimes (%)	Yes (%)																																
12	50	14	36																																
13	45	17	38																																
14	52	10	38																																
15	30	24	46																																
16	30	15	55																																
17	24	12	64																																
18	17	18	65																																
9	686	<p>33 At the entrance to a harbour there are two lights. A red light flashes every 5 seconds. A green light flashes every 7 seconds. The red light and the green light both flash together at 7:00 am. How many more times will the lights both flash at the same time in the next 3 minutes?</p> <input type="text"/>	5																																


Band	NAPLAN scale score	Item	Key / key string
10	738	<p>38 Suki makes a regular hexagon from six identical triangular tiles. Each tile has an area of 3.9 cm^2.</p>  <p>Suki then adds more tiles to make a hexagon with double the side length of this hexagon.</p> <p>What will be the area of this larger hexagon?</p> <p><input type="radio"/> 7.8 cm^2 <input type="radio"/> 23.4 cm^2 <input type="radio"/> 46.8 cm^2 <input type="radio"/> 93.6 cm^2</p>	D

Table 55: Reading example items in reporting bands

Dingle's game

Dingle needed a wash—not good news for Abbey and her brother Michael. Dingle was a big dog. A really big dog. His coat was shaggy and golden and his ears hung over his head like a pair of loose earmuffs. He always stood with his eyes bright and his legs ready to spring in any direction at any time—which he usually did.

The old iron wash tub was brimming with soapy water. It waited for Dingle on the one patch of green grass at the back of the house.

'I bags his front legs,' called Michael.

'All right, I'll take the back,' Abbey grudgingly agreed.

Abbey and Michael herded Dingle warily around the yard, steering him towards the small patch of lawn. A metre out, Michael took a chance and sprang towards Dingle. The big dog thought it was a great game and jumped in the opposite direction. Michael went down into a somersault before landing in a cloud of red dust. Abbey gave chase and Dingle let out a woof of delight. *This was fun.* Abbey ducked left and Dingle went right. Abbey ducked right and he went left. Then she just managed to scoop a hand under his collar and held on. It was a wild ride. She bounced across the yard as Dingle woofed again and took her in a wide circle around Mum's vegetable garden.

Dingle loved the game of chasey. He often played it with the hens or the sheep and sometimes with Mum's car coming up the drive, but now he was getting tired. As soon as Dingle (with Abbey attached) started to slow down, Michael was ready. He ran up behind Dingle and grabbed hold of the dog's haunches. That just seemed to give the massive hound a fresh burst of energy and he kept going, loving it all. Abbey and Michael, holding on tightly, heads down, didn't see what was coming.

When Dingle sailed over the tub, his hind legs kicked the surface of the water, and a wall of warm soapy spray lifted into the air and caught the sun. As the children swiped at the suds, they saw Dingle disappearing through the garden gate.



Band	NAPLAN scale score	Stimulus text	Item
3	343	Dingle's game	<p>Which word describes Dingle's size?</p> <p>That just seemed to give the massive hound a fresh burst of energy and he kept going, loving it all.</p>
4	394	Dingle's game	<p>The writer compares Dingle's ears to <i>loose earmuffs</i> to suggest that</p> <ul style="list-style-type: none"><input type="radio"/> Dingle cannot hear very well.<input type="radio"/> Dingle's ears are round.<input type="radio"/> Dingle's ears are very warm.<input checked="" type="radio"/> Dingle's ears are floppy.
5	462	Dingle's game	<p>This text is about</p> <ul style="list-style-type: none"><input type="radio"/> a very clean dog called Dingle.<input type="radio"/> how two children washed their pet dog.<input checked="" type="radio"/> a dog turning bath time into a game.<input type="radio"/> how you should wash your dog.

Band	NAPLAN scale score	Stimulus text	Item
6	497	Dingle's game	<p>Paragraph 1 suggests that Dingle</p> <ul style="list-style-type: none"><input type="radio"/> is too big to wash.<input checked="" type="radio"/> is difficult to wash.<input type="radio"/> has not been washed before.<input type="radio"/> is scared of being washed.
7	578	Dingle's game	<p>Why didn't Abbey and Michael see <i>what was coming?</i> (second last paragraph)</p> <ul style="list-style-type: none"><input type="radio"/> The sun was shining brightly in their eyes.<input type="radio"/> Dingle's head was blocking their view.<input checked="" type="radio"/> They were not looking where they were going.<input type="radio"/> Dingle made them dizzy.

A great southern secret—*two views*

View 1

Journeys not only take us out into the world; journeys inspire, delight and reawaken our souls. For a journey that will take you to a place of inspiring, awesome natural beauty without getting too far off the beaten track, go to where the Waychinicup River meets the Southern Ocean.

The name Waychinicup is loosely translated as ‘place where the emus came into being’. Although emus are no longer found in the area, it is not difficult to imagine the estuary as a place of creation. River and sea meet in an intense contrast; in the river mouth huge granite rocks, like broken giant’s teeth, are pounded by the Southern Ocean and through these the river is silently sieved out to sea.

The Waychinicup is one of the few rivers on the south coast not to have a sand bar, and on either side of the river the steep slopes are carpeted in thick impenetrable coastal scrub. Scattered across this carpet rear enormous, smooth, bone-coloured boulders, so inexplicably smooth, they are like finely carved sculptures. You cannot but suspect some earlier presence here. Who arranged these stones this way? Who smoothed them so? There is a large stone, sepulchral grey, with hundreds of smaller pink pebbles, flat and even as saucers, wedged into its side, keeping it vertical, forbidding it hurtling into the oblivion of black river water. And twin columns, like struts of an ancient altar, sit perfectly atop the skyline, looking down on the giant’s playground below.

View 2

Waychinicup is just a 50-minute trip from Albany. Head out on the road to Cheynes Beach for about 40 minutes and then onto a gravel road for 10 or so minutes, depending on how you and your car enjoy gravel corrugation. Every part of your load seems to challenge gravity on these corrugations before you arrive at a neat ring ‘road’ that has little tracks, like spokes on a wheel, radiating from it to numbered campsites. Apart from the tracks, an information board and a well-maintained bush toilet, there is really nothing else human-made that is permanently here.

Campers soon encounter the wildlife. Between June and October, whales calve close to shore and breaching whales are a common sight. Closer to camp, the brush-tailed possum is like the camp cat, roaming at will, but never too near. It will discover your rubbish bag wherever you put it. Quenda are far more shy, and seen only by the vigilant.

This is a place to experience uncomplicated life. There are no sounds except those of nature; no phones, televisions or internet pulling at your senses. Every day is a bad hair day, but you are oblivious because it is just you, the blue dome sky and an exceptional view. For a few days you feel like there are no other people on Earth.



Band	NAPLAN scale score	Stimulus text	Item
7	545	A great southern secret – two views	<p><i>Who arranged these stones this way? Who smoothed them so? (View 1)</i></p> <p>Why are these ideas expressed as questions?</p> <p><input type="radio"/> to introduce an explanation</p> <p><input checked="" type="radio"/> to produce a sense of wonder</p> <p><input type="radio"/> to outline areas for further investigation</p> <p><input type="radio"/> to question the importance of such matters</p>
8	589	A great southern secret – two views	<p>What do both views appeal to, in order to persuade the reader to visit Waychinicup?</p> <p><input type="radio"/> a sense of local pride</p> <p><input type="radio"/> an appreciation of history</p> <p><input type="radio"/> a love of camping</p> <p><input checked="" type="radio"/> a desire to escape ordinary life</p>
9	637	A great southern secret – two views	<p>Which comparison of View 1 and View 2 is the most accurate?</p> <p><input type="radio"/> View 1 is more detailed than View 2.</p> <p><input type="radio"/> View 1 is more humorous than View 2.</p> <p><input type="radio"/> View 2 is more biased than View 1.</p> <p><input checked="" type="radio"/> View 2 is more practical than View 1.</p>

Band	NAPLAN scale score	Stimulus text	Item
10	727	A great southern secret – two views	<p>In View 1, what is the main point of contrast between the river and the sea?</p> <p> <input checked="" type="radio"/> sound <input type="radio"/> depth <input type="radio"/> colour <input type="radio"/> beauty </p>


Table 56: Grammar and punctuation example items in reporting bands

Band	NAPLAN scale score	Item	Key / key string
1	215	<p>Place the correct word in the box to complete this sentence.</p> <p> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> </p> <p> <input type="text"/> or <input type="text"/> so <input type="text"/> for </p> <p>I like baking cakes <input type="text"/> I do not like cleaning up afterwards.</p>	but
2	283.1	<p>Place the correct ending in the box to complete this sentence.</p> <p> <input type="text"/> <input type="text"/> <input type="text"/> </p> <p> <input type="text"/> swimming with friends <input type="text"/> if she has time <input type="text"/> because it is hot </p> <p>Every day after school, <input type="text"/>.</p>	Jill helps her dad
3	328.8	<p>Choose the word that describes how the man walked.</p> <p> <input type="text"/> Slowly <input type="text"/> the <input type="text"/> old <input type="text"/> man walked <input type="text"/> down <input type="text"/> the hall and then <input type="text"/> wearily <input type="text"/> climbed into bed. </p>	Slowly

Band	NAPLAN scale score	Item	Key / key string
4	420	<p>Place the correct word in the box to complete this sentence.</p> <div style="border: 1px solid #ccc; padding: 5px; margin: 10px 0;"> hard hardly hardest </div> <p>It is harder to ride a horse than a bike.</p>	harder
5	458	<p>Place the correct word in the box to complete this sentence.</p> <div style="border: 1px solid #ccc; padding: 5px; margin: 10px 0;"> each much every </div> <p>The teacher asked how many parents would come to the concert.</p>	many
6	515	<p>Which of these sentences uses brackets correctly?</p> <ul style="list-style-type: none"> <input type="radio"/> My recipe for (pumpkin) soup uses 500 ml 2 cups of chicken stock. <input type="radio"/> My recipe for pumpkin soup uses (500 ml) 2 cups of chicken stock. <input type="radio"/> My recipe for pumpkin soup uses 500 ml 2 cups of (chicken) stock. <input checked="" type="radio"/> My recipe for pumpkin soup uses 500 ml (2 cups) of chicken stock. 	D
7	566	<p>Which is a complete sentence?</p> <ul style="list-style-type: none"> <input type="radio"/> Later, when we get the final numbers for the competition. <input checked="" type="radio"/> As Ben is coming too, I will make extra sandwiches. <input type="radio"/> Which I think is very interesting and helpful to us. <input type="radio"/> As they like going to the game and cheering on their team. 	B

Band	NAPLAN scale score	Item	Key / key string																									
8	618	<p>Choose one checkbox in each row of the table to show the correct word class for each word taken from this sentence.</p> <p>The chilly wind blows wildly.</p> <table border="1"> <thead> <tr> <th></th> <th>adverb</th> <th>adjective</th> <th>verb</th> <th>noun</th> </tr> </thead> <tbody> <tr> <td>chilly</td> <td><input type="checkbox"/></td> <td><input checked="" type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> <tr> <td>wind</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input checked="" type="checkbox"/></td> </tr> <tr> <td>blows</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input checked="" type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> <tr> <td>wildly</td> <td><input checked="" type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> </tbody> </table>		adverb	adjective	verb	noun	chilly	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	wind	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	blows	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	wildly	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
	adverb	adjective	verb	noun																								
chilly	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																								
wind	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>																								
blows	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>																								
wildly	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																								
9	655	<p>Place the correct punctuation mark in each sentence.</p> <p style="text-align: right;">: ;</p> <p>Rover lost his collar ; he was swimming in the dam.</p> <p>Our fitness has improved ; it has taken many hours of training.</p> <p>I have finally learnt the secret to success : believe in yourself.</p> <p>I love everything Dad cooks : steak, pizza and chicken pasta.</p>																										
10	731.2	<p>Which adverb in this sentence describes when an action happens?</p> <p>Henry arrived early for training, dropped his bag hurriedly and ran quickly to the oval where his coach was waiting patiently for the rest of the team.</p>	early																									

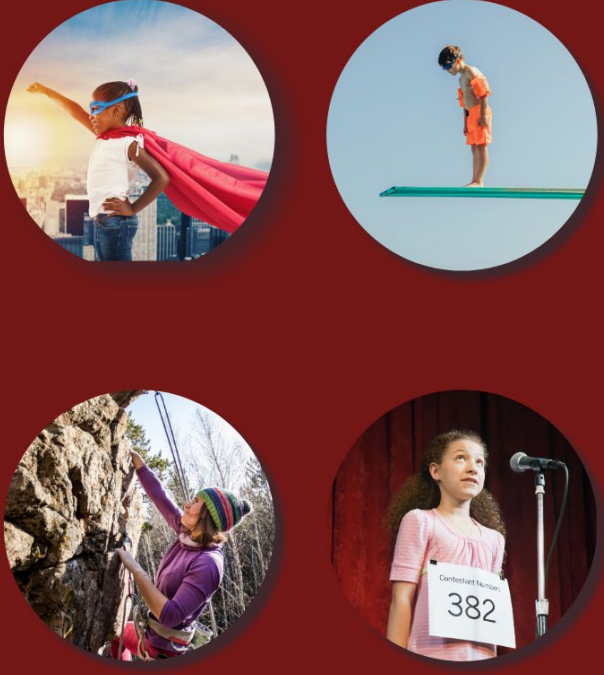
Table 57: Spelling items in bands

Band	NAPLAN scale score	Band	Key / key string
1	256.0	<p>They were giving out apples for _____.</p> <p>Click on the play button to hear the missing word.</p>  <p>Type the correct spelling of the word in the box.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">free</div>	free
2	325.7	<p>The spelling mistake in this sentence is underlined.</p> <p>The toy began to spinn around.</p> <p>Type the correct spelling of the underlined word.</p> <div style="border: 1px solid black; height: 40px; width: 100%; margin: 10px auto;"></div>	spin
3	362.6	<p>The spelling mistake in this sentence is underlined.</p> <p>He kickd the football through the goals.</p> <p>Type the correct spelling of the underlined word.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">kicked</div>	kicked

Band	NAPLAN scale score	Band	Key / key string
4	398.2	<p>The spelling mistake in this sentence is underlined.</p> <p>A dog is much <u>bigga</u> than a mouse.</p> <p>Type the correct spelling of the underlined word</p> <input type="text"/>	bigger
5	430.0	<p>The spelling mistake in this sentence is underlined.</p> <p>One <u>rool</u> in our class is to raise your hand to ask a question.</p> <p>Type the correct spelling of the underlined word in the rule</p> <input type="text"/>	rule
6	516.6	<p>The spelling mistake in this sentence is highlighted.</p> <p>The children saved the day and were heros.</p> <p>Type the correct spelling of the highlighted word</p> <input type="text"/>	heroes

Band	NAPLAN scale score	Band	Key / key string
7	534.3	<p>The spelling mistake in this sentence is highlighted.</p> <p>A rock band often has a <u>gitar</u> player.</p> <p>Type the correct spelling of the highlighted word in the box.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">guitar</div>	guitar
8	611.2	<p>There is one spelling mistake in this sentence.</p> <p>The students had a very <u>efficiant</u> method for completing their projects.</p> <p>Type the correct spelling of the word in the box.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">efficient</div>	efficient
9	654.6	<p>There is one spelling mistake in this sentence.</p> <p>The performance was given <u>spontaineous</u> applause.</p> <p>Type the correct spelling of the word in the box.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">spontaneous</div>	spontaneous
10	716.3	<p>The spelling mistake in this sentence is underlined.</p> <p>The mouse was a <u>nuisance</u> when it chewed through the electrical wires.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div style="border: 3px double black; padding: 5px; width: fit-content; margin: 10px auto;">nuisance</div>	nuisance

Table 58: Example writing prompt

<p style="text-align: center;">Brave</p> <p>Write a narrative (story) about a character who does something brave.</p> <p>Maybe the character helps someone or they do something that is hard but important to them</p> <p>You can use an idea on this page or your own idea about being brave.</p> <p>Think about:</p> <ul style="list-style-type: none"> • the characters and where they are • the complication or problem to be solved • how the story will end. <p>Remember to:</p> <ul style="list-style-type: none"> • plan your story before you start • choose your words carefully • write in sentences • pay attention to your spelling, punctuation and paragraphs • check and edit your writing. 	
--	--

Setting branching rules

In the NAPLAN online tailored tests, students are branched to easier or harder testlets, based on their number of correct responses on the previous testlet(s). Branching rules for sending students to testlets that are best matched to their ability level were determined before administration of the NAPLAN tests.

The branching method implemented in the NAPLAN multistage tailored test design was based on the Approximate Maximum Information (AMI) method (Leucht, Brumfield & Breithaupt 2006). In the AMI method, the intersection of the testlet information curves for the 2 adjacent testlets represents the branching cut-off. This approach is analogous to the maximum information item selection method in Computerised Adaptive Test (CAT) (Breithaupt & Hare 2007). The location of the intersection in logits (using estimated item difficulties from the item trial and previous NAPLAN assessments) was transformed into the number of correct responses using the test characteristic function. The final branching cut score was determined by truncating the result to an integer.

Adams and Lazendic (2013) showed that the AMI method provided effective and valid branching solutions for the NAPLAN online tailored test design. The AMI principle guided the development of the testlet targeting and boundaries, in addition to the decision regarding the ease of access condition that stipulated that testlet A must provide enough easy entry items to engage students at the lower end of the ability scale. NAPLAN tailored tests contained only 2 testlets in the second stage of the test (ignoring the option for students who failed to engage with the test to be routed to testlet C) and thus from the perspective of the AMI method, the ideal separation of the testlet information curves for testlets B and D would be a solution in which these 2 curves intersect at the point that will route 50% of students to each of these testlets, which was the mean of the student ability distribution.

However, the student ability and item difficulty means are not always aligned; therefore, in translating the intersection of the test information curves on to the student ability scale, care was taken to account for such mistargeting. The investigation showed that the empirical distributions

of the ability estimates did not differ significantly across year level and domains, when the measurement scale was case-centred within year level (that is, when the mean of student ability was set to zero). Consequently, the same set of item difficulty estimates for NAPLAN online testlets could be used across year levels for the grammar and punctuation, numeracy and reading domains. The final testlet boundaries and parameters were developed and empirically investigated in a series of simulations to establish the feasibility and robustness of overall NAPLAN online test parameters for reading and numeracy tests.

Domain specific branching rules are discussed in the remaining of this section.

Branching rules for numeracy, reading, and grammar and punctuation tests

Figure 3 illustrates a 3-stage tailored test design (1–2–3) with one node (A) in Stage 1; 2 nodes (B and D) in Stage 2; and 3 nodes (C, E and F) in Stage 3. These 6 testlets form 7 pathways (ABC, ABE, ABF, ADC, ADE, ADF and ACB), which are shown in Figure 3.

All students at each year level and domain started with a testlet in node A (Stage 1). Once this testlet was completed, a decision was made to branch a student to either an easier testlet (node B) or a harder testlet (node D), which was the *first branching point*. Assuming that a student was sent to a testlet in node D and completed this testlet, then another decision was made to branch this student to a testlet in node C (low complexity items), a testlet in node E (items with average complexity) or a testlet in node F (high complexity items), which was the *second branching point*. If a student was branched to node E, pathway ADE (shown in Figure 3) was completed. As discussed earlier, students with very low performance on a testlet in node A were first assigned the easiest testlet in node C as a second testlet before finally being assigned testlet B as the third testlet (pathway ACB). This allowed low-performing students to demonstrate their knowledge with items that matched their test performance and to engage more efficiently through the test.

A rational approach to setting these branching rules was to use the test information function (Lord and Novick 1968). The test information function describes the level of precision that a test can provide at each level of ability.

The information functions for testlets in nodes C, B and D are illustrated in Figure 5. As this figure shows, the peak of the information function for testlets in nodes B and D was about -1 and 1 logits, respectively. This means that the items were allocated to B and D so that D was more suited to more able students and B was more suited to less able students. In fact, given that the curves intersect at about 0.0 logits, these information functions show that if a student's ability was below 0.0 logits, then testlet B was expected to work best for them; whereas if a student's ability was above 0.0 logits, then testlet D was expected to work best for them. Similarly, this figure shows that testlet C (green curve) provides more information for students with an ability less than -1.5 logits. Given that the testlets C and B curves intersect at about -1.6 logits, if a student's ability was below -1.6 logits, then testlet C was expected to work best for that student.

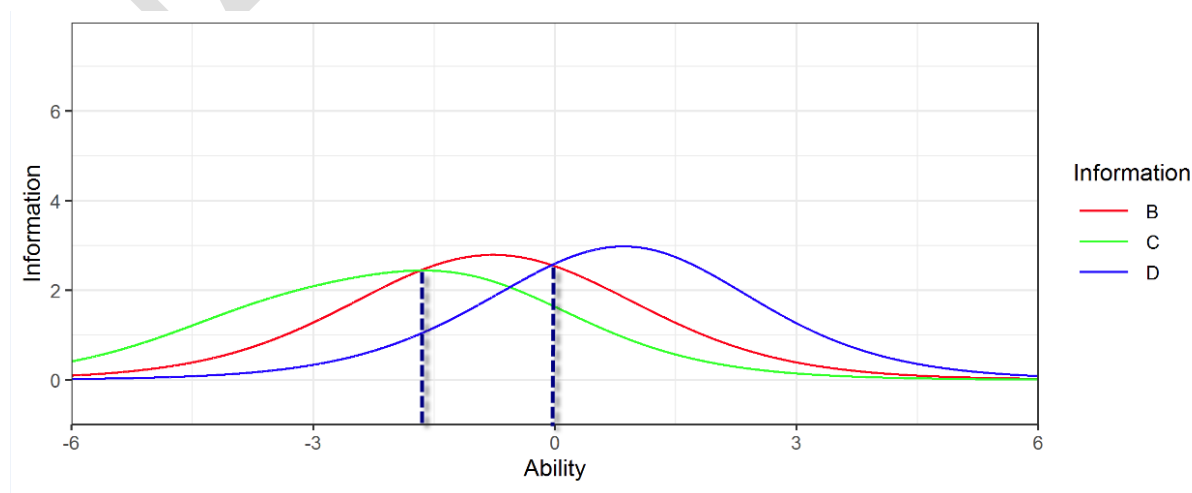


Figure 5: Test information functions: curves for testlets C, B and D

Once suitability of each testlet to students' ability was known, the location of the intersections in logits could be transformed into a raw score, or the number of correct responses on the previous testlet(s).

Figure 6 illustrates how the test characteristic curve for one testlet (in node A) can be used to find the raw scores that correspond to the cut-points between testlet information functions. The test characteristic curve for testlet A is shown on the same axis as the information functions for testlets C, B and D. If a student has a raw score of 4 or less on testlet A, then their ability estimate is in a region for which testlet C provides most precision; whereas if a student has a raw score greater than 4 and less than 9 on testlet A, then their ability estimate is in a region for which testlet B provides most precision. Similarly, students with a raw score of 9 or more will be assigned testlet D, which provides most precision.

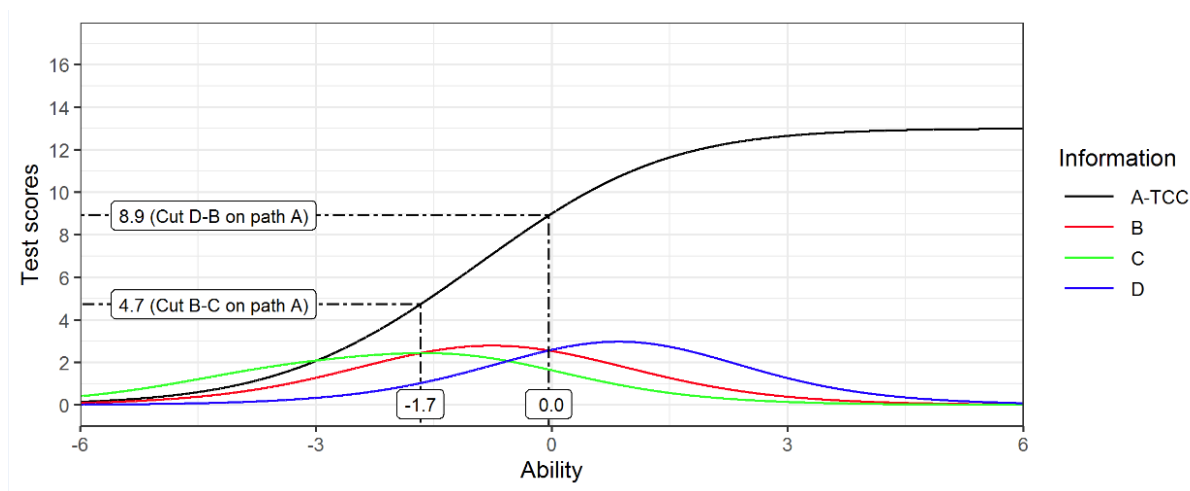


Figure 6: Stage 1 testlet A-C|B|D cut scores

The branching rules for the first branching point discussed above are presented in Table 59.

Table 59: Stage 1 cut scores (testlet A to C|B|D)

Rule Id	Lower Bound (inc)	Upper Bound (inc)	Logit	RawScore
AC	0	4	-1.673	4.740
AB	5	8	-0.040	8.914
AD	9	13	6.000	13.000

The same approach was taken to set the rules (cut scores) for the second branching point (Figure 7 and Table 60).

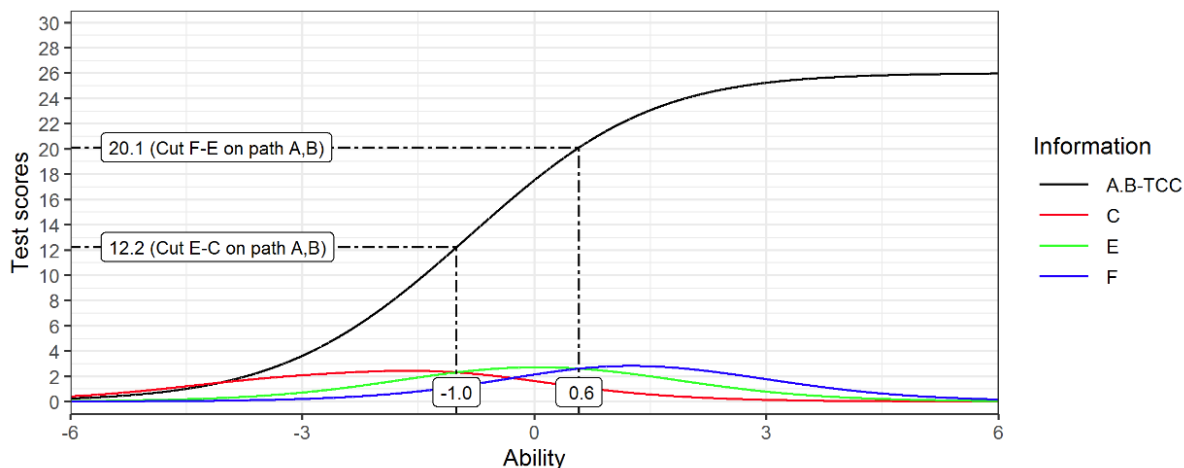


Figure 7: Stage 2 testlet AB–C|E|F cut scores

In Figure 7, the test characteristics curve for testlet AB is shown on the same axis as the information functions for testlets C, E and F. If a student had a cumulative raw score of 12 or less on testlets A and B, then their ability estimate was in a region for which testlet C provided most precision; whereas if a student had a cumulative raw score greater than 12 but less than 21 on testlets A and B, then their ability estimate was in a region for which testlet E provided most precision. Finally, students with a cumulative raw score of 21 or more were assigned Testlet F, which was designed for high-performing students. The branching rules for the second branching point after students completed testlets A and B are presented in Table 60.

Table 60: Stage 2 cut scores (testlet AB to C|E|F)

Rule Id	Lower Bound (inc)	Upper Bound (inc)	Logit	RawScore
ABC	0	12	-1.007	12.245
ABE	13	20	0.577	20.125
ABF	21	26	6.000	26.000

In Figure 8, the test characteristics curve for testlet AD is shown on the same axis as the information functions for testlets C, E and F. If a student had a cumulative raw score of 8 or less on testlets A and D, then their ability estimate was in a region for which testlet C provided most precision; whereas if a student had a cumulative raw score greater than 8 but less than 17 on testlets A and D, then their ability estimate was in a region for which Testlet E provided most precision. Finally, students with a cumulative raw score of 17 or more were assigned Testlet F, which contained the most challenging items. The branching rules for the second branching point after students completed testlets A and D are presented in Table 61.

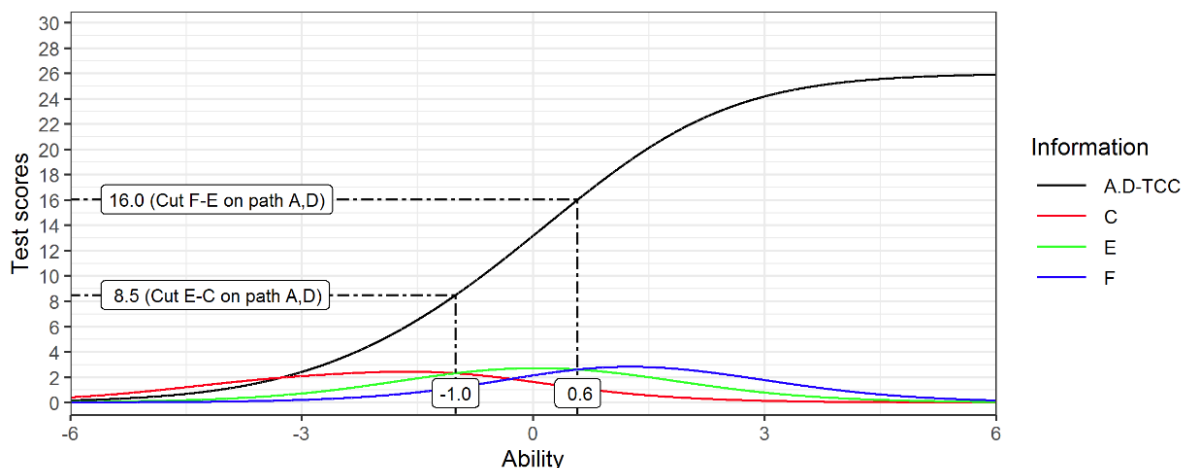


Figure 8: Stage 2 testlet AD-C|E|F cut scores

Table 61: Stage 2 cut scores (testlet AD-C|E|F)

Rule Id	Lower Bound (inc)	Upper Bound (inc)	Logit	RawScore
ADC	0	8	-1.007	8.504
ADE	9	16	0.577	16.044
ADF	17	26	6.000	26.000

Branching rules for spelling

The right-hand side of Figure 4 illustrates a 3-stage tailored test design (1-2-2) for spelling with one testlet in Stage 1, 2 testlets in Stage 2 and 2 testlets in Stage 3. These 5 testlets formed 4 pathways (SA-SD-PD, SA-SD-PB, SA-SB-PD, SA-SB-PB).

As in the numeracy, reading, and grammar and punctuation tailored test design, every student started with testlet SA (Stage 1). Once testlet SA was completed, a decision was made to branch a student to either an easier testlet SB or a harder testlet SD, which was the *first branching point*. If a student was sent to testlet SD and completed this testlet, then another decision was made to branch this student to testlet PB (low complexity items), or testlet PD (high complexity items), which was the *second branching point*. If a student was branched to testlet PD, pathway SA-SD-PD was completed.

Figure 9 shows that 2 decisions were made before branching students to the final stage in the multistage tailored tests: 1) after completion of testlet SA, and 2) after completion of testlets SA-SB or SA-SD. These decisions were made before the multistage test was administered. The same rationale, applied to setting branching rules for reading and numeracy tests, was utilised in spelling. The branching rules for spelling are illustrated in Figure 9, Figure 10 and Figure 11.

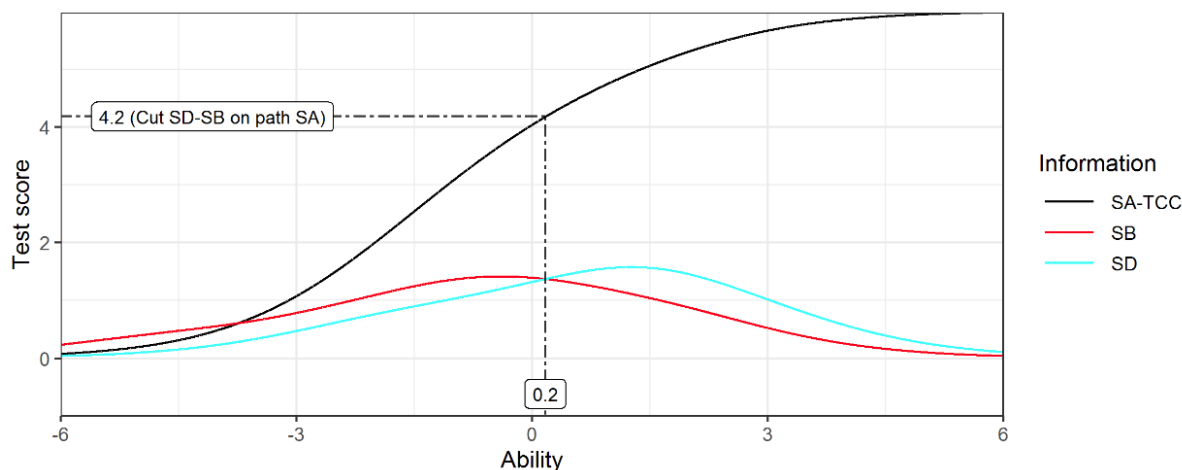


Figure 9: Stage 1 testlet SA-SB/SD cut scores

In Figure 9, the test characteristics curve for testlet SA is shown on the same axis as the information functions for testlets SB and SD. If a student had a raw score of 4 or less on testlet SA, then their ability estimate was in a region for which testlet SB provided most precision; whereas if a student had a raw score greater than 4 on testlet SA, then their ability estimate was in a region for which testlet SD provided most precision. The branching rules for the first branching point in spelling are presented in Table 62.

Table 62: Stage 1, testlet SA-SB/SD cut scores

Rule Id	Lower Bound (inc)	Upper Bound (inc)	Logit	RawScore
SASB	0	4	0.168	4.175
SASD	5	6	6.000	6.000

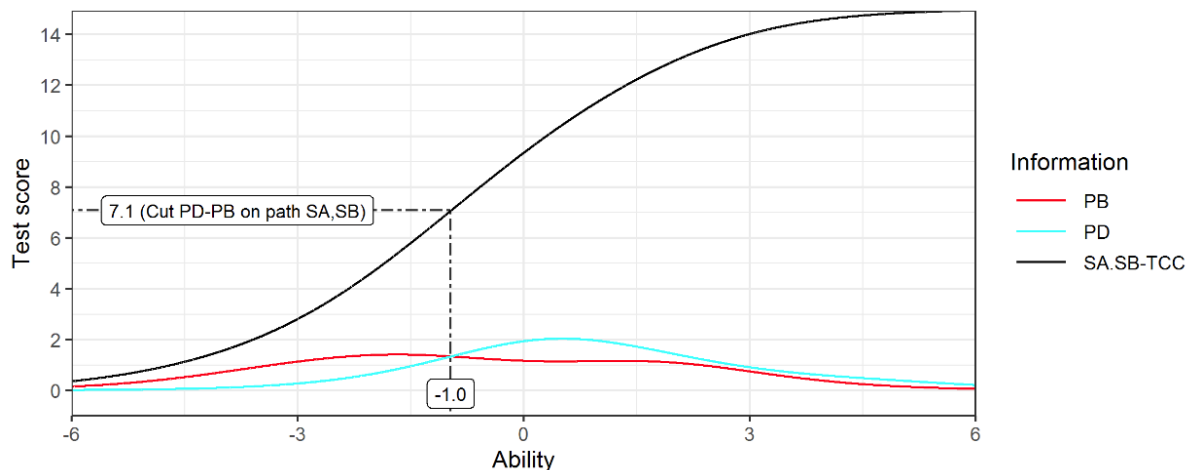


Figure 10: Stage 2 testlet SA-SB to PB/PD cut scores

In Figure 10, the test characteristics curve for testlet SA-SB is shown on the same axis as the information functions for testlets PB and PD. If a student had a cumulative raw score of 7 or less on testlets SA and SB, then their ability estimate was in a region for which testlet PB provided most precision; whereas if a student had a cumulative raw score greater than 7 on testlets SA and SB, then their ability estimate was in a region for which testlet PD provided most precision. The branching rules for the second branching point in spelling are presented in Table 63.

Table 63: Stage 2, testlets SA–SB to PB|PD cut scores

Rule Id	Lower Bound (inc)	Upper Bound (inc)	Logit	RawScore
SASBPB	0	7	-0.965	7.076
SASBPD	8	15	6.000	15.000

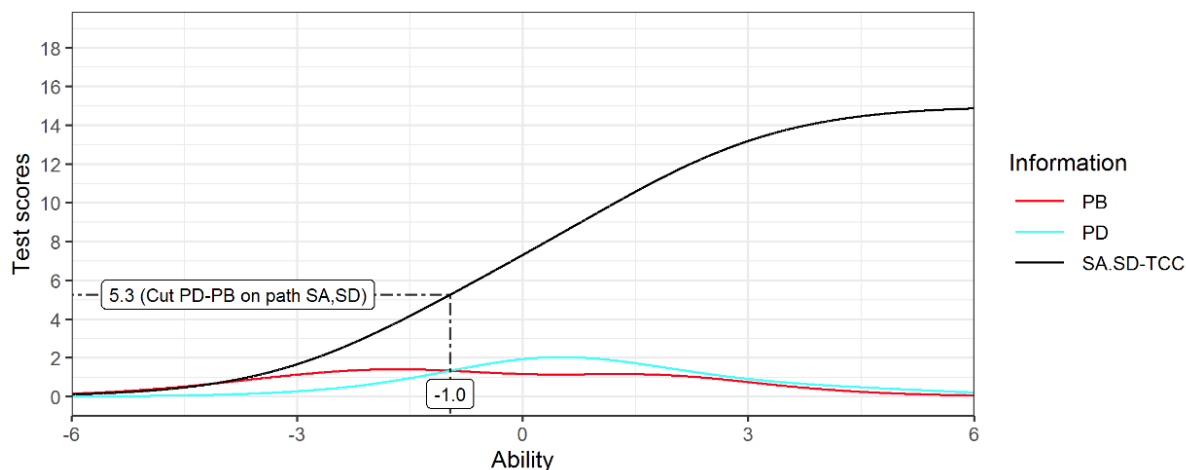


Figure 11: Stage 2 testlets SA–SD to PB|PD cut scores

In Figure 11, the test characteristics curve for testlet SA–SD is shown on the same axis as the information functions for testlets PB and PD. If a student has a cumulative raw score of 5 or less on testlets SA and SD, then their ability estimate is in a region for which testlet PB provides more precision; whereas if a student has a cumulative raw score greater than 5 on testlets SA and SD, then their ability estimate is in a region for which testlet PD provides more precision. The branching rules for the second branching point in spelling are presented in Table 64.

Table 64: Stage 2, testlet SA–SD to PB|PD cut scores

Rule Id	Lower Bound (inc)	Upper Bound (inc)	Logit	RawScore
SASDPB	0	5	-0.965	5.27
SASDPD	6	15	6.000	15.00

Pathway utilisation

This section describes how different pathways were utilised in NAPLAN 2022 online tests, using Year 3 numeracy as an example. The results for other year levels and domains are presented in Appendix A.

The percentage of students assigned to each pathway, and ability distributions at each stage for Year 3 numeracy are shown in Figure 12 and Figure 13.

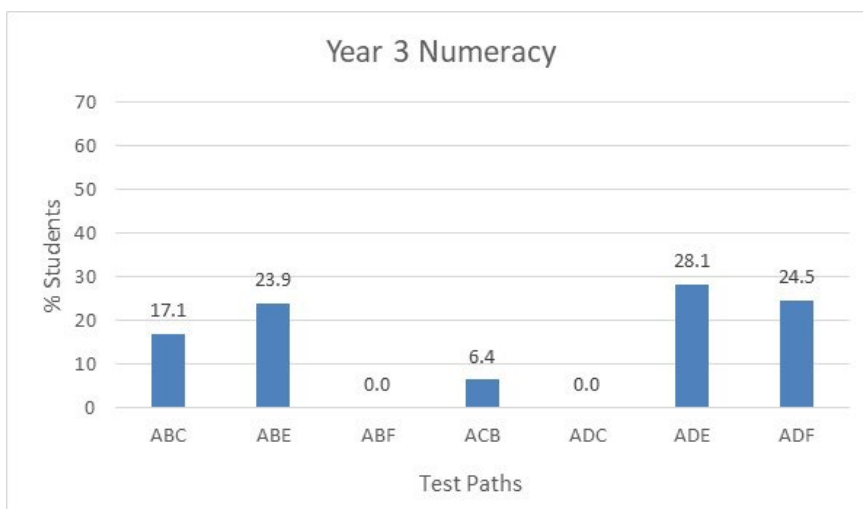


Figure 12: Percentage of students assigned to each pathway in Year 3 numeracy

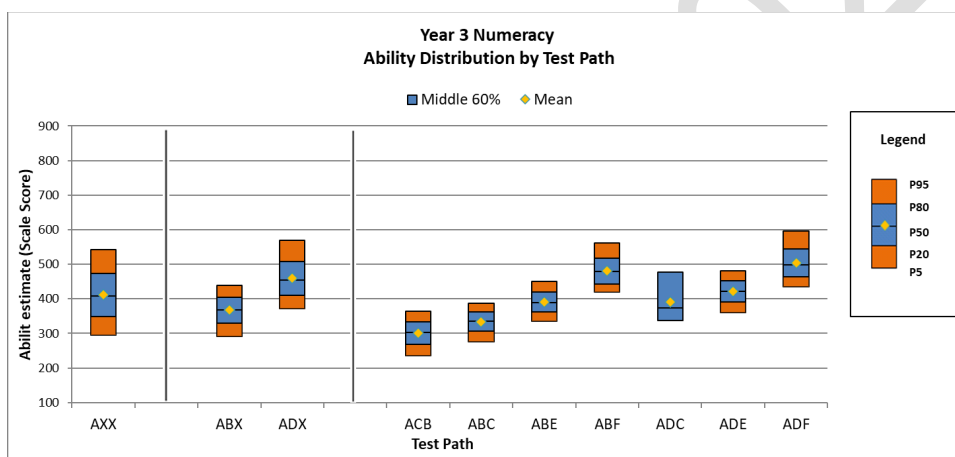


Figure 13: Ability distribution by pathway for Year 3 numeracy

As Figure 12 shows, the ideal separation of the testlet information curves for testlets B¹ and D has been achieved, so approximately 50% of students have been sent to each of these testlets. The number of students assigned to each path varied from 0% for ADC and ABF pathways to approximately 28% in ADE pathway. To some extent, the very low proportions in the ADC and ABF pathways were expected since, for example, going through the ADC pathway would require high performance on testlet A followed by very poor performance on testlet D. Similarly, a very low percentage (0.0) for ABF pathway was expected since it would require low performance on testlet A followed by high performance on testlet B. This chart also shows approximately 6.5% of students were sent to Testlet C immediately after completing Testlet A.

Ability distributions by pathway are illustrated in Figure 13. Patterns of ability distributions across pathways were roughly as expected. That is, students ending in testlet F had the highest ability distribution and students who were administered testlet C immediately after completing Testlet A (ACB) had the lowest ability distributions. Furthermore, the ability distribution in the second stage shows that, to a large degree, high- and low-performing students were sent to testlet D and testlet B, respectively. Figure 13 also shows that pathways overlapped in abilities.

¹ B testlets include pathways ABC, ABE and ACB.

Chapter 5: Data collection and preparation

This chapter describes data collection and delivery, data validation and data preparation for NAPLAN 2022. The first part of the chapter focuses on how data for paper and online tests are collected by test administration authorities (TAAs) from each jurisdiction and delivered to ACARA. The second part of the chapter describes how data are validated and prepared by the contractor before performing the analysis.

Data collection and delivery

TAAs are responsible for:

1. implementing and administering the NAPLAN tests in their jurisdiction, following “National protocols for test administration” provided by ACARA
2. collecting NAPLAN test and student background data in their jurisdiction, performing quality assurance on data before providing it to ACARA. ACARA then performs quality assurance on the final data received from each jurisdiction.

Student background data plays an important role in different phases of NAPLAN analysis. Therefore, it is especially important for schools and school systems to collect this information in a consistent way.

The purpose of the Data Standards Manual: Student Background Characteristics¹ is to provide guidance to schools and school systems in the collection of information on student background characteristics, using the nationally agreed standard measures of the characteristics. The manual is to be used by schools and school systems when enrolling students for the first time in the school year, or when collecting information, via special data collection forms, on those students participating in national assessments.

The nationally agreed student background characteristics collected are:

- gender
- Indigenous status
- parental occupation and education
- language background other than English (LBOTE).

Test response data were delivered to ACER in 4 main batches:

- staggered delivery of online test data including both scored and raw response data (used for item calibration)
- delivery of the second version of the Student Master File (SMF), online Writing Scores File (WSF) and Item Response File (paper data for those jurisdictions that sat NAPLAN tests on paper and online)
- delivery of the third version of the SMF, IRF, WSF and online test data (NAEs), previously called Stage 1 census data, for analysis to produce the NAPLAN 2022 summary results
- delivery of the final SMF / IRF / WSF / NAEs, previously called Stage 2 complete census data, to produce the NAPLAN 2022 National Report.

NAPLAN 2022 Stage 1 and Stage 2 data flow are shown in Figure 14 and Figure 15.

¹ www.acara.edu.au/reporting/data-standards-manual-student-background-characteristics

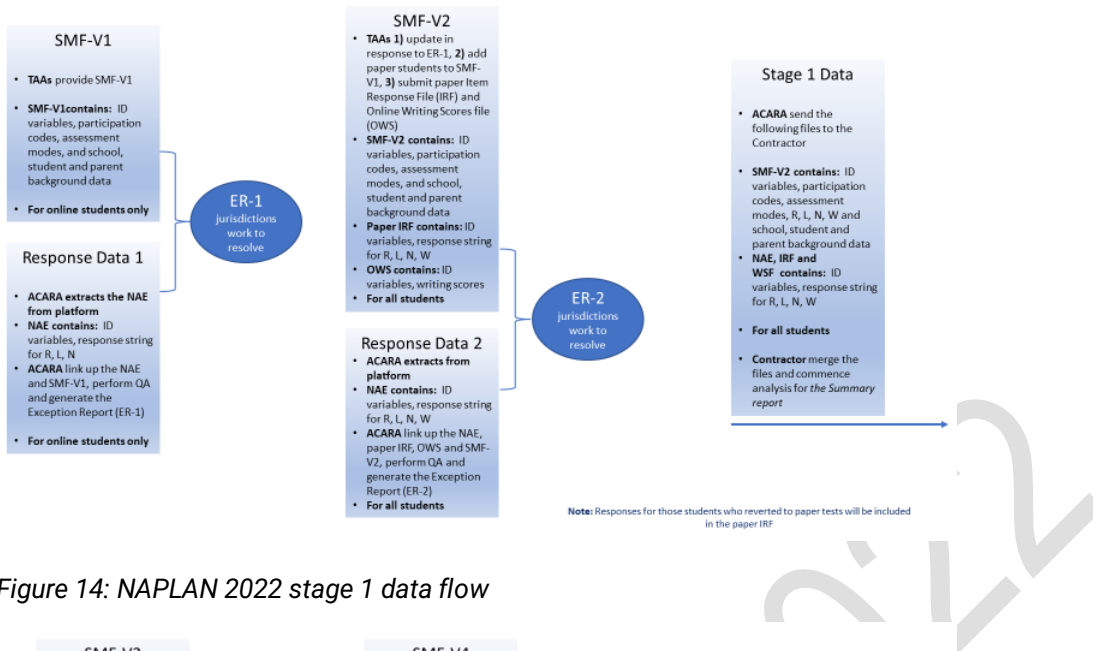


Figure 14: NAPLAN 2022 stage 1 data flow

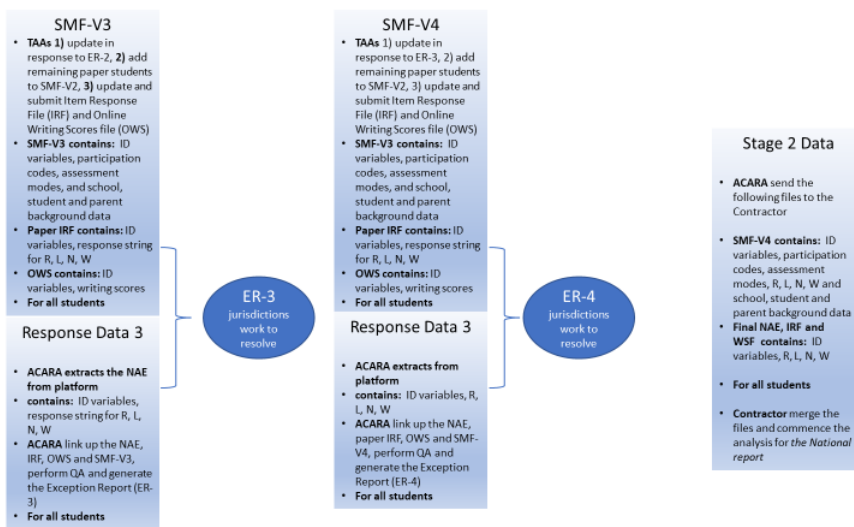


Figure 15: NAPLAN 2022 stage 2 data flow

Paper tests

Data collection for paper tests was undertaken by the TAAAs in the jurisdictions. A systematic process involving data checking was used by ACARA to ensure that each dataset was consistent with national code frames and data dictionaries. There are several types of exception rules implemented in the NAPLAN QA scripts such as structural, show-stopper, advisory and statistical. A sample of the exception rules is included in Appendix P.

Online tests

Education Services Australia (ESA) managed the online national assessment platform on which the NAPLAN 2022 online tests were delivered. The Australian Council for Educational Research (ACER) received the online test data extracted from the platform directly from ACARA by domain as those became available. With the tight timeline between the online assessments and the delivery of School and Student Summary Reports (SSSRs), quality assurance checks of online data extracted from the platform along with the SMF and IRF started in late May. The preparation for online data checking and management and for the analysis of online data followed the quality assurance check. Data integrity checking included verification that online data files conformed to their data dictionary and coding conventions (supplied by ACARA) and that item responses in the data files conformed to the valid codes specified in the code frames.

Data cleaning validation process

All data files were checked for invalid codes and inconsistencies. Data were cleaned and recoded. Any concerns about data were communicated to the relevant TAA directly and rectified as necessary. Recoded data files were generated and verified in preparation for data analysis. This was carried out for both the paper-based tests and the online tests.

Data preparation

The recoding of test data was conducted by the contactor prior to data analysis.

In 2022, responses to multiple-choice items were indicated by the number of the chosen response option for each item; that is, 1, 2, 3, 4 or 5. Responses for students not participating on a particular test or testlet were recoded to “R” and treated as *not administered*. Multiple responses to multiple-choice item on paper tests (“7”) were treated as *incorrect*. Embedded missing responses were coded as “9” and treated as *incorrect*. Trailing missing responses were also coded as “9” for the first unanswered item and treated as *incorrect*, while the remaining trailing missing items were recoded as “M” and treated as *not reached*. These not-reached items were treated as *not administered* items for item calibration to obtain an appropriate estimate of the item difficulty (for students who had a chance to respond). However, these not-reached responses were treated as *incorrect* for the final estimation of student abilities. Finally, students who were present but did not attempt any item (“non-attempts”) had their responses recoded to “R” and treated as *not administered*. In summary:

- 7 multiple/invalid response
- 9 embedded missing
- M not reached
- R not administered/ non attempt.

Data for partial-credit items were indicated by ordered categories starting with 0 up to the maximum possible value. Short-answer items were given scores of 0 or 1. The rules for data coding are provided in Table 65.

Table 65: Rules for data coding

Participation code	Data recoding rule
P – present	<p>Data string (i.e. item responses) expected. Any embedded missing responses are indicated with a 9, invalid responses with a 7.</p> <p>The first trailing missing response is kept as a 9; subsequent trailing missing responses are retained as trailing-missing responses, and are recoded as a M. Any embedded missing responses within the data string are kept as a 9.</p> <p>Students who are present but do not attempt any question (“non-attempts”) have their responses recoded to a string of Rs.</p> <p>Additionally, for the online tailored test data, responses for items in those testlets that were not administered to the students are coded as a R.</p>
A – absent	<p>A data string of all 8s for that test was expected from the TAA. Item response data are recoded as a string of Rs (this is like “not-administered”).</p>
S – sanctioned abandonment	<p>Response data are recoded as a string of Rs. This is specifically used to indicate students who unexpectedly abandon the test due to illness or injury. See National Protocols for Test Administration, section 5.5.</p>
W – withdrawn	<p>A data string of all 8s for that test. See National Protocols for Test Administration, section 5.4. Response data are coded as a string of Rs.</p>

E – exempt	A data string of all 8s for that test. See National Protocols for Test Administration, section 5.2.
C – cancelled	
N – no longer enrolled	These students are not included in the calibration or the calculation of means. Item data are recorded as a string of Rs.

Students who did not attempt all 3 testlets of the online tests had incomplete pathways. In these cases, predefined rules were applied to assign stage 2 and stage 3 testlets to a student's pathway. Responses to items in these testlets were coded as not reached (M). The rules are listed in Table 66. For example, students who only attempted some items in testlet A were assigned to pathway ABE. Similarly, students who aborted the test while attempting testlet B or D during stage 2 were assigned testlet E in stage 3.

Table 66: Pathway assignment rules to incomplete online tests

Domain	Last item attempted		Assigned pathway
Numeracy, Reading, Grammar & Punctuation	None		ACB
Numeracy, Reading, Grammar & Punctuation	Stage 1	A	ABE
Numeracy, Reading, Grammar & Punctuation	Stage 2	B	ABE
Numeracy, Reading, Grammar & Punctuation	Stage 2	C	ACB
Numeracy, Reading, Grammar & Punctuation	Stage 2	D	ADE
Spelling	None		SASBPB
Spelling	Stage 1	A	SASBPB
Spelling	Stage 2	B	SASBPB
Spelling	Stage 2	D	SASDPB

Distribution of not reached items

Ensuring that tests were designed so that the vast majority of students had sufficient time to submit valid responses to all items was an important consideration. This section provides percentage of trailing missing responses across all students for a given online test pathway.

Not reached items in online tests

Figure 16 to Figure 19 show the percentage of trailing missing responses by year levels and test pathways in numeracy, reading, spelling, and grammar and punctuation for the online tests. In these charts, the trailing missing responses were shown for one set of parallel testlets (for example, testlets A1 to F1 for numeracy, reading, and grammar and punctuation, and testlets SA1 to PD1 for spelling). Across domains, grammar and punctuation had the lowest trailing missing rates. In numeracy and spelling, trailing missing responses started to appear from the third testlet of a test, and increased towards the end of a test. Across test paths, the most difficult pathway A1-D1-F1 had the highest trailing missing rates in Years 5 and 7 numeracy tests. In spelling, the easiest pathway SA1-SB1-PB1 had the highest trailing missing rates in Years 3, 5, 7 and 9. In Year 5 and 9 reading, and Years 3, 5, 7 and 9 grammar and punctuation, the pathway A1-C1-B1 had the highest trailing missing rates. This is consistent with students branching to the easiest testlet (C) from A and subsequently branching to a harder testlet (B). Similar patterns of trailing missing responses were found in other parallel testlets.

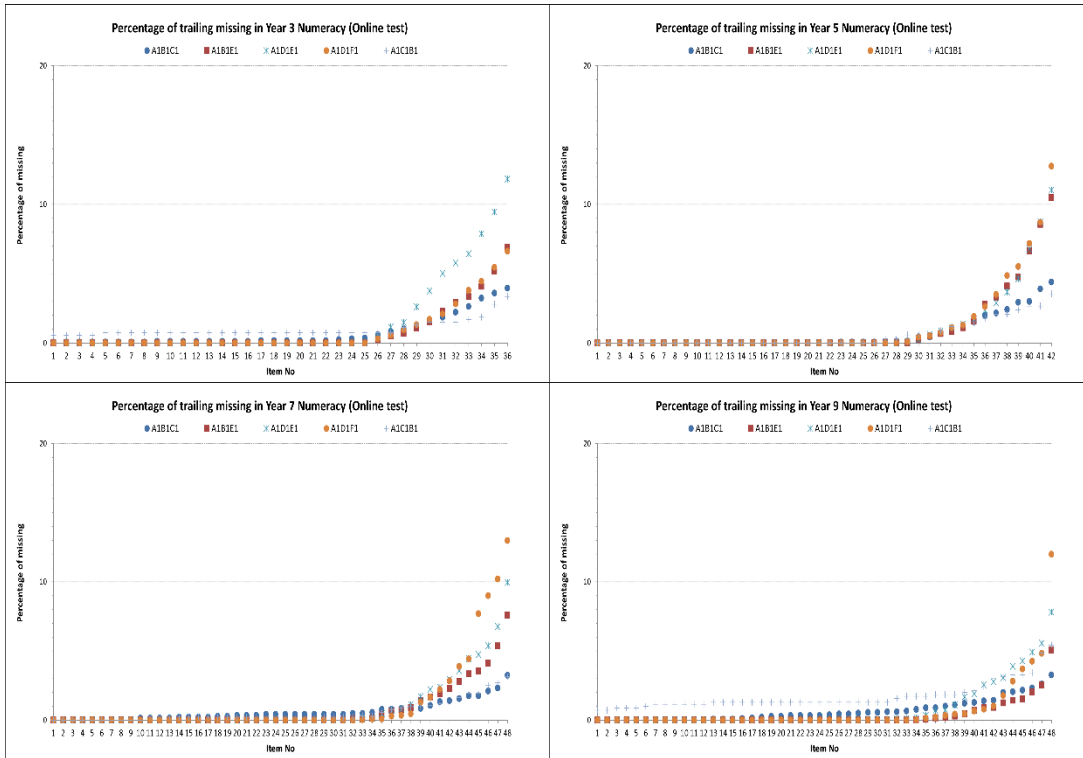


Figure 16: Trailing missing percentage in numeracy

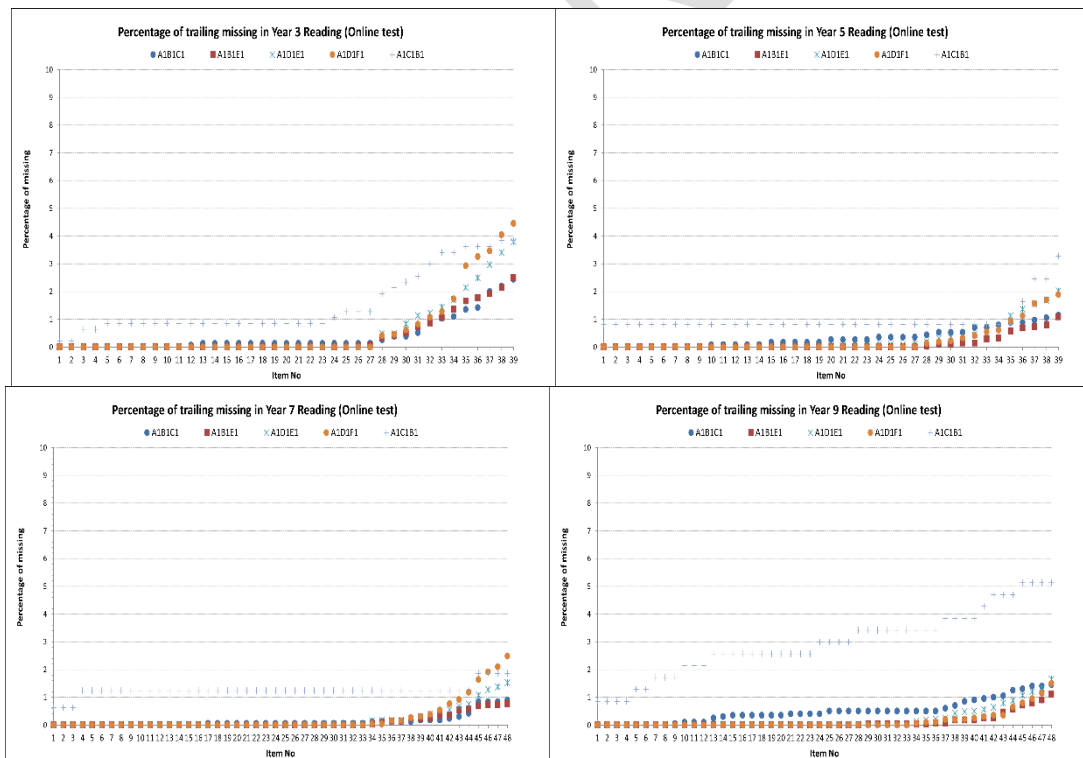


Figure 17: Trailing missing percentage in reading

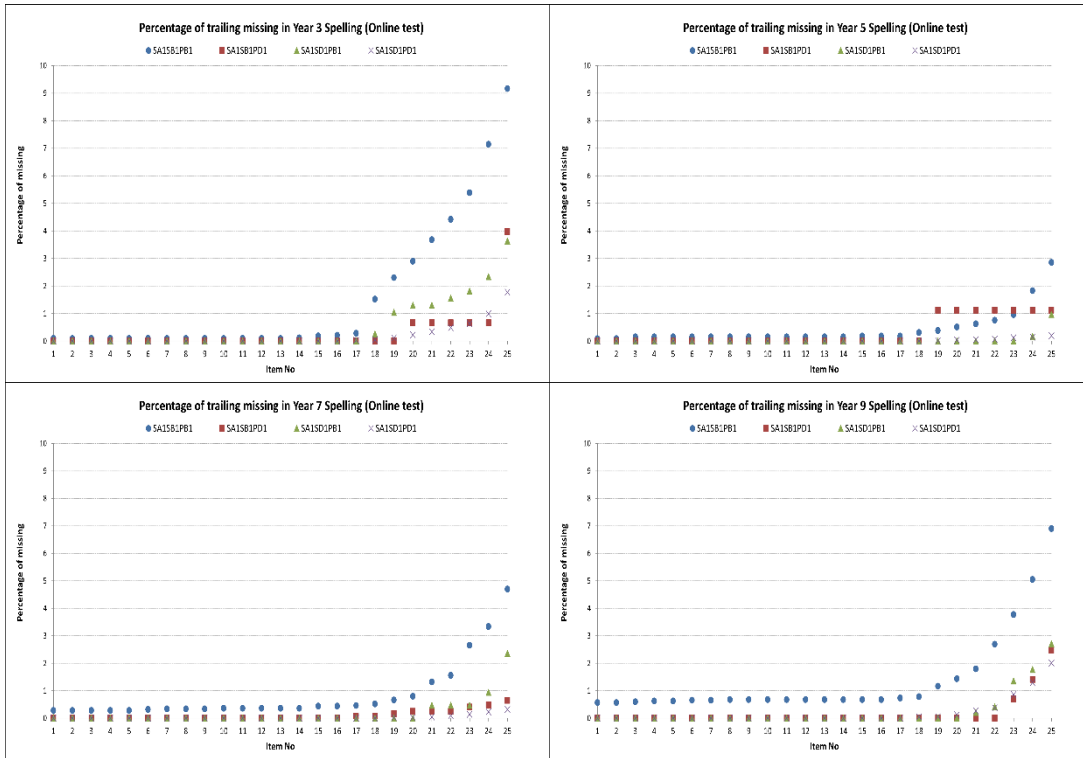


Figure 18: Trailing missing percentage in spelling

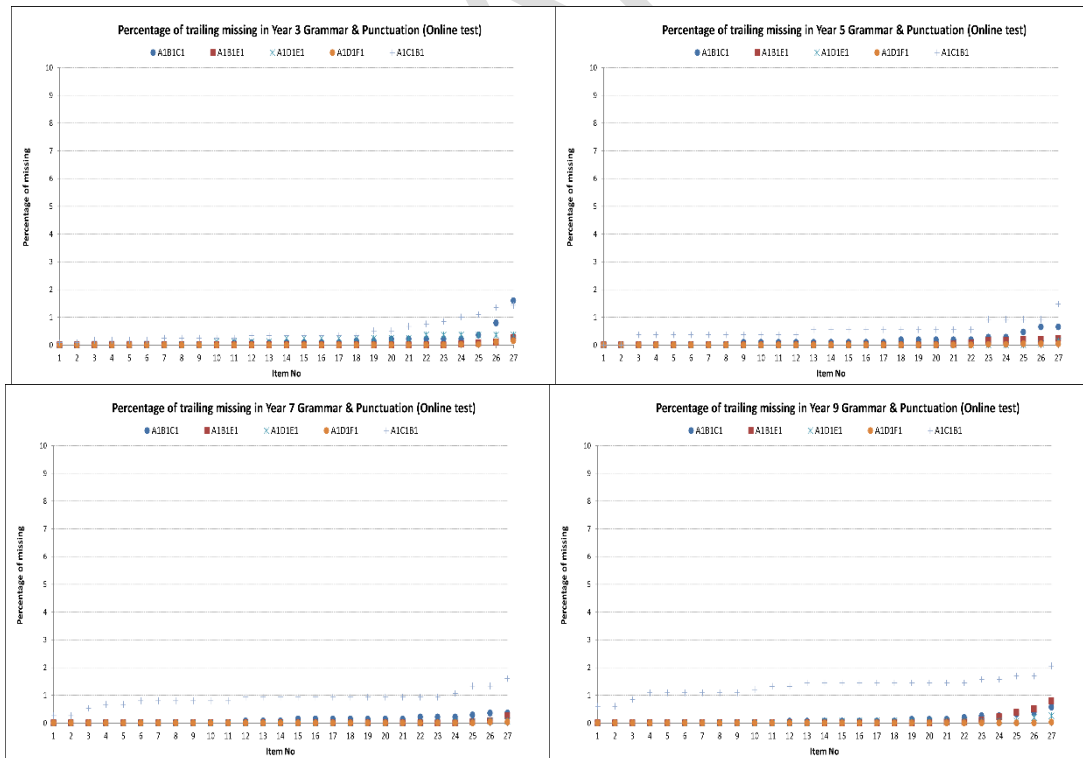


Figure 19: Trailing missing percentage in grammar and punctuation

Final student participation rates

The participation category diagram for NAPLAN 2022, with data file participation code shown in parentheses, is shown in Figure 20. Participating students include present (assessed, non-attempts) and not present (exempt) students. Final student participation rates for NAPLAN 2022 are presented in Table 67. The participation rate standard was 90% at national and jurisdictional level to ensure unbiased population statistics. Results in the National Report were annotated if the standard was not met. These percentages are coloured red in Table 67.

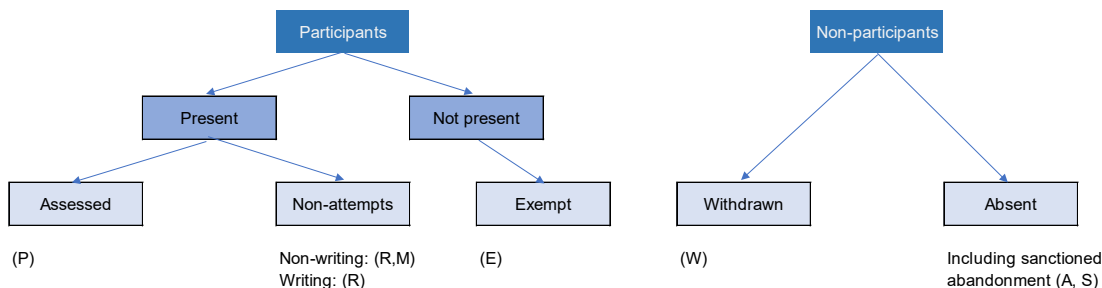


Figure 20: NAPLAN 2022 participation categories

Table 67: Student participation rate

TAA	Year level	Numeracy (%)	Reading (%)	Writing (%)	Spelling (%)	Grammar and punctuation (%)
NSW	3	94.5	96.5	93.8	95.5	95.5
Vic.	3	93.6	94.6	91.4	93.6	93.6
Qld	3	90.4	92.6	91.1	91.3	91.3
WA	3	94.6	95.4	94.6	94.8	94.8
SA	3	92.8	94.2	92.4	93.1	93.1
Tas.	3	93.5	95.4	93.8	94.5	94.5
ACT	3	92.0	93.5	90.7	92.5	92.5
NT	3	78.7	81.4	79.3	80.3	80.3
Aus.	3	93.1	94.7	92.4	93.7	93.7
NSW	5	94.8	96.8	95.9	95.9	95.9
Vic.	5	94.0	95.3	94.4	94.1	94.1
Qld	5	90.4	92.6	92.0	91.4	91.4
WA	5	95.2	96.2	96.2	95.4	95.4
SA	5	93.1	94.7	94.2	93.7	93.7
Tas.	5	94.1	95.8	95.3	94.7	94.7
ACT	5	91.8	93.7	92.6	92.6	92.6
NT	5	78.4	81.1	81.6	79.5	79.5
Aus.	5	93.3	95.1	94.4	94.0	94.0
NSW	7	92.4	94.9	94.2	93.5	93.5

TAA	Year level	Numeracy (%)	Reading (%)	Writing (%)	Spelling (%)	Grammar and punctuation (%)
Vic.	7	91.4	93.6	92.8	91.4	91.4
Qld	7	84.9	87.8	86.9	85.5	85.5
WA	7	92.4	94.5	94.4	92.6	92.6
SA	7	90.8	93.0	92.4	91.4	91.4
Tas.	7	90.5	93.6	92.6	91.5	91.5
ACT	7	87.4	90.8	89.6	87.6	87.6
NT	7	75.3	78.0	78.7	76.0	76.0
Aus.	7	90.1	92.6	91.9	90.7	90.7
NSW	9	88.3	91.4	90.9	89.6	89.6
Vic.	9	87.5	89.7	89.0	87.2	87.2
Qld	9	77.4	80.4	79.8	78.0	78.0
WA	9	90.1	92.1	92.0	89.8	89.8
SA	9	86.2	89.0	88.5	86.8	86.8
Tas.	9	84.7	88.8	88.2	85.9	85.9
ACT	9	82.3	86.6	84.4	83.6	83.6
NT	9	67.8	71.6	72.4	69.4	69.4
Aus.	9	85.4	88.2	87.6	85.9	85.9

Chapter 6: Scaling methodology and outcomes

This chapter describes the processes and methodologies used in the NAPLAN 2022 central analysis, as well as the outcomes of the scaling analysis. The psychometrics and scaling methods used are methods that have been applied in many large-scale assessment programs, including the Programme for International Student Assessment (PISA).

Scaling model

Test calibrations and scaling for both the online tests and the paper tests were undertaken with the Rasch model, as was the case in previous administrations.

For multiple-choice items and constructed-response items with a category score 1 for correct responses and 0 for incorrect responses, the Rasch model predicts the probability of a correct response given the latent trait (θ_n) and the item difficulty or location (δ_j). This is expressed as

$$P_i(1|\theta_n) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad (1)$$

where $P_i(1|\theta_n)$ is the probability of person n to score 1 on item i . θ_n is the estimated latent trait of person n , and δ_i the estimated location of item i on this dimension. For each item, responses are modelled as a function of the latent trait θ_n .

In the case of items with more than 2 categories, this model can be generalised to the Partial Credit Model (Masters 1982) as

$$P(X_{ni} = x|\theta_n) = \frac{\exp \sum_{j=0}^x (\theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i} \exp \sum_{j=0}^h (\theta_n - \delta_i + \tau_{ij})} \quad x = 0, 1, \dots, m_i \quad (2)$$

where $P(X_{ni} = x|\theta_n)$ is the probability of person n to score x on item i . θ_n denotes the person's latent trait estimate, the item parameter δ_i gives the location of the item on the latent continuum, and τ_{ij} is a step parameter of score j on item i .

It should be noted that both item (difficulty) and person (ability) parameters are measured on the same scale: in the case of dichotomous items with just 2 categories (correct and incorrect), for students with an ability (θ_n) equal to the difficulty of an item (δ_i), the probability of giving a correct response is 0.5.

Software used for analyses

For the Rasch scaling analysis, the software *ACER ConQuest 5* (Adams et al. 2022) was used. *ACER ConQuest 5* provides tools for the estimation of a variety of item response models and regression models. It was used for test calibrations, for generating weighted likelihood estimates (WLEs) used for the score-equivalence tables, and for drawing plausible values (PVs) based on a multidimensional item response model with latent regression. The marginal maximum likelihood (MML) estimation method was used for test calibrations and for generating the plausible values. When calibrating items from multistage adaptive test designs, it has previously been shown that MML estimation produces unbiased estimates (Eggen and Verhelst, 2011, Adams and Lazendic 2013).

Item calibration

Item response data for the online calibration of non-writing domains was extracted as soon as data was collected for 40% of students within each jurisdiction for all year levels. In total, the number of students included in the estimation of each domain was between 150,000 and 190,000 per year level.

For 2022 NAPLAN online tests, the numeracy, reading, spelling, and grammar and punctuation tests were calibrated separately by domain, year level, resulting in 16 separate calibrations. For each of the 4 non-writing online tests, items from all testlets within a domain and year level were calibrated in a concurrent analysis. In 2022, there were only a small number of students who completed NAPLAN paper tests, and it was not possible to construct a representative national calibration sample due to the student distribution. Therefore, no paper test calibration was carried out. Since all questions in the paper tests are included in the online test, item parameters in the paper test were anchored to their values from the online test.

For 2022 writing, the resulting scripts from students who responded on paper (mainly Year 3 students) or online from different tasks were scored using the same marking rubric based on 10 criteria. The scored writing data from Years 3, 5, 7 and 9 were calibrated concurrently using a sample of approximately 100,000 students' data, based on the partial credit model with the latent distribution conditioned on year level and assessment mode. The reason for the concurrent calibration was that some scores did not occur for some year levels. The calibration results obtained from the 2022 calibration were compared with parameters from previous NAPLAN cycles.

In the estimation of parameters, unreached-missing (M) and responses from an absent student (R, including *absent*, *withdrawn* and *exempt*) were treated as *not administered*, and embedded-missing (9) and invalid responses (7 in paper tests) were treated as *incorrect* responses. Non-attempts (students who were present for the test but did not answer any items) have only Rs, no 9s. Online items that were not included in a student's pathway and therefore not presented to students (R) were treated as *not administered* in all analyses.

Only students with complete test paths were included in the calibration data. The senate weight was used for calibrating the online numeracy, reading, spelling, and grammar and punctuation tests to ensure each jurisdiction was equally represented.

For each jurisdiction, a senate weight was calculated for online calibration according to the following equation:

$$SenateWeight_{jurisdiction} = \frac{StudentWeight_{jurisdiction}}{Sum(StudentWeight_{jurisdiction})} \times Sum(StudentWeight_{NSW}) \quad (3)$$

The student weight is equal to 1 for each student. This means for each jurisdiction, the sum of the senate weights was equal to the sum of the senate weights for the jurisdiction with the largest student population: New South Wales.

For the writing item calibration, equal representation of each jurisdiction was achieved by selecting a random sample from each of the remaining 7 TAAs to match the number of students in Northern Territory.

Review of test and item characteristics

The ACER ConQuest 5 item analysis results for NAPLAN 2022 online tests are given in Appendix B. This is an item-by-item tabular display of classical item statistics: item facility, discrimination and point-biserial statistics, counts and percentages of each response option (for multiple-choice items), score-points (for scored items), Rasch item parameters and infit mean square fit statistics. The item parameters shown in these tables are case-centred (that is, the mean of case estimates is set to zero) within each domain and year level.

Any summary statistics (e.g. Mean) shown at the end of the item analysis results for the online numeracy, reading, spelling, and grammar and punctuation tests are to be ignored. This is because these were not for any one test form but were for the whole item pool at each year level, meaning their interpretation is not straightforward.

The Rasch item parameter estimates and statistics are summarised in Appendix C for the online items in each of the 16 item pools for the numeracy, reading, spelling, and grammar and punctuation tests across all 4 year levels. The item parameters shown in these tables are delta-

centred for each test (that is, the mean of item difficulties for each scale are set to zero). The 95% confidence interval for the expected value of the mean square infit is also provided for each item.

Item Characteristic Curves (ICCs) for all online items are shown in Appendix D. The ICC plot shows a comparison of the empirical ICC based on observations from 8 ability groupings (broken line joining 8 dots) and the expected model-based ICC (smooth line). Equal-distance grouping was used for each test node (generic testlet) for online tests with different ability range. The 2 curves should display small or no disparities for an item that has good fit to the model. Since the ICC for a multiple-choice item also shows the proportion of students in each of the 8 groups who responded to each distractor in the category characteristic curves, the performance of distractors can be examined using the item analysis results and the response curves in the ICC plots.

Expected Score Curves for the online writing test criteria are shown in Appendix E. These show a comparison of the observed and the modelled expected score curve for each criterion.

Test reliability

Table 68 shows the IRT-based reliabilities (WLE and EAP/PV) of each online test and for the writing test.

The WLE reliability coefficients were between 0.91 and 0.94 for the numeracy tests, between 0.88 and 0.91 for the reading tests, between 0.90 and 0.93 for the spelling tests, and between 0.81 and 0.85 for the grammar and punctuation tests. The EAP/PV reliability coefficients were between 0.88 and 0.95 for the numeracy tests, between 0.83 and 0.87 for the reading tests, between 0.88 and 0.90 for the spelling tests, and between 0.78 and 0.84 for the grammar and punctuation tests. The reliability coefficient for the writing test was 0.96 and 0.92 for WLE reliability and EAP/PV reliability, respectively. In general, the WLE reliability is higher than the EAP/PV reliability, except for the year 9 numeracy and spelling tests, where EAP/PV was slightly higher or equal.

Table 68: Reliability (EAP/PV, WLE) for NAPLAN 2022 tests

Year level	Numeracy		Reading		Spelling		Grammar and punctuation		Writing*	
	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV
3	0.91	0.88	0.91	0.85	0.93	0.89	0.85	0.83		
5	0.92	0.89	0.88	0.83	0.92	0.88	0.81	0.78	0.96	0.92
7	0.93	0.93	0.89	0.84	0.92	0.89	0.81	0.78		
9	0.95	0.95	0.90	0.87	0.90	0.90	0.85	0.84		

*For Years 3, 5, 7 and 9 together

Test targeting and item spread

The purpose of the item-person map (or Wright map) is to compare the distribution of student locations (on the left side of the map) and the item thresholds (on the right side of the map). Item, step and person parameters are plotted on a common scale on a map. Appendix F provides the item-person maps for each domain at each year level for the online tests. It is important to note that for the online tests, the item-person maps are not for specific testlets or pathways but instead display the distribution of student locations against the item difficulties of all the items (in all testlets) within the domain online item pool at a year level.

For dichotomously scored non-writing tests, the item-person maps are constructed so that a student has a 50% chance of answering an item correctly when the item is at a difficulty level that is at the same level as the student's ability. On each item-person map, the mean of the case estimates was centred at zero. Students at the top end of the distribution had higher proficiency estimates, while items at the top end were the more difficult items.

Figure 21 displays the item-person map for Year 3 numeracy online test. That map indicates that the current tests targeted the average numeracy achievement level of the student group quite well. The distribution of student abilities (each X represents approximately 267 students) matched up well with the distribution of item difficulties.

For the polytomously scored writing tests, the criterion difficulty of each of the 10 rating criteria is plotted in Figure 22 with the latent ability distribution on the left-hand side. Figure 23 shows locations of the Thurstonian thresholds of each item and again with the latent ability distribution on the left-hand side. The notation $a.b$ indicates threshold b of criterion a . The location of the threshold indicates the ability level required for a student to have 50% chance of achieving category b on criterion a . The maps show that the thresholds are well spread out and well separated.



Figure 21: Wright map for Year 3 numeracy online test (an example)

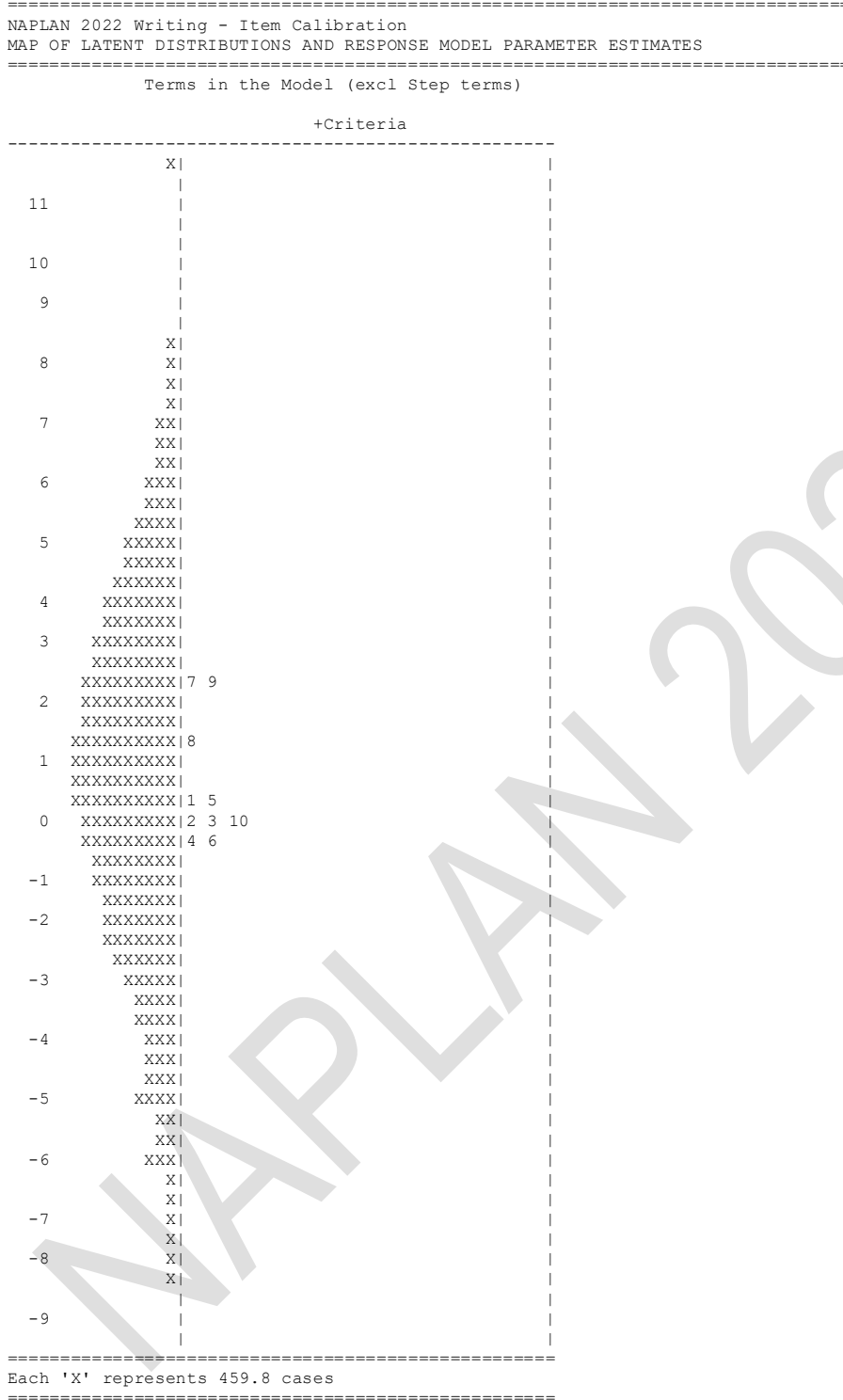


Figure 22: Wright map for writing test (a polytomous example)



Figure 23: Thurstonian thresholds for writing test

Item fit

The evaluation of goodness of fit to the Rasch model for individual items was based on the weighted mean square (infit mean square) statistics. Infit compares the observed residual variance with the expected residual variance if the data fit the model. Infit mean square is an IRT-based index for the degree an item discriminates between low- and high-achieving students. Values larger than 1 indicate low discrimination (or flatter ICC slope than expected) and values smaller than 1 indicate high discrimination (or steeper ICC slope than expected). We used an infit value of 1.20 as the criterion value for evaluating the goodness of fit, or the discrimination, of each item (that is, infit values greater than 1.20 indicate item misfit). We also calculated classical item statistics (that is, item-rest score correlation and facility) for the purpose of item fit evaluation, specifying criterion values for discrimination (based on item-rest score correlation) less than 0.25 and facility outside the range of 0.10 to 0.90. The infit mean square and classical item statistics of items included in NAPLAN 2022 tests can be found in Appendices B and C.

As mentioned earlier, the ICC of each item shows a comparison of the empirical ICC based on observations from 8 ability groupings (broken line joining 8 dots) and the expected model-based ICC (smooth line), and the 2 curves should display small or no disparities for an item that has a good fit to the model. The ICCs for all items can be found in Appendix D.

Item fit to the Rasch model was closely examined for numeracy, reading, spelling, and grammar and punctuation at each of the 4 NAPLAN year levels. As all items were trialled and examined before inclusion in NAPLAN tests, a few items are expected to show misfit. Because of the large size of the calibration sample, the confidence intervals for the infit mean squares were rather narrow.

Table 69 presents a summary of item statistics in the NAPLAN 2022 tests. It presents the number of items having infit mean square greater than 1.20. It also presents the number of items with a facility rate outside the range of 0.10 to 0.90.

As seen from Table 69, there were 23 out of 3,098 items from 16 non-writing online tests having infit greater than 1.20. There were 77 items with a facility rate higher than 0.90 and 39 items with a facility rate less than 0.10. Figure 24 shows the ICC of one numeracy Year 3 item (item Id: x00114420) with an infit statistic equal to 1.00. In contrast, Figure 25 shows the ICC of one Year 9 reading item (item Id: x00037734) with an infit statistic (1.26) higher than the criterion value (1.20) for evaluating the goodness of fit of each item. The item parameter estimates and statistics from item calibration are included in Appendix C for each of the 16 online tests and writing test.

The evaluation of goodness of fit to the Rasch model for individual writing criteria was also based on the weighted mean square statistics. The paragraphing and punctuation criteria exhibited misfit to the Rasch partial credit model; that is, infit are 1.51 and 1.64, respectively. None of the other criteria exhibited misfit to the Rasch partial credit model. Inspection of the ICCs did not reveal large differences between the empirical and the expected curves for each of the 10 criteria. The ICCs of the 10 writing criteria are included in Appendix D.

Table 69: Summary of item statistics in NAPLAN 2022 online tests

Domain	Year level	Total number of items	Number of items with Infit > 1.2	Number of items with	
				Facility > 0.90	Facility < 0.10
Numeracy	3	212**	1	4	0
	5	248**	0	7	1
	7	271	3	3	2
	9	272	4	2	1

Domain	Year level	Total number of items	Number of items with Infit > 1.2	Number of items with	
				Facility > 0.90	Facility < 0.10
Reading	3	234	2	1	0
	5	228	1	12	0
	7	288	0	13	0
	9	288	1	5	1
Spelling	3	123	3	1	9
	5	125	3	4	7
	7	127	2	9	5
	9	125	2	5	5
Grammar and punctuation	3	140	0	2	3
	5	132	0	3	1
	7	143	0	4	2
	9	142	1	2	2
Writing	3, 5, 7 & 9	10*	2	n/a	n/a

* Item in Writing is criterion.

** 213 items in original test design with one item deleted.

** 249 items in original test design with one item deleted.

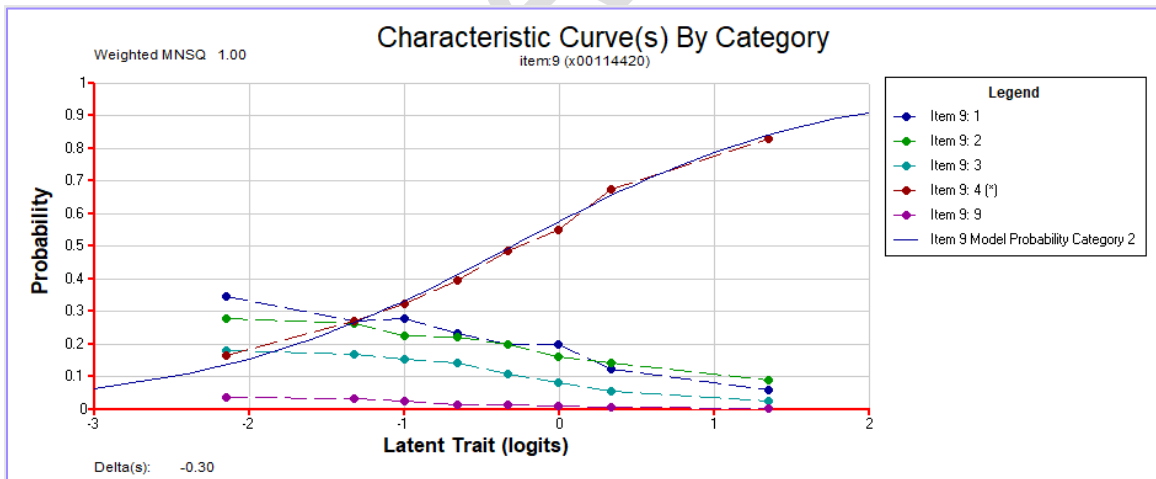


Figure 24: Item characteristic curves for an item with *infit* = 1.00

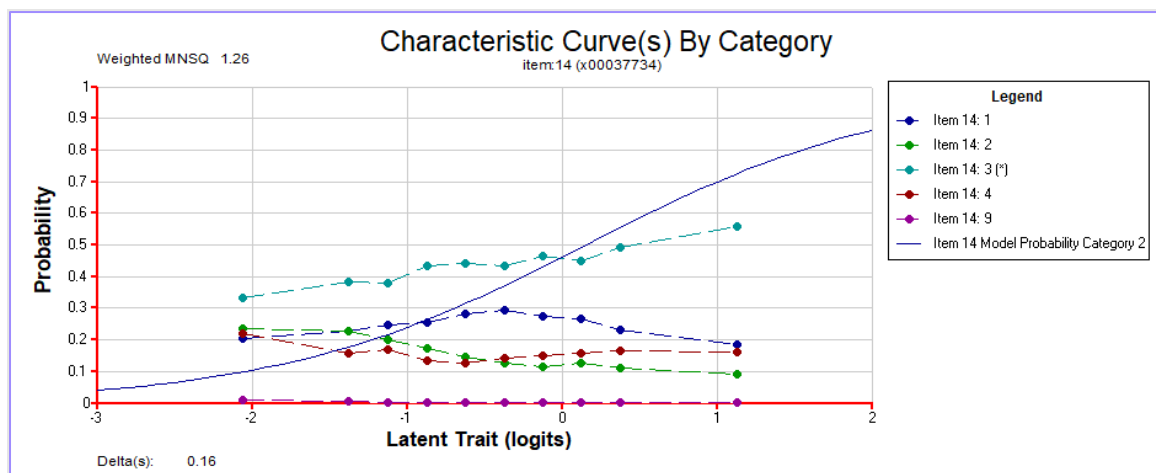


Figure 25: Item characteristic curves for an item with $infit = 1.26$

Differential Item Functioning (DIF) analyses

The functioning of the items was also evaluated through various DIF analyses. DIF occurs when groups of students with the same overall ability have different probabilities of responding correctly to an item (or of attaining certain item scores, in the case of polytomously scored items). Using the common example of gender DIF, if girls have a higher probability of success on a given item than boys with the same ability, the item is said to exhibit DIF, in this case favouring girls. It is important to monitor DIF, because DIF is a violation of an assumption of the Rasch model and can cause bias in the estimates. DIF by subgroup and DIF by jurisdiction analyses were performed for the online tests.

According to Camilli and Shepard (1994), item response theory can be used to assess DIF. Specifically,

[i]tem characteristic curves provide a means for comparing the responses of two different groups ... to the same item. A difference between the ICCs of two groups indicates that ... examinees [for the two groups] at the same ability level do not have the same probability of success on the item. More technically, DIF is said to occur whenever the conditional probability, $P(\theta)$, of a correct response differs for two groups. (Camilli and Shepard 1994)

In the analysis for NAPLAN, subgroups were arbitrarily categorised as either reference or focal groups. While males, non-LBOTE students and non-Indigenous students were assigned to the reference group, females, LBOTE students and Indigenous students were assigned to the focal group for DIF analyses. Independent Rasch analyses were then performed over the same set of items for each subgroup to examine any DIF that exists between 2 subgroups (for example, males versus females). The mean item difficulty for each subgroup was centred at zero to adjust for group differences in ability. The difference in the relative item difficulties after adjustment is referred to as the adjusted difference.

For visual depiction of DIF, item locations of the reference group are plotted against those of the focal group as seen from Appendices G, H and I (that is, gender, LBOTE and Indigenous status, respectively). Each item is represented by one point on the plot. An identity line ($y=x$) is plotted as the reference line. If the relative item difficulty for an item is not different between the 2 groups after taking their relative performance on the test into account, the point representing the item is on the reference line. The distance of a point from the diagonal reflects the magnitude of DIF. Due to the large sample sizes, confidence bands were very narrow and were not plotted on the charts.

Gender DIF

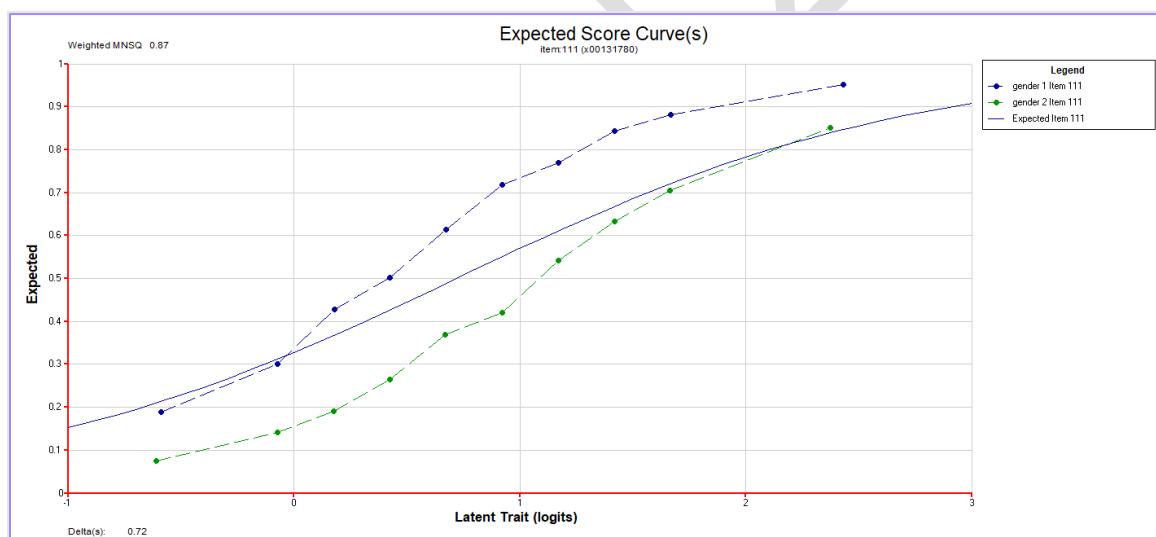
Appendix G presents the scatter plots for examining gender DIF in the 5 domains. The plots for numeracy, reading, spelling, grammar and punctuation are presented by year levels. The writing gender DIF was performed by combining all 4 NAPLAN year levels. On the whole, the plots

indicate that there are few items that exhibit gender differences in the adjusted item estimates and that any differences are not large and thus were not of great concern.

Table 70 identifies the number of items (out of the total number of items) that show gender DIF with an absolute difference of 0.50 or greater for numeracy, reading, spelling, grammar and punctuation, and writing¹. Figure 26 shows as an example, one Year 3 numeracy item (Item Id: x00131780) with an absolute difference of 0.50 or greater. This item was relatively easy (mean difference = -1.11) for male students. Appendix G includes DIF plots that show for each of the items the observed curves by gender group compared with the expected ICC.

Table 70: Number of items showing gender DIF by domain by year level

Year level	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
3	28/212	4/234	9/123	0/140	0/10
5	25/248	4/228	10/125	0/132	
7	22/271	13/288	14/127	2/143	
9	13/272	10/288	24/125	9/142	



† 'gender 1' indicates 'male' and 'gender 2' indicates 'female'.

Figure 26: Example of item characteristic curves displaying gender DIF†

Language background DIF

Appendix H shows scatter plots for examining DIF due to language background in the 5 domains by the 4 NAPLAN year levels. Writing LBOTE DIF was performed by combining all 4 NAPLAN year levels. These plots indicated that there were not many items that showed notable differences in the relative item difficulties.

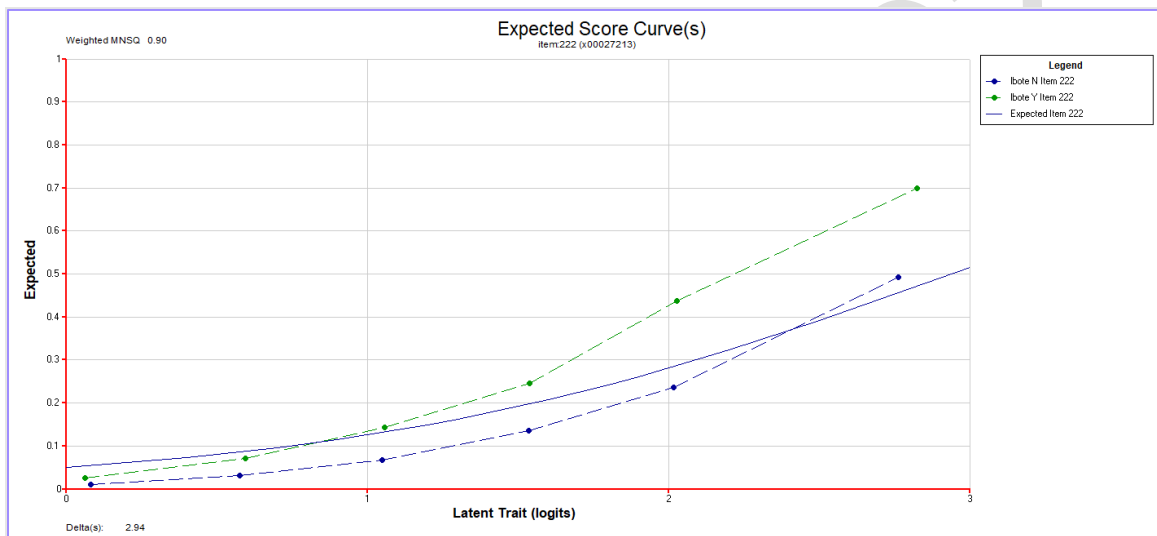
Table 71 indicates the number of items that show LBOTE DIF with an absolute adjusted difference of 0.50 or greater for numeracy, reading, spelling, grammar and punctuation, and writing. Figure 27 depicts one Year 5 numeracy online test item (item Id: x00027213) with an absolute mean

¹ For writing, item referred is marking criterion. This is applied throughout the report.

difference of 0.50 or greater. This item was relatively easy (mean difference = -0.88) for LBOTE students.

Table 71: Number of items showing LBOTE DIF by domain by year level

Year level	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
3	6/212	2/234	15/123	4/140	0/10
5	14/248	2/228	11/125	10/132	
7	8/271	1/288	13/127	9/143	
9	9/272	5/288	12/125	14/142	



† 'lbote Y' indicates 'LBOTE group' and 'lbote N' indicates 'non-LBOTE group'.

Figure 27: Example of item characteristic curves displaying LBOTE DIF†

Indigenous status DIF

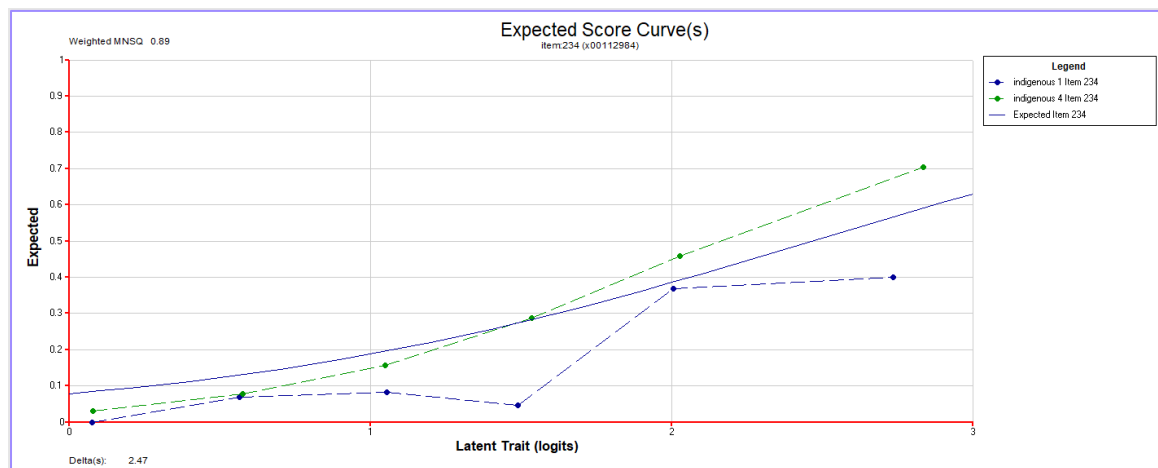
Appendix I includes scatter plots for examining Indigenous DIF in the 5 domains for online tests. Writing Indigenous DIF was performed by combining all 4 year levels. These plots showed that there were not many items that showed notable differences in the relative item difficulties for tests.

Table 72 lists the number of items that show Indigenous DIF with an absolute adjusted difference of 0.50 or greater for numeracy, reading, spelling, grammar and punctuation, and writing. Figure 28 depicts one Year 5 numeracy online test item (item Id: x00112984) with an absolute mean difference of 0.50 or greater. This item was relatively easy (mean difference = 0.71) for non-Indigenous students.

Appendix I provides the item DIF plots for items listed in Table 72. The plots show, for each of the items, the observed curves by Indigenous group compared with the expected ICC. In interpreting the plots, it should be noted that there may not be many Indigenous students along parts of the ability range. As a result, one would expect larger variability of empirical probabilities (that is, the dots connected by dashed lines) about the model-based curve (the solid curves).

Table 72: Number of items showing Indigenous DIF by domain by year level

Year level	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
3	4/212	7/234	1/123	9/140	0/10
5	13/248	9/228	3/125	8/132	
7	12/271	11/288	3/127	8/143	
9	13/272	6/288	1/125	6/142	



† 'indigenous 1' indicates 'Indigenous group' and 'indigenous 4' indicates 'non-Indigenous group'.

Figure 28: Example of item characteristic curves displaying Indigenous DIF†

DIF values of individual items for gender, LBOTE, Indigenous status, jurisdiction and device are presented in Appendix J.

Jurisdictional DIF

In order to determine whether state/territory DIF exists, all tests were calibrated independently by state/territory and year level. The relative item difficulties (or criterion difficulties for writing) were compared to the national item difficulty estimates obtained from the item calibration for the online tests. The following procedures were applied:

- Items were calibrated by jurisdiction, by domain and year level; item parameters were then delta-centred.
- The national delta-centred item parameter estimates from the item calibration were used.
- The parameter difference for item(i) between a state/territory and the national item parameter was calculated as:

$$Difference(i) = Item\ Parameter(i) - National\ Item\ Parameter(i) \quad (4)$$

- The difference was tested for statistical significance by dividing it by twice the standard error of the item parameter. If the absolute value obtained is greater than 1.96 then a statistically significant difference exists.
- Statistically significant differences were then compared against an effect size of 0.25.
- If the difference for an item between a state/territory and the national average was statistically significant and greater than 0.25 logits, then the item was deemed harder for the state/territory. If the difference was statistically significant and less than -0.25 logits, then the item was deemed easier for the state/territory.

The number of items showing statistically significant (and above 0.25 logits) state/territory related DIF in online numeracy, reading, spelling, grammar and punctuation, and writing are shown in Table 73. In the headings of Table 73, “E” indicates that the item is relatively easy for the jurisdiction and “H” indicates that the item is relatively hard for the jurisdiction. Table 73 can be read in conjunction with Appendix K, which contains item DIF plots for items showing state/territory related DIF for items listed in Table 73. The plots show, for each of these items, the observed curves by TAAs compared with the expected ICC. Figure 29 depicts one Year 3 numeracy online test item (item Id: x00075150) showing DIF among TAAs. This item was relatively easy for Qld and WA students, and relatively hard for SA and Tas students.

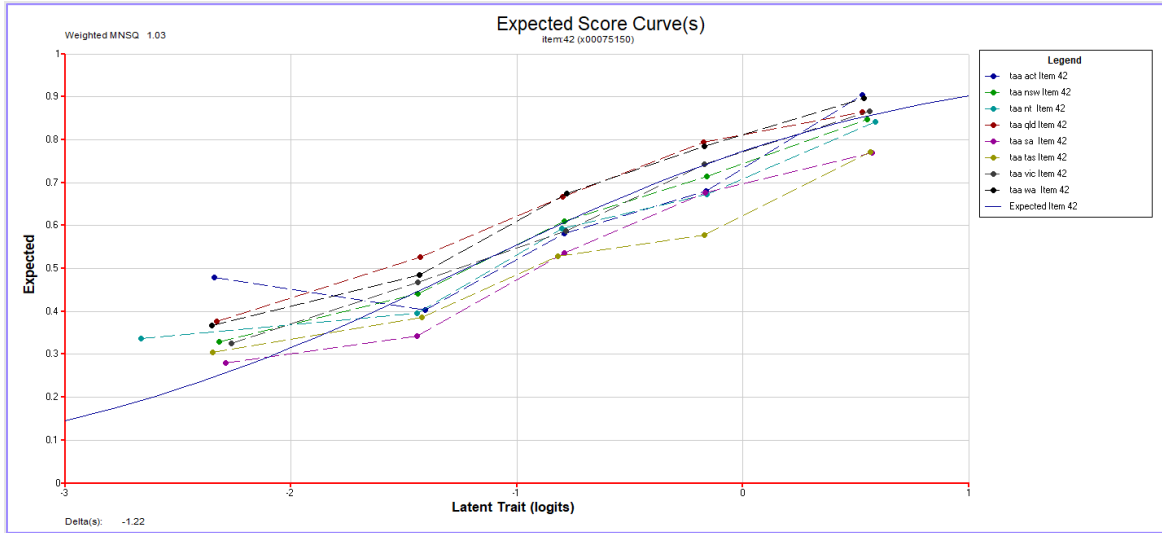


Figure 29: Example of item characteristic curves displaying jurisdictional DIF

Table 73: Number of items showing state/territory DIF by domain by year level

Domain	Year	ACT		NSW		NT		Qld		SA		Tas		Vic		WA	
		E	H	E	H	E	H	E	H	E	H	E	H	E	H	E	H
Numeracy	3	-	1	-	-	-	1	8	1	1	3	1	2	-	2	3	-
	5	-	2	8	5	-	1	12	3	-	1	-	2	7	6	1	-
	7	2	-	7	2	1	-	5	-	1	2	1	4	5	-	2	-
	9	2	-	13	5	1	-	2	3	-	1	-	3	6	1	8	2
Reading	3	-	1	-	-	-	-	1	-	2	-	-	1	1	1	1	-
	5	-	1	4	-	-	-	1	-	-	-	-	1	9	1	-	-
	7	-	-	5	-	-	1	2	-	-	-	-	14	1	-	1	-
	9	1	1	1	-	-	-	-	1	1	-	1	-	1	-	8	4
Spelling	3	1	-	1	3	-	-	2	1	7	3	-	-	-	4	5	7
	5	-	1	2	2	-	-	5	3	-	3	2	2	2	3	3	3
	7	3	2	2	4	1	-	2	3	2	1	2	2	1	1	3	4
	9	1	2	1	2	-	-	2	2	3	1	1	-	4	-	4	4
Grammar and punctuation	3	-	1	1	-	1	1	2	2	6	2	3	-	6	2	2	-
	5	-	3	4	-	-	3	8	2	2	-	2	4	8	1	-	1
	7	2	2	2	-	3	2	7	2	2	-	-	-	7	3	-	-
	9	3	-	-	-	-	-	3	1	-	-	-	1	1	-	2	3
Writing	3,5,7 & 9	-	-	1	-	2	3	2	3	-	-	1	2	1	2	-	-

Note. 'E' indicates that the item is relatively easy for the jurisdiction, and 'H' indicates that the item is relatively hard for the jurisdiction.

Device DIF

For online tests, a device DIF analysis was also carried out for non-writing domains¹ as there were different devices used by different students. There were 4 different types of devices: Chromebook, iOS, Mac and Windows. The same method used to determine jurisdictional DIF was used for determining device DIF. Table 74 shows the number of students using each device type at each year level and domain as used for the device DIF analysis. These numbers were based on the information recorded – not all students recorded device information.

For each type of device, items were calibrated separately, and then item parameters from each device were compared with pooled online item parameters. An item parameter demonstrating a significant difference greater than 0.25 logits was deemed as exhibiting DIF. A summary of device DIF is shown in Table 75. Table 75 shows that Mac devices had the most items demonstrating DIF, especially in numeracy, reading, and grammar and punctuation for Year 7 and Year 9 students, while Windows devices had shown the fewest items with DIF. Appendix L includes scatter plots for examining Device DIF across the different non-writing domains.

¹ Device DIF was not investigated for writing as some students completed the test on paper while others completed the test online.

Table 74: Number of students by device

Domain	Year level	Chromebook	iOS	Mac	Windows
Numeracy	3	35,275	67,242	2,658	86,381
	5	36,156	50,363	6,243	96,344
	7	16,752	15,284	30,447	125,739
	9	14,308	11,802	33,015	114,165
Reading	3	33,598	65,643	2,616	87,329
	5	34,029	47,615	5,773	92,598
	7	16,139	15,001	27,345	120,922
	9	12,528	11,145	29,337	103,699
Spelling	3	32,613	62,099	2,400	80,179
	5	33,499	47,163	5,519	89,163
	7	15,551	13,707	28,140	113,292
	9	12,761	10,741	30,278	99,387
Grammar and punctuation	3	33,046	62,659	2,425	81,826
	5	33,595	47,276	5,529	89,569
	7	15,592	13,732	28,167	113,690
	9	12,788	10,749	30,304	99,685

Table 75: Number of items showing device DIF by domain and year level

Domain	Year level	Chromebook		iOS		Mac		Windows	
		E	H	E	H	E	H	E	H
Numeracy	3	-	-	1	-	-	1	-	-
	5	6	3	-	-	2	1	-	-
	7	2	-	2	-	15	9	-	-
	9	1	1	2	1	21	9	-	-
Reading	3	3	-	-	-	-	-	-	-
	5	5	-	4	-	1	1	-	-
	7	1	2	2	-	24	6	2	-
	9	-	1	-	1	16	9	-	-
Spelling	3	1	2	-	-	-	-	-	-
	5	-	-	-	-	1	2	-	-
	7	1	-	-	1	-	3	-	-
	9	-	-	-	-	2	3	-	-
Grammar and punctuation	3	3	-	1	-	-	-	-	-
	5	4	-	2	-	1	5	3	-
	7	-	-	1	1	10	5	-	-
	9	-	-	-	-	8	4	-	-

Estimation of student ability and generation of PVs

For student and school-level reporting, weighted likelihood estimates (WLE) (Warm 1989) were produced. WLEs are point estimates of student achievement. Every student with the same raw score on the same set of items receives the same WLE score. Therefore, they are discrete scores. These estimates are unbiased for individual student scores, unless the test was too easy or too difficult for a student. However, population estimates based on WLEs may be biased. Population variances and covariances are overestimated when using WLEs.

For that reason, plausible values methodology was applied for producing population estimates. This approach, developed by Mislevy and Sheehan (1987) and based on the imputation theory of Rubin (1987, 1991), produces consistent estimators of population parameters. Instead of a point estimate, the most likely range is estimated for each student. This range is called the *posterior distribution*. Plausible values (PVs) are random draws from this distribution. For NAPLAN, a set of 5 plausible values was drawn for each domain for each student.

Score-equivalence tables based on WLEs in logits were generated for each test path of the online tests by domain and year level based on delta-centred item parameters. Similarly, score-equivalence tables based on WLEs in logits were also generated for each of the paper tests by anchoring item parameters on the online test item parameters. Transformations were applied to the logit scores for conversion to NAPLAN reporting scale scores on the historic NAPLAN scales, as was done in previous years.

For the estimation of population statistics, rather than using the WLE estimates, 5 sets of PVs of student latent proficiency estimates were drawn using *ACER ConQuest 5* based on imputation techniques and a multidimensional item response model (partial credit model) with latent regression (Adams et al. 2022) for students in each of the year levels for each of numeracy, reading, spelling, grammar and punctuation, and writing.

In drawing the plausible values, conditioning variables were used as regressors in the model. The plausible values were drawn by TAAs and by year level for both online and paper students together. The regression model used in 2022 was the same as that used in previous NAPLAN cycles with an additional regressor for test mode. The conditioning variables used in the model were gender, LBOTE status, Indigenous status, parental education, parental occupation, dummy variables based on sector by geolocation interactions, the school reading WLE average score (adjusted for the student's own score) as a measure of average proficiency at the school level and test mode. A diagrammatic representation of the multidimensional model is shown in Figure 30.

The categorical variables (gender, LBOTE status, Indigenous status, parental education, parental occupation, interaction dummy variables of school sector by school geolocation and test mode) were included in the model using what are referred to as *indicator variables*. In this approach, a single categorical variable was recoded by multiple indicator variables that were coded with a "1" to denote the presence of a category level, and a "0" to denote the absence of the category level. In general, it takes $k - 1$ indicator variables to recode k category levels. For example, the variable gender was designated as having 3 categories, namely, *male*, *female* and *missing*. The categories of gender were recoded for each student, using one indicator variable to denote *female* and a second indicator variable to denote *missing*. If the pair of indicator variables had the values 1 and 0 respectively, this meant that the gender category for the student was *female*; when the indicator variables had the values of 0 and 1, then the gender category was *missing*. When both indicators were 0, this indicated that the gender category for the student was *male*. In a similar fashion, this approach was applied to the other categorical variables used in the model. For each student, the school mean reading WLE score was calculated excluding that particular student.

Adding background variables as regressors to the conditioning model does not change the meaning of the constructs; only the item responses define the construct. Instead, conditioning on background variables increases the precision of population estimates and allows the analysis of relationships between proficiency estimates and background variables. The plausible values were drawn separately for each jurisdiction for all students (including absent students, withdrawn students and non-attempts) except for students who were exempt from NAPLAN testing.

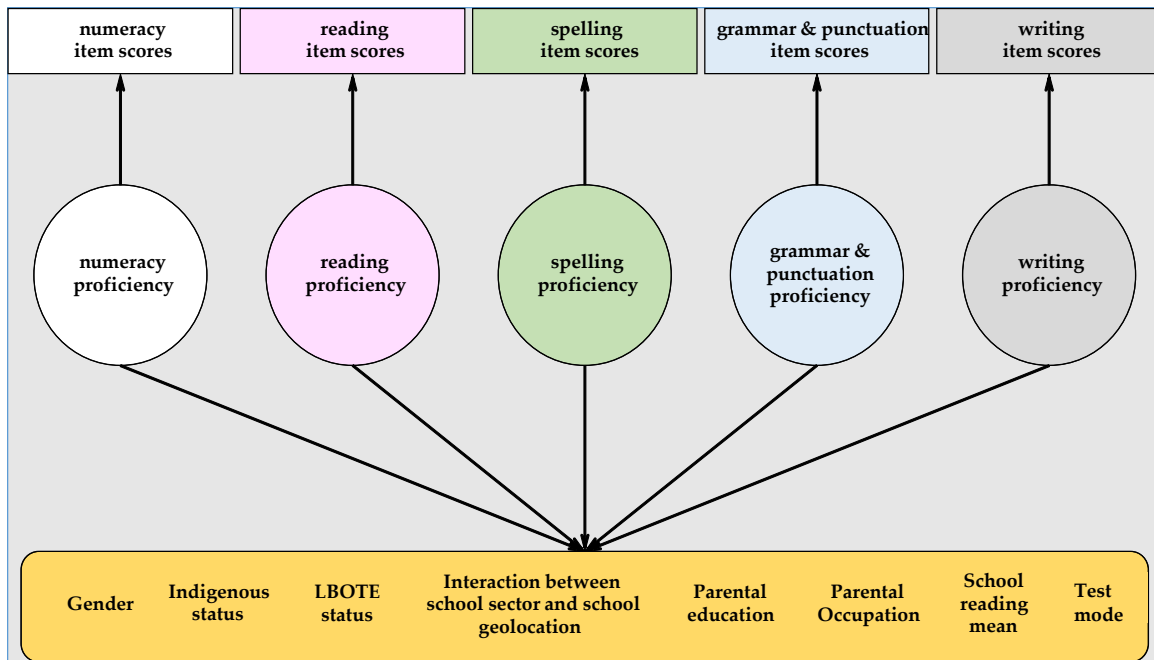


Figure 30: Conditioning variables for the multidimensional item response model with latent regression model

Chapter 7: Equating procedures

This chapter describes the process of equating the 2022 online tests onto the NAPLAN historical scales, and the procedure that located the paper tests onto the NAPLAN historical scales. The first section describes equating procedures for the numeracy, reading, spelling, and grammar and punctuation online test results. This is followed by a description of the method used for locating paper test results onto the NAPLAN scales. The chapter finishes with a description of the equating procedures for writing. It should be noted that a different equating design and methodology was applied for the writing domain.

Equating of numeracy, reading, spelling, and grammar and punctuation results

NAPLAN results are reported using 5 national achievement scales, one for each of the assessed domains of literacy – reading, writing, spelling, and grammar and punctuation – and one for numeracy. The horizontal and vertical equating design for the 2022 online tests is represented schematically in the data matrix in Table 76.

The 2022 online tests were linked to the historical scales by a set of common items used in the 2021 tests. Additionally, all items included in the 2022 paper tests were also in the 2022 online tests. Therefore, horizontal equating was based on a common item equating design that first placed the 2022 online tests onto the NAPLAN 2021 online test scales, and then transformed them onto the NAPLAN historical scales by applying the shifts and transformations that were used to locate the 2021 online tests onto the NAPLAN historical scales. The 2022 year level NAPLAN tests were also linked to each other by a set of common items between adjacent year levels. However, the vertical equating and horizontal-vertical regression (HVR) equating shifts were only used to evaluate the horizontal shifts as an additional quality assurance procedure.

Table 76: Equating design for online tests

NAPLAN 2022 online test items – horizontal links							
Items	Y3		Y5		Y7		Y9
Y3 2021 test							
Y5 2021 test							
Y7 2021 test							
Y9 2021 test							
NAPLAN online test items – vertical links							
Students	Y3	Y3&5	Y5	Y5&7	Y7	Y7&9	Y9
Y3 population							
Y5 population							
Y7 population							
Y9 population							

The 5 NAPLAN scales (one per domain) were established in 2008 by placing all year levels on the same scales using vertical link items. For the purpose of monitoring student achievement over time, the NAPLAN 2022 scale for each domain needs to be horizontally equated to the historical NAPLAN reporting scale. The horizontal links between the NAPLAN 2022 online tests and NAPLAN 2021 online tests included a large number of common items. Common item equating

was used as the final horizontal equating method to bring the NAPLAN 2022 scale onto the NAPLAN historical scale for online tests.

In theory, no vertical link items were needed after 2008, when all year levels were placed on the same scales, because each year level could be shifted onto the historical scales by common student equating using the equating tests (see ACARA, 2022 for an explanation of the equating tests and their use in equating). However, vertical link items were used in all subsequent years to check and adjust the horizontal shifts for each year level. This method was labelled the horizontal-vertical regression (HVR) equating method and can be found in previous years' technical reports. Appendix M presents the 2022 vertical link item locations (Rasch difficulty parameters) for the relevant year levels, standard errors, and differences in the item locations by domain and year level.

Before calculating the horizontal equating shifts, the quality of the common items in terms of their functioning as equating links was systematically reviewed. Only items that showed satisfactory and similar psychometric properties across test forms were used as link items.

A common item was considered for omission (that is, not to be used for linking purposes) based on the fit of the item and evidence of Differential Item Functioning (DIF) between test forms. Review of the horizontal link items was undertaken as follows:

- Initial cross-test form scatterplots with all items were examined to ascertain the overall correlation and to note any patterns and outliers.
- Items were omitted if they showed cross-test form DIF. To evaluate test form DIF, difficulties of the set of common items were centred on zero for each test form. For each pair of linked tests, one set of item difficulties (e.g. of 2022 Year 3 link items) was then plotted against the other set of item difficulties (e.g. of 2021 Year 3 link items). Two plots were presented in the following sections for each review: one plot for the set of link items to be reviewed and one plot for the retained link items after review and selecting good link items. On the plots, each dot represents a common item. Links were broken in 3 steps. Any link items from different nodes (A, B, C, D, E or F) were broken first, due to item positioning and potential differences in the subpopulation of students responding to items within each node. Outliers (absolute difference larger than 0.9 of a logit) were then broken. Any other items with an absolute difference of more than 0.4 were broken in the third step, and the process was repeated if necessary. For each set of linked test scales, mean item difficulties of the link items were calculated for each of the 2 test forms. The equating shift is the difference between the 2 means.
- In addition to relative item difficulty and node of the link items, item facility, (average) position of the item in the pathway, infit MNSQ and gender DIF were compared between the 2 linked tests.

The scatter plot was inspected with a focus on the agreement of bivariate data with the identity line. The ratio of the standard deviations of the item locations was checked for each linked test form (e.g. 2022 Year 3 SD / 2021 Year 3 SD). Ideally the ratio should fall between 0.85 and 1.15. A scaling factor was considered only if the ratio between the SDs of the link items was not between 0.85 and 1.15. Once a 2022 online test scale was shifted onto the 2021 scale, the same transformations were applied as in 2021 to further shift it onto the NAPLAN historical scale.

The link-item review procedure for vertical link items was similar to the review procedure for reviewing horizontal links. Links were broken in 2 steps. Outliers (absolute difference larger than 0.9 of a logit) were broken first. Any other items with an absolute difference of more than 0.4 were broken in the second step, and the process was repeated if necessary.

Horizontal equating shifts of the online tests

As already noted, there were 2 steps involved in equating the NAPLAN 2022 online tests to the NAPLAN historical tests. First, the 2022 NAPLAN online tests were equated to the NAPLAN 2021 online tests. This placed the NAPLAN 2022 online tests onto the NAPLAN 2021 online delta centred scales. Second, the equating parameters that were previously applied in 2021 to place the

2021 online test scales onto the NAPLAN historical test scales were applied to the NAPLAN 2022 online tests. This step resulted in the NAPLAN 2022 online tests being placed onto the historical NAPLAN scales. The top section of Table 76 shows the horizontal equating design for each of numeracy, reading, spelling, and grammar and punctuation at each year level.

Figure 31 to Figure 46 show the comparisons of the 2021 item parameter estimates with the 2022 item parameter estimates, for each of the 16 online tests. For link items that did not change in relative item difficulty, the bivariate points were on the identity line (a green dotted line on each graph). A thin solid line on each figure shows the linear line of best fit through the dots in each scatterplot and dashed black line shows the 95% confidence interval.

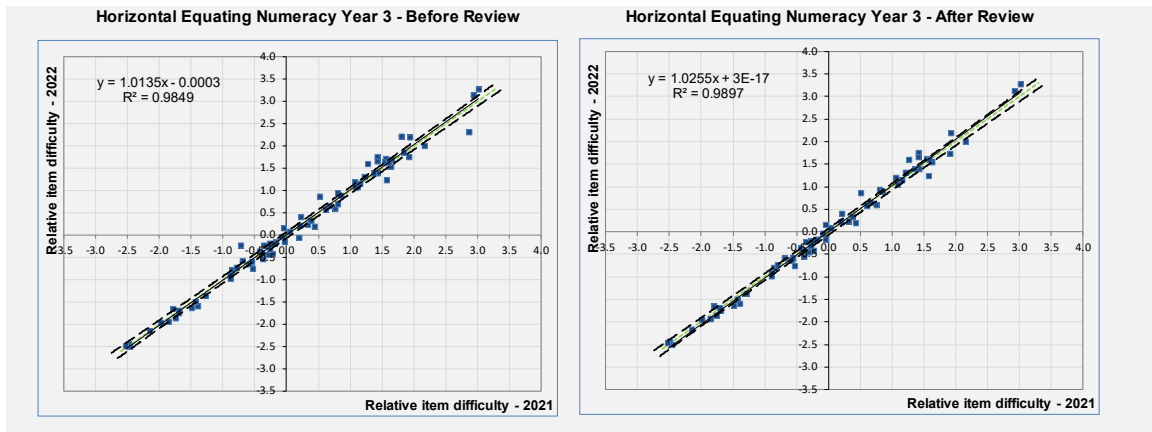


Figure 31: Scatterplot of numeracy, horizontal equating items between 2022 and 2021 for Year 3 online students

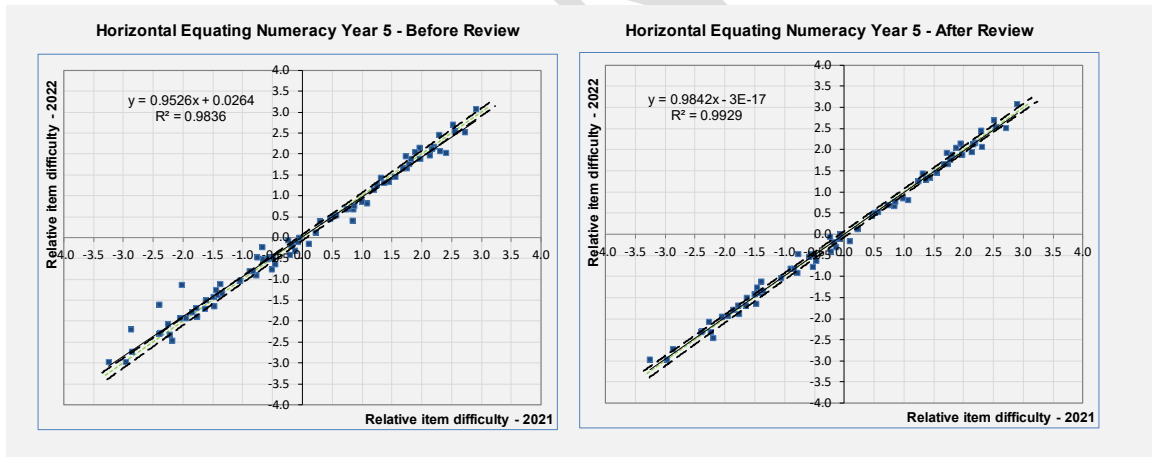


Figure 32: Scatterplot of numeracy, horizontal equating items between 2022 and 2021 for Year 5 online students

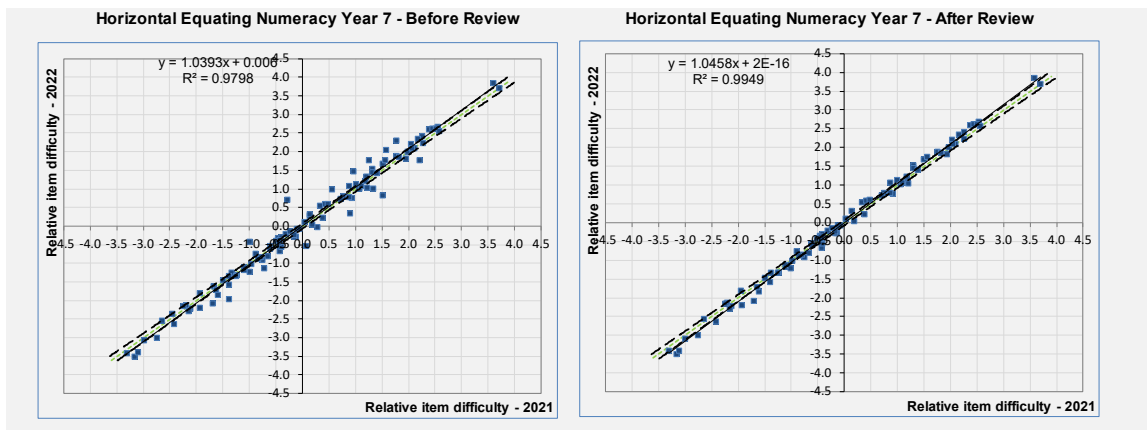


Figure 33: Scatterplot of numeracy, horizontal equating items between 2022 and 2021 for Year 7 online students

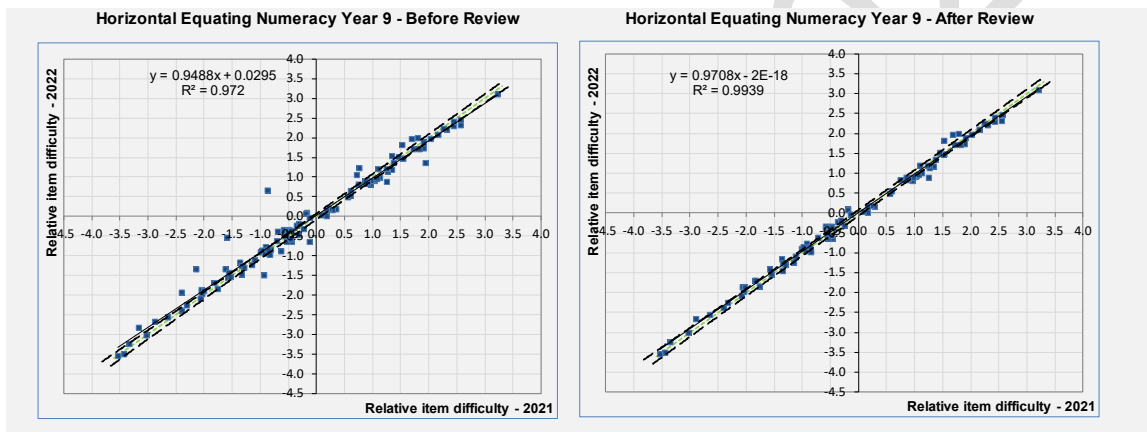


Figure 34: Scatterplot of numeracy, horizontal equating items between 2022 and 2021 for Year 9 online students

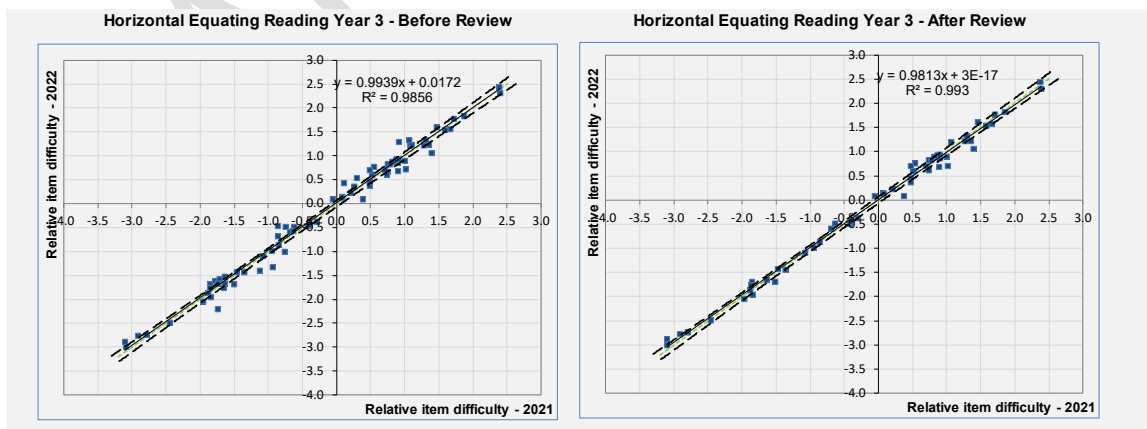


Figure 35: Scatterplot of reading, horizontal equating items between 2022 and 2021 for Year 3 online students

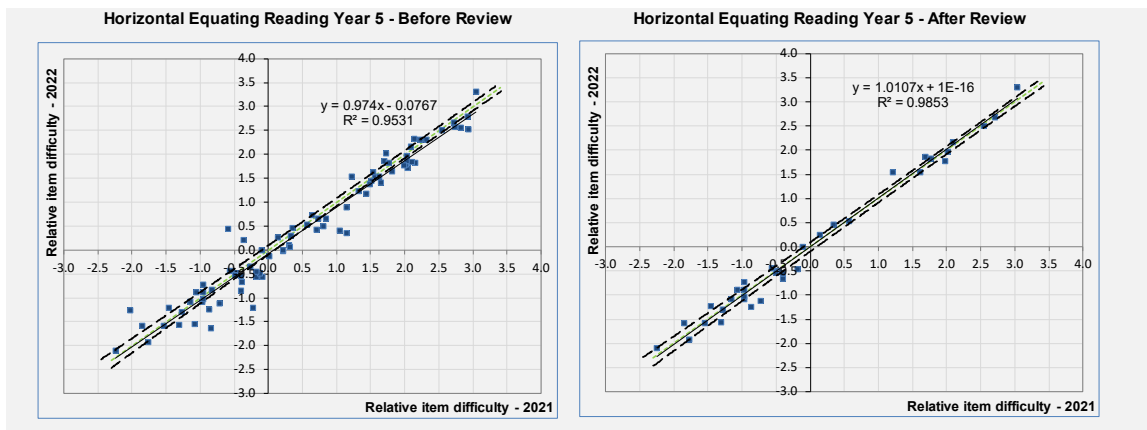


Figure 36: Scatterplot of reading, horizontal equating items between 2022 and 2021 for Year 5 online students

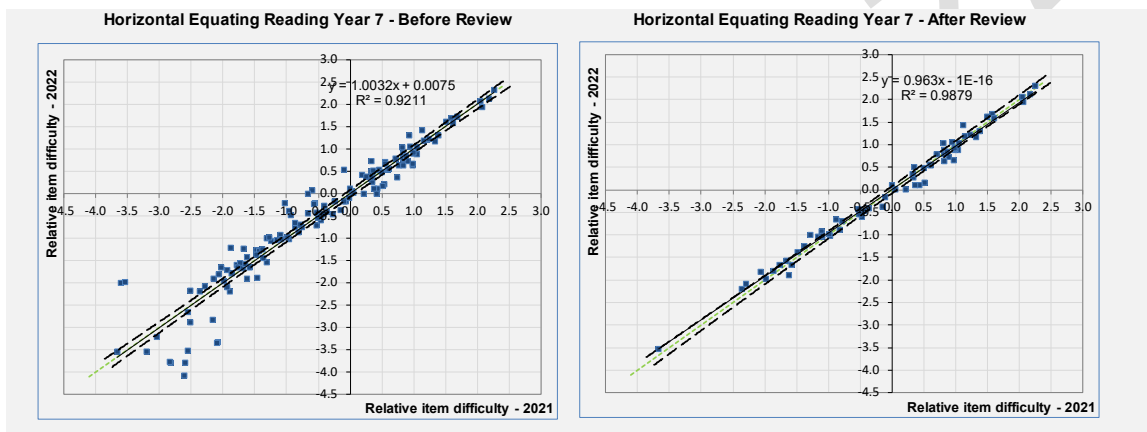


Figure 37: Scatterplot of reading, horizontal equating items between 2022 and 2021 for Year 7 online students

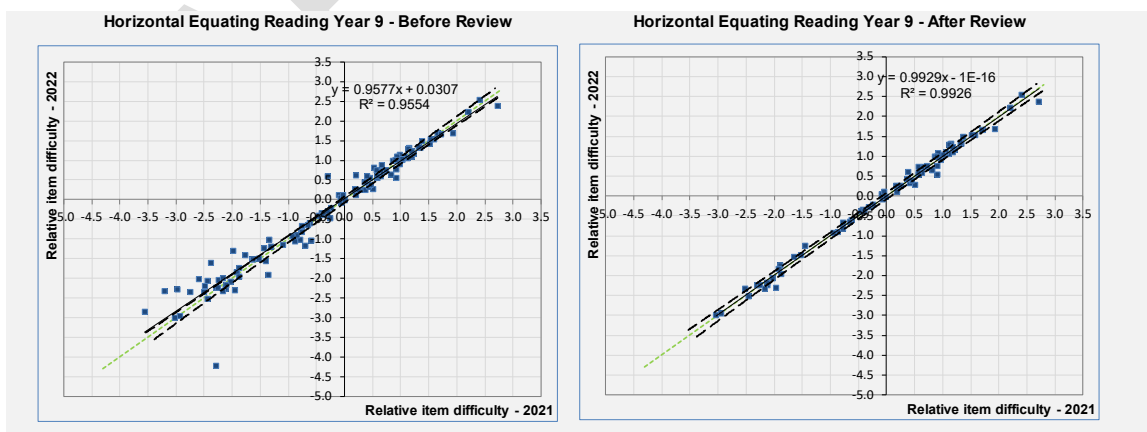


Figure 38: Scatterplot of reading, horizontal equating items between 2022 and 2021 for Year 9 online students

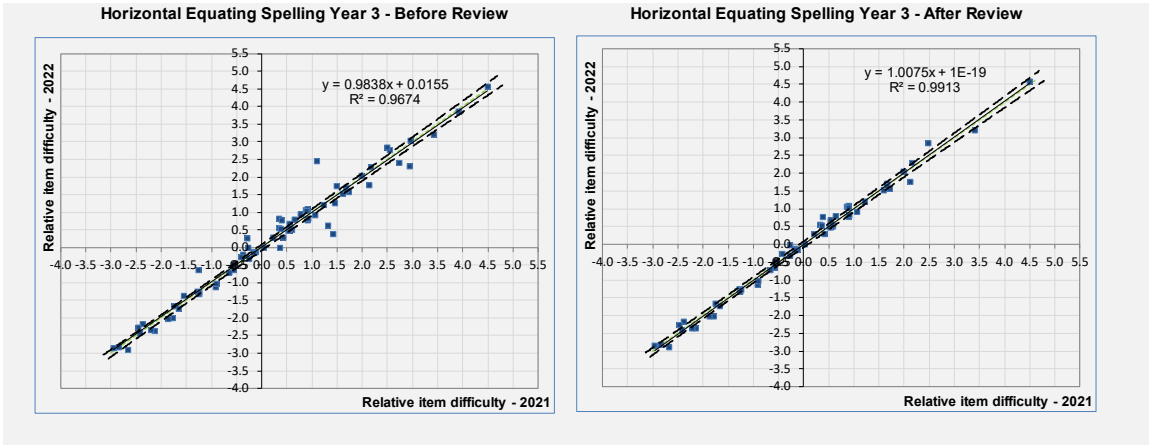


Figure 39: Scatterplot of spelling, horizontal equating items between 2022 and 2021 for Year 3 online students

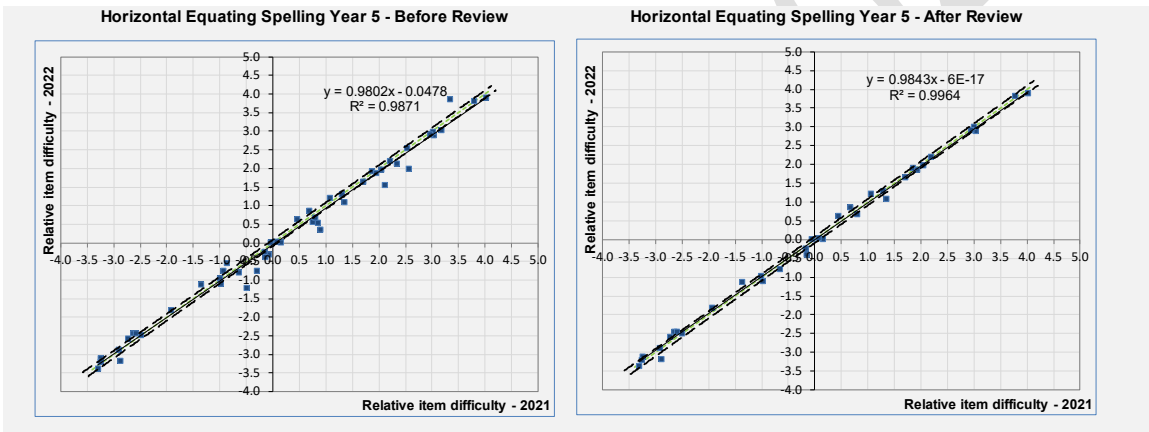


Figure 40: Scatterplot of spelling, horizontal equating items between 2022 and 2021 for Year 5 online students

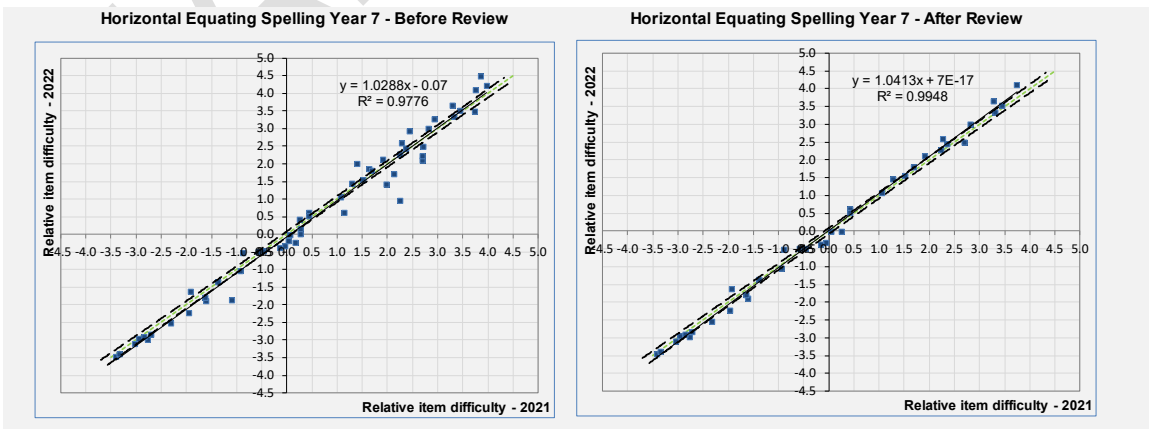


Figure 41: Scatterplot of spelling, horizontal equating items between 2022 and 2021 for Year 7 online students

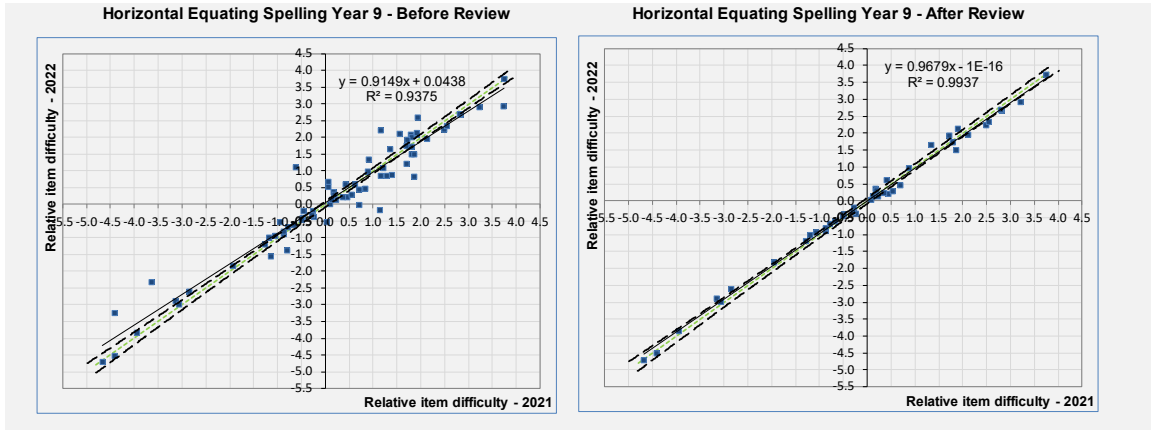


Figure 42: Scatterplot of spelling, horizontal equating items between 2022 and 2021 for Year 9 online students

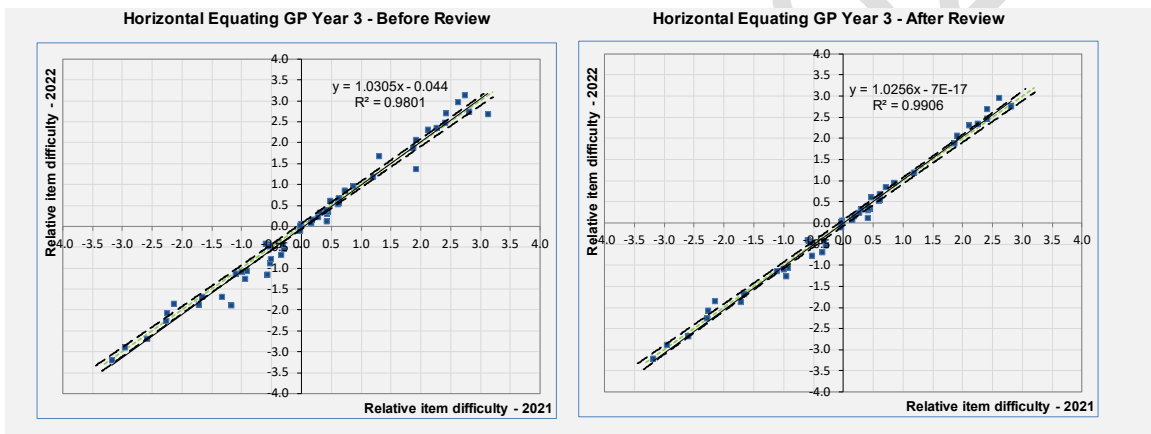


Figure 43: Scatterplot of grammar and punctuation, horizontal equating items between 2022 and 2021 for Year 3 online students

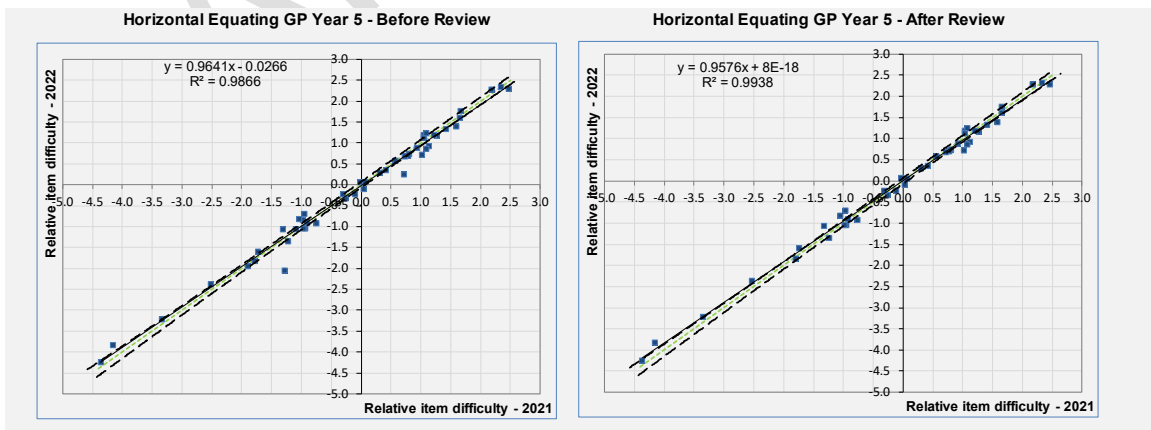


Figure 44: Scatterplot of grammar and punctuation, horizontal equating items between 2022 and 2021 for Year 5 online students

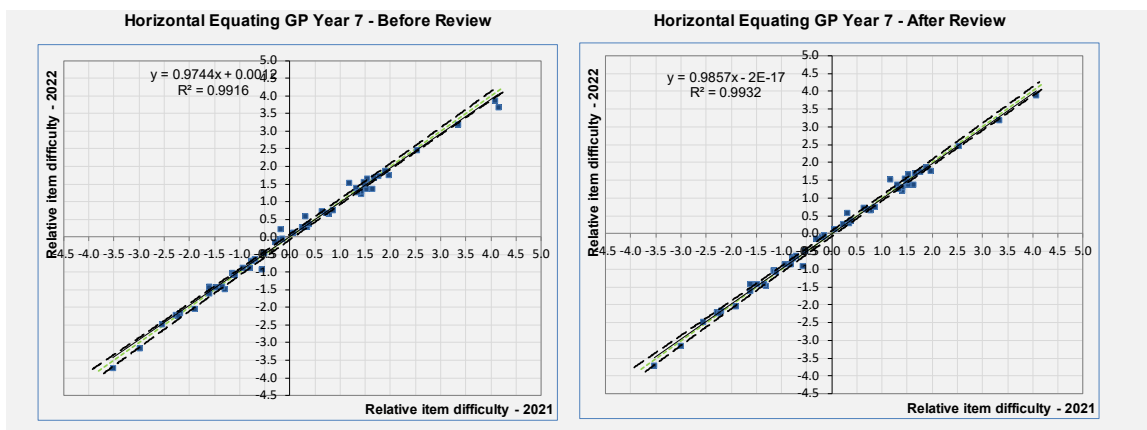


Figure 45: Scatterplot of grammar and punctuation, horizontal equating items between 2022 and 2021 for Year 7 online students

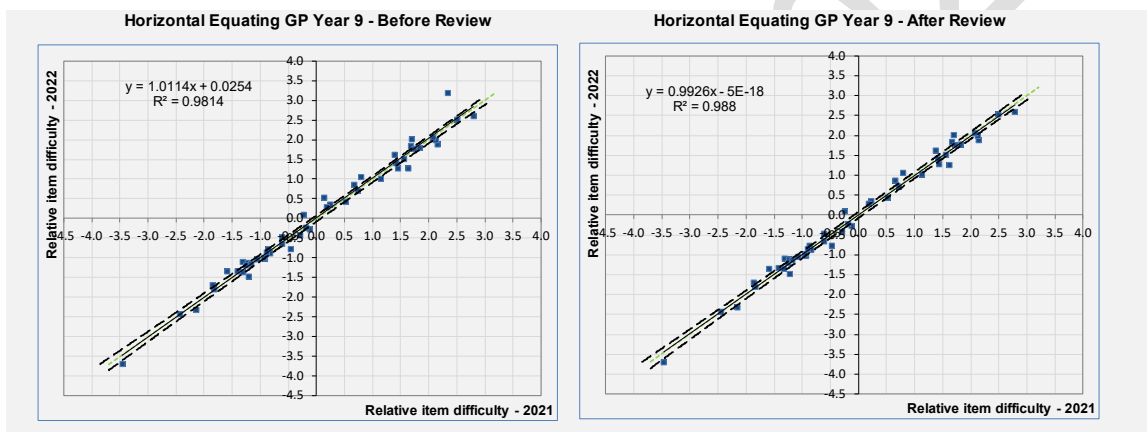


Figure 46: Scatterplot of grammar and punctuation, horizontal equating items between 2022 and 2021 for Year 9 online students

After the review and evaluation of the equating items between the 2022 and 2021 online tests, a final set of link items was identified for each domain and year level. The final sets of link items were used to calculate the preliminary horizontal shifts from 2022 to 2021.

In previous years, except for the 2021 online tests, horizontal-vertical regression (HVR) adjustment shifts were used as the final shifts to equate the NAPLAN tests to the NAPLAN historical scales. In 2021, HVR shifts were reviewed and found to be unnecessary for equating the online tests. In 2022, vertical equating and HVR shifts were again reviewed as a quality assurance check, and were found to be unnecessary for equating the online tests. This is due to the fact that in the current online test design, a much higher number of link items can be included with a wider range of item parameters compared to the earlier paper test design. Therefore, the preliminary horizontal shifts were the final shifts used to place the 2022 tests onto 2021 delta-centred scales. Then, the parameters that were used to equate the 2021 online tests to the NAPLAN historical scales were applied, placing the 2022 online tests onto the NAPLAN historical scales. The number of horizontal link items used and retained for each online test are shown in Table 77 and the horizontal shift-constants for each domain at each year level are summarised in Table 78.

Appendix N presents the 2022 horizontal link item locations (Rasch difficulty parameters), standard errors, and differences in the item locations by domain and year level.

Table 77: Horizontal link review summary for online tests

Year level	Numeracy Retained/ total	Reading Retained/ total	Spelling Retained/ total	Grammar and punctuation Retained/ total
3	67/76	59/79	56/73	45/53
5	69/78	35/82	35/49	42/47
7	86/107	67/136	37/55	52/54
9	86/105	73/110	42/72	47/49

Table 78: Horizontal equating shifts (Shift22to21) between 2022 item locations and 2021 item locations by year level by domain for online tests

Year level	Numeracy	Reading	Spelling	Grammar and punctuation
3	-0.034	-0.058	-0.050	-0.171
5	0.149	-0.101	0.068	0.054
7	0.191	-0.128	0.354	-0.001
9	0.025	0.072	-0.129	-0.184

Equating paper tests

There were 1570 students from 44 schools across 6 TAAs who completed the 2022 NAPLAN paper tests and who were originally designated to the paper test mode. In addition, there were some other students who completed the 2022 NAPLAN paper tests for various reasons. In total, there were approximately 4700 students, including absent, withdrawn and exempt students and non-attempt students, as well as mixed-mode¹ students across 4 NAPLAN year levels. It was not viable to construct a representative sample from this group of students in the paper test mode. Therefore, it was not possible to follow the normal procedure to calibrate and to equate the 2022 paper tests onto the NAPLAN historical scales.

The process of locating the 2022 paper tested students on the NAPLAN scales was as follows:

- equate to the historical NAPLAN scale
- align the 2022 results for students from the 44 schools testing on paper to their 2021 results, by domain and year level

The parameters used for locating 2022 paper test scale scores on the 2021 scale are given in Table 79. The transformations used in 2021 were then applied to the paper tested students to put them onto the NAPLAN historical scales.

¹ A mixed-mode student means a student who sat some domains online and some domains in paper test form.

Table 79: Parameters for locating 2022 paper test scales on the 2021 scales by year level and domain

Parameter	Year level	Numeracy	Reading	Spelling	Grammar and punctuation
$a_p = \frac{SD_{2021}}{SD_{2022}}$	3	0.82512	0.91622	0.52239	0.82207
	5	0.89636	0.73060	0.47014	0.81573
	7	0.92074	0.97639	0.76826	1.00030
	9	0.82095	0.89283	0.53272	0.70351
$b_p = Mean_{2021} - a_p * Mean_{2022}$	3	51.52510	49.64092	223.78345	87.24950
	5	52.18902	168.55309	294.34083	117.09080
	7	54.55906	19.84422	147.13304	-2.70506
	9	111.51297	75.61775	296.13356	205.17613
c_p	3	-2.61436	0.58576	-3.35488	0.50522
	5	-6.59062	-1.12926	0.55773	-3.38744
	7	-3.38806	0.80291	-0.63533	0.82371
	9	-2.24848	1.22629	-2.38120	1.65550

Scaling factors

Applying a scaling factor is sometimes necessary due to the potential impact that differences in test reliability can have on the spread of student scores. As the NAPLAN tests measure the same construct within a domain, it is expected to result in the same latent distribution for the same group of students. In this case, the scaling factor would be very close to 1. However, due to differences in NAPLAN test reliabilities, between the current year test and previous year test, between the tests across year levels, or historically between the equating test and the NAPLAN test, the spreads of scores between samples of 2 equated tests can be quite different for some year levels and domains. The scaling factor is defined as the standard deviation ratio between the 2 tests being equated. In 2022, no scaling factors were required to be applied.

Equating of writing results

Instead of applying an equating shift from the current scale to the historical scale, the anchoring method was used for equating writing to the historical scale. Before anchoring the item (criterion) difficulties to their historical values, the appropriateness of this method was assessed in 2 ways. First, the relative item difficulty steps were compared with those from 2021. Second, achievement drift caused by any systematic changes in marking over time was examined.

To review the stability of item difficulty steps, the 2022 writing data were freely calibrated and compared to the item difficulties of the 2021 online tests since the writing genre was narrative in both 2022 and 2021. The scatterplot between the 2 calendar years is shown in Figure 47. They indicate that the consistency of relative difficulties was supported using the anchoring method in 2022.

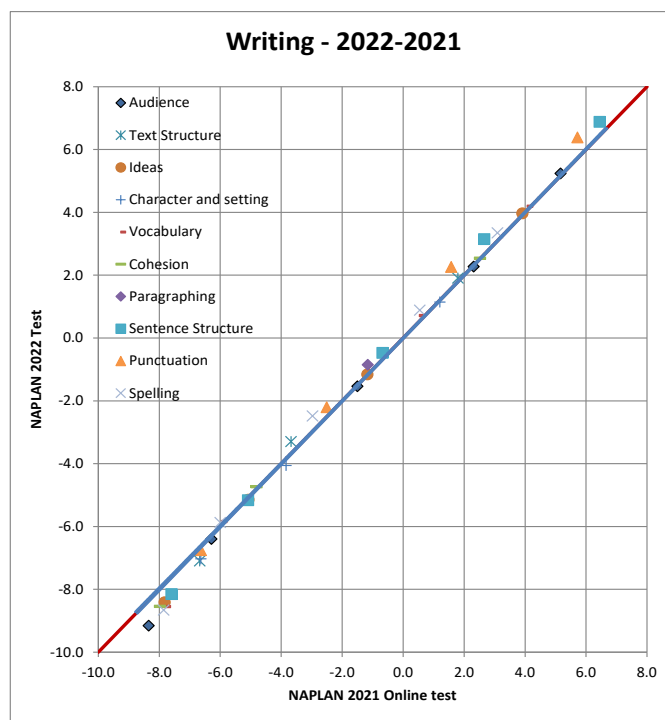


Figure 47: Scatterplot for writing criteria between 2022 and 2021 online and paper tests

In addition to comparing relative item difficulties, an equating verification study was conducted using pairwise comparisons of scripts to investigate if a shift in marking may have occurred. More information about the pairwise comparison methodology can be found in Humphry and McGrane (2015).

The pairwise study for writing in the NAPLAN 2022 assessment created a 2022 scale for 2022 and 2021 writing scripts. Using this scale, performance estimates could be compared directly to the 2021 scale estimates. The comparison of these 2 scales provides direct evidence on the validity of continuing the identity equating, which has been used for several years. This comparison of scales forms a key component of the equating verification.

The equating design involved internal comparisons of all 297 writing samples from 2022 provided by the TAAs. The breakdown of 2022 scripts per task is: 69 paper task 1, 70 online task 1, 158 online task 2. These scripts were selected using an approximately uniform score distribution. Roughly equal numbers of scripts were used for the 2 task groups (paper task 1 + online task 1, and online task 2).

Comparisons of 2021 scripts against 2022 scripts were judged using all 297 sampled scripts from 2022 and 282 scripts from 2021 (of the 2021 scripts, 72 were paper form tasks and 210 were online tasks).

For the 2022 pairwise equating project, 35 judges compared 27,274 pairs of scripts in total. Of these, there were 13,492 comparisons of 2022 against 2022 scripts, and 13,782 comparisons of 2022 against 2021 scripts. There were 828 comparisons made between 2021 paper scripts and 2022 paper scripts. The statistical fit index termed outfit mean square was used to test whether or not each judge agreed (on aggregate) with the consensus of all judges. All judges had outfit values of less than 1.46, indicating good consistency of judgements across the set of judges.

A joint 2016/2021/2022 pairwise scale was formed by adding comparisons from the previous NAPLAN writing pairwise project to the judgments from the 2022 pairwise project for the purpose of calculating the equating error. In total, 52,846 comparisons were analysed using the Bradley-Terry-Luce model to form this scale.

The purpose of the pairwise study is to ascertain whether there are differences in rubric marks that are inconsistent with the results of direct comparisons of scripts. Specifically, the design allows evaluation of whether, for a given scale location based on pairwise comparisons, a similar rubric score is predicted for 2021 and 2022 scripts or whether different scores are predicted from the common pairwise scale. For this objective, paper and online components are compared across years separately, as well as both components combined. Thus, the purpose of the pairwise study is to obtain a common frame of reference by which to compare marking in 2021 with marking in 2022 (paper and online), and in addition to reference to the 2016 scale. In particular, the objective is to examine whether there is evidence for differences in marker harshness that might affect the comparability of results.

It is noted that in the procedure, prompts are selected to minimise task effects to the extent possible. It is also noted that exemplars are used in the writing marking guide to help anchor score points over time.

Pairwise study results

To evaluate fit to the Bradley-Terry-Luce model used to analyse the data, judge outfit indices were calculated after removing extreme observations (comparisons for which the standardised residuals were greater than 7). For the 2022 pairwise study, all judges had good outfit indices (less than 1.46).

Figure 48 shows the pairwise scale locations of the 2021 scripts, comparing the estimates from the 2021 project (y-axis) and the estimates from the 2022 project (x-axis). Figure 48 shows a very strong linear correspondence between 2021 estimates and 2022 estimates, with points scattered very close to the identity line, indicating excellent comparability of the scales. The correlation between 2021 estimates and 2022 estimates is greater than 0.99. The regression model indicates almost perfect correspondence in the spread of the scale between the 2 equating years. The correspondence shows that 2021 script locations are essentially the same whether based on 2022 paired comparisons or based on comparisons used in 2021. This means the locations were robust over time and formed a strong basis for checking marking consistency in 2022.

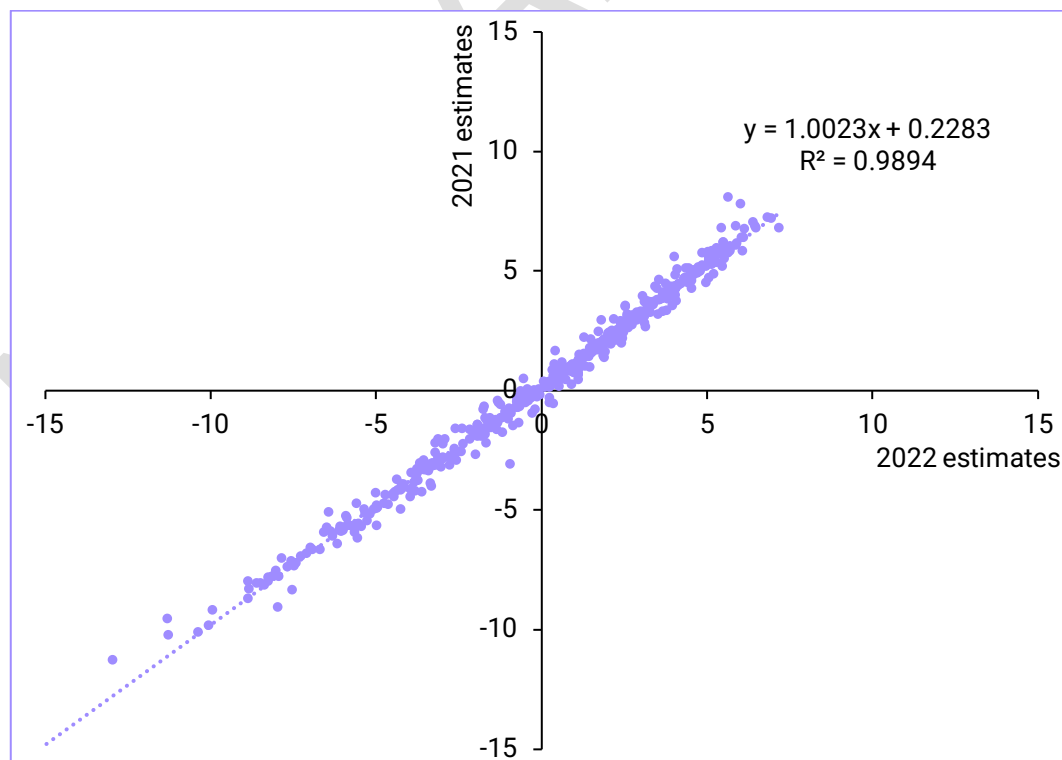


Figure 48: Pairwise location estimates from the 2021 project plotted against the estimates from the 2022 project for the 2021 scripts

Figure 49 shows the pairwise scale locations (x-axis) plotted against the NAPLAN rubric locations (y-axis) for the 2021 and 2022 scripts, across writing tasks. The pairwise scale locations show the ordering of the scripts based on direct comparisons, whereas the NAPLAN scale locations are based on rubric marking. The overall linear correlation between the pairwise and rubric locations is 0.92.

The fitted curves in Figure 49 are somewhat curvilinear as in previous years of the program and show a very close relationship between the 2021 scripts and the 2022 scripts. Rubric locations for the 2022 performances are based on the same correspondence table, between raw scores and logits, as the rubric locations for the 2021 performances. The close agreement of both fitted curves and data points for the 2021 data and 2022 data provide evidence that marking in 2022 was consistent with marking in 2021.

The correlation and nature of the relationship are relatively similar for both of these calendar years to the relationship observed in previous calendar years of NAPLAN.

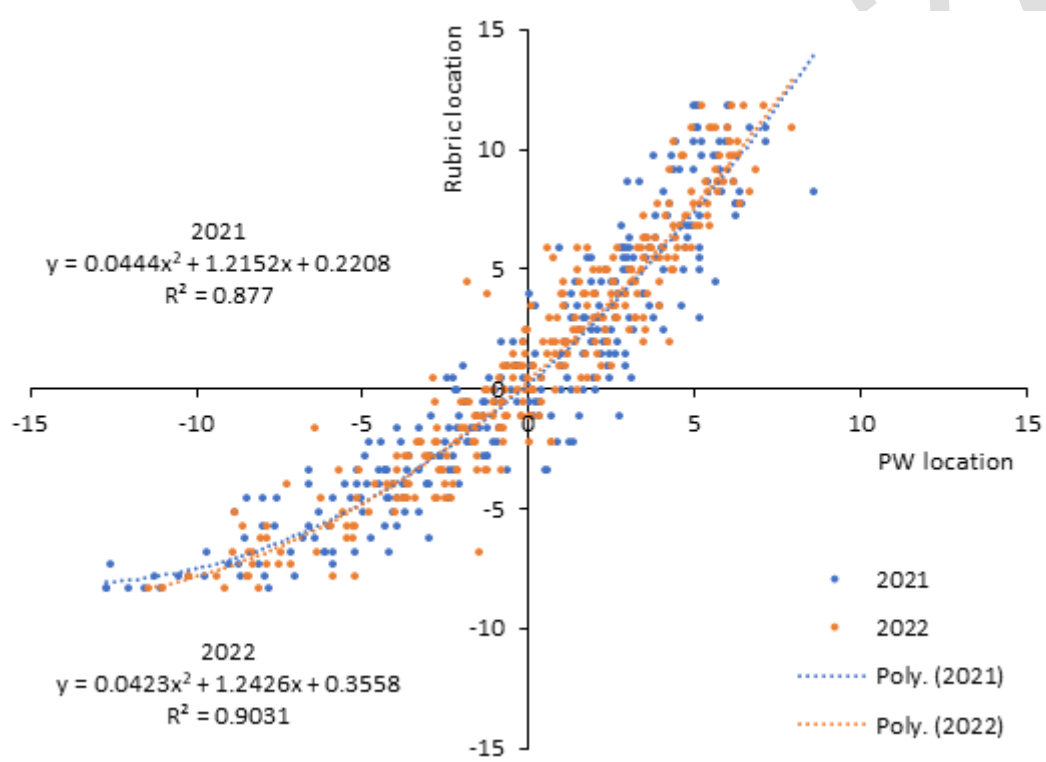


Figure 49: Rubric location estimates plotted against the pairwise location estimates from the 2022 project for the 2021 and 2022 scripts.

Figure 50 shows the pairwise scale locations (x-axis) plotted against the NAPLAN rubric locations (y-axis) for the 2021 and 2022 paper scripts only. The data points in Figure 50 show the locations of the Year 3 scripts only, as Year 3 students only completed the paper-based assessment in 2022. The data points for the 2021 scripts are shown in a different colour to the data points for the 2022 scripts, and separate regression curves are shown.

It can be seen in Figure 50 that the regression curves diverge somewhat with increasing pairwise location. There are relatively few data for this comparison (72 paper scripts from 2021 and 69 paper scripts from 2022). On average, 2022 paper scripts were marked somewhat higher in this sample but the difference is not statistically significant (using $p = 0.05$) when outliers are omitted. Statistical significance was tested with the equating method and slope estimation.

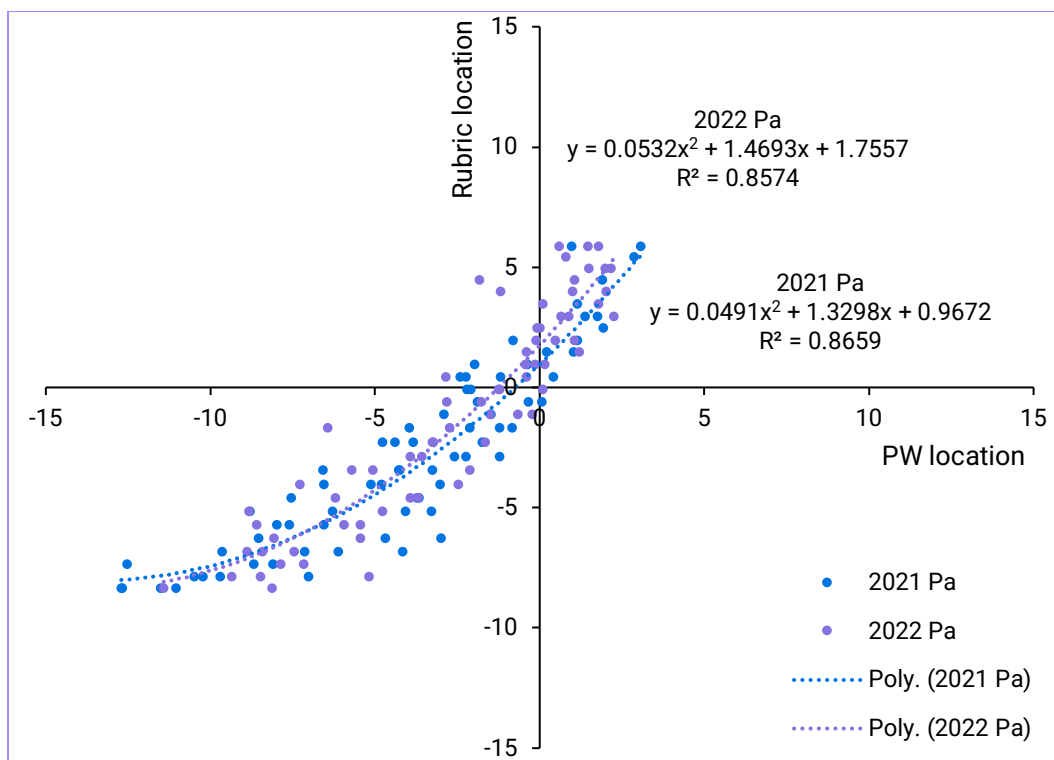


Figure 50: Rubric location estimates plotted against the pairwise location estimates from the 2022 project for the 2021 and 2022 year 3 paper scripts.

Overall, the 2022 pairwise study showed that for the selected sample, rubric scores in 2022 are highly consistent with rubric scores in 2021. The results showed that 2021 and 2022 performances scaled together well to form a single scale, which indicates that there is a common pairwise scale. Figure 49 shows that for any given location along the paired comparison scale (x-axis) the predicted rubric scores for 2021 and 2022 are highly similar, and the distribution and range of actual rubric scores is generally similar. Together these observations imply that with the pairwise comparison scale as a reference, marking for the selected sample was consistent across 2022 and 2021.

Summary of equating parameter estimates for NAPLAN 2022

In 2022, the NAPLAN online and paper results were first placed onto the 2021 delta-centred scales separately, then the 2-step formula, which was used to place the 2021 online results onto the NAPLAN historical scales, was used to place the 2022 results onto the NAPLAN historical scales as below:

Online results:

$$\theta_{22on21}^* = \theta_{22} + Shift_{22to21} \quad (5)$$

$$\theta_{22on19}^* = SF_{21}(\theta_{22on21}^* - LM_{21}) + LM_{21} + Shift_{21to19} \quad (6)$$

$$\theta_{22}^* = 100 * (SF_{19}(\theta_{22on19}^* - LM_{19}) + LM_{19} + HVR_{19} - MN_{\theta_{08}}) / SD_{\theta_{08}} + 500 \quad (7)$$

Where θ_{22on21}^* is the equated 2022 achievement score onto the 2021 scale, θ_{22on19}^* is the equated 2022 achievement score onto the 2019 scale, θ_{22}^* is the equated 2022 achievement score onto the NAPLAN historical scale, θ_{22} is the 2022 original achievement score in logits, SF_{21} and SF_{19} are the scaling factors applied to the online tests in 2021 and 2019, LM_{21} and LM_{19} are the local means of the online tests in 2021 and 2019, $Shift_{22to21}$ and $Shift_{21to19}$ are the 2022 to 2021 horizontal shift and 2021 to 2019 horizontal shift, HVR_{19} the 2019 shift for the online tests, $MN_{\theta_{08}}$

the mean achievement in logits of all students in 2008, and $SD_{\theta_{08}}$ the standard deviation in logits of all students in 2008.

For selected domains and year levels, these procedures were followed by equipercentile equating, using the formula

$$\theta_{22}^{**} = a + b * (\theta_{22}^*)^2 + c * \theta_{22}^* \tag{8}$$

Paper results:

$$\theta_{22on21}^* = \theta_{22} + Shift_{22to21} \tag{9}$$

$$\theta_{22}^* = 100 * (SF_{21}(\theta_{22on21}^* - LM_{21}) + LM_{21} + Shift - MN_{\theta_{08}}) / SD_{\theta_{08}} + 500 \tag{10}$$

Where θ_{22on21}^* is the equated 2022 achievement score onto the 2021 scale, θ_{22}^* is the equated 2022 achievement score onto the NAPLAN historical scale, θ_{22} the original achievement score in logits, SF_{21} the scaling factor of the paper test, LM_{21} the local mean, $Shift_{22to21}$ the shift from 2022 to 2021, $Shift$ the 2021 shift for the paper tests, $MN_{\theta_{08}}$ the mean achievement in logits of all students in 2008, and $SD_{\theta_{08}}$ the standard deviation in logits of all students in 2008. These procedures were followed by regression of θ_{22}^{**} on θ_{22}^* using the formula

$$\theta_{22}^{**} = a_p * \theta_{22}^* + b_p + c_p \tag{11}$$

Where a_p , b_p and c_p are defined as per Table 79.

Parameters for transforming the 2022 online and paper scores to NAPLAN reporting scales are presented in Table 80 and Table 81 respectively.

Table 80: Summary of parameters for transforming the 2022 online logit scores to the NAPLAN reporting scales

Domain & year	Shift _{22to21}	LM ₂₁	SF ₂₁	Shift _{21to19}	LM ₁₉	SF ₁₉	HVR ₁₉	MN _{θ₀₈}	SD _{θ₀₈}	a	b	c
N3	-0.034	0.12854	1.00000	0.08946	0.2832	1.0293	-1.0910	0.8102	1.6652	100.49140	0.00048	0.56178
N5	0.149	0.26065	1.00000	0.02940	0.2744	0.8408	0.3039	0.8102	1.6652	145.33553	0.00058	0.42416
N7	0.191	0.23353	0.95819	-0.30456	-0.0116	0.9673	1.6987	0.8102	1.6652	0.00000	0.00000	1.00000
N9	0.025	0.06790	1.00000	-0.32354	-0.2495	0.9782	2.4713	0.8102	1.6652	253.30670	0.00035	0.36937
R3	-0.058	0.08570	1.00000	-0.12757	-0.1172	1.1951	0.1399	1.1629	1.4867	134.01356	0.00059	0.43734
R5	-0.101	0.25824	1.00000	-0.02776	0.1707	0.9642	1.0718	1.1629	1.4867	0.00000	0.00000	1.00000
R7	-0.128	0.07500	1.00000	-0.09311	0.0485	0.9742	1.7694	1.1629	1.4867	0.00000	0.00000	1.00000
R9	0.072	0.23665	1.00000	-0.29216	-0.0411	1.1291	2.3102	1.1629	1.4867	35.61401	-0.00015	1.03585
S3	-0.050	-0.02228	1.00000	0.50267	0.3575	0.9877	-1.6518	0.9406	2.6241	0.00000	0.00000	1.00000
S5	0.068	0.07738	0.97071	0.60279	0.4816	1.0624	0.2715	0.9406	2.6241	114.13480	0.00018	0.69006
S7	0.354	0.29852	1.00000	0.25374	0.5037	0.9068	1.5922	0.9406	2.6241	0.00000	0.00000	1.00000
S9	-0.129	-0.08522	1.00000	0.23621	0.2447	0.9209	2.8255	0.9406	2.6241	0.00000	0.00000	1.00000
G3	-0.171	-0.00411	1.15198	1.10486	0.0000	1.0000	-0.7518	1.2529	1.3605	95.63552	0.00041	0.58622
G5	0.054	0.24575	1.00000	0.85955	0.0000	1.0000	0.2612	1.2529	1.3605	125.42115	0.00053	0.46176
G7	-0.001	0.02195	1.00000	0.66904	0.0000	1.0000	0.9034	1.2529	1.3605	121.30717	0.00043	0.55426
G9	-0.184	-0.02288	1.00000	0.61441	0.0000	1.0000	1.7331	1.2529	1.3605	-25.55819	-0.00012	1.10480
W3	0.000	0.00000	1.00000	0.00000	0.0000	1.0000	0.0000	1.1160	3.3679	0.00000	0.00000	1.00000
W5	0.000	0.00000	1.00000	0.00000	0.0000	1.0000	0.0000	1.1160	3.3679	0.00000	0.00000	1.00000

Domain & year	Shift _{22to21}	LM ₂₁	SF ₂₁	Shift _{21to19}	LM ₁₉	SF ₁₉	HVR ₁₉	MN ₀₈	SD ₀₈	a	b	c
W7	0.000	0.00000	1.00000	0.00000	0.0000	1.0000	0.0000	1.1160	3.3679	0.00000	0.00000	1.00000
W9	0.000	0.00000	1.00000	0.00000	0.0000	1.0000	0.0000	1.1160	3.3679	0.00000	0.00000	1.00000

Table 81: Summary of parameters for transforming the 2022 paper logit scores to the NAPLAN reporting scales

Domain & year	LM ₂₁	SF ₂₁	Shift	MN ₀₈	SD ₀₈
N3	0.04375	1.10615	-0.782	0.8102	1.6652
N5	0.30951	0.92438	0.58969	0.8102	1.6652
N7	0.53365	1	1.25462	0.8102	1.6652
N9	0.41191	0.81621	1.94652	0.8102	1.6652
R3	0.71441	1.12442	-0.37957	1.1629	1.4867
R5	0.76304	1.02541	0.6934	1.1629	1.4867
R7	0.38195	1	1.55132	1.1629	1.4867
R9	0.76346	1	1.62156	1.1629	1.4867
S3	-0.0537	1.04731	-0.79029	0.9406	2.6241
S5	0.17112	0.9096	0.86677	0.9406	2.6241
S7	0.29636	1	2.18209	0.9406	2.6241
S9	-0.06482	1	3.41619	0.9406	2.6241
G3	0.34426	1.22924	0.10715	1.2529	1.3605
G5	0.3544	1	1.08651	1.2529	1.3605
G7	0.31983	1.07673	1.46936	1.2529	1.3605
G9	0.62801	1	1.78709	1.2529	1.3605
W3	0	1	0	1.116	3.3679
W5	0	1	0	1.116	3.3679
W7	0	1	0	1.116	3.3679
W9	0	1	0	1.116	3.3679

Estimating equating errors

As with all statistics, equating shifts have an associated level of uncertainty. Had a different set of link items been chosen, the equating shifts would have been slightly different. As a consequence, there is an uncertainty associated with the equating, which is due to the choice of link items, similar to the uncertainty associated with the sampling of schools and students.

The uncertainty that results from the selection of a subset of link items is referred to as *equating error*. This error should be taken into account when making comparisons between the results from different data collections across time (see Chapter 9). The exact magnitude of the equating error cannot be determined. We can, however, estimate the likely range of magnitudes for this error and take this error into account when interpreting results. As with sampling or measurement errors, the likely range of magnitude for the combined errors is represented as a standard error of each reported statistic.

In 2022, 2 sets of equating errors were determined for comparing student achievement for numeracy, reading, spelling, and grammar and punctuation: the equating error between 2022 and 2021, and between 2022 and the base year. The equating of 2022 NAPLAN tests was through the 2021 NAPLAN online tests. Hence, the equating error between 2022 and the base year was a combination of the equating error between 2022 and 2021 online tests and the equating errors between 2021 and the base year that were estimated in 2021 (ACARA, 2022), with the assumption that they are independent.

The errors considered in the equating processes over the course of the program are shown in Figure 52.

Figure 51: A schematic of the equating errors accumulated across NAPLAN administrations

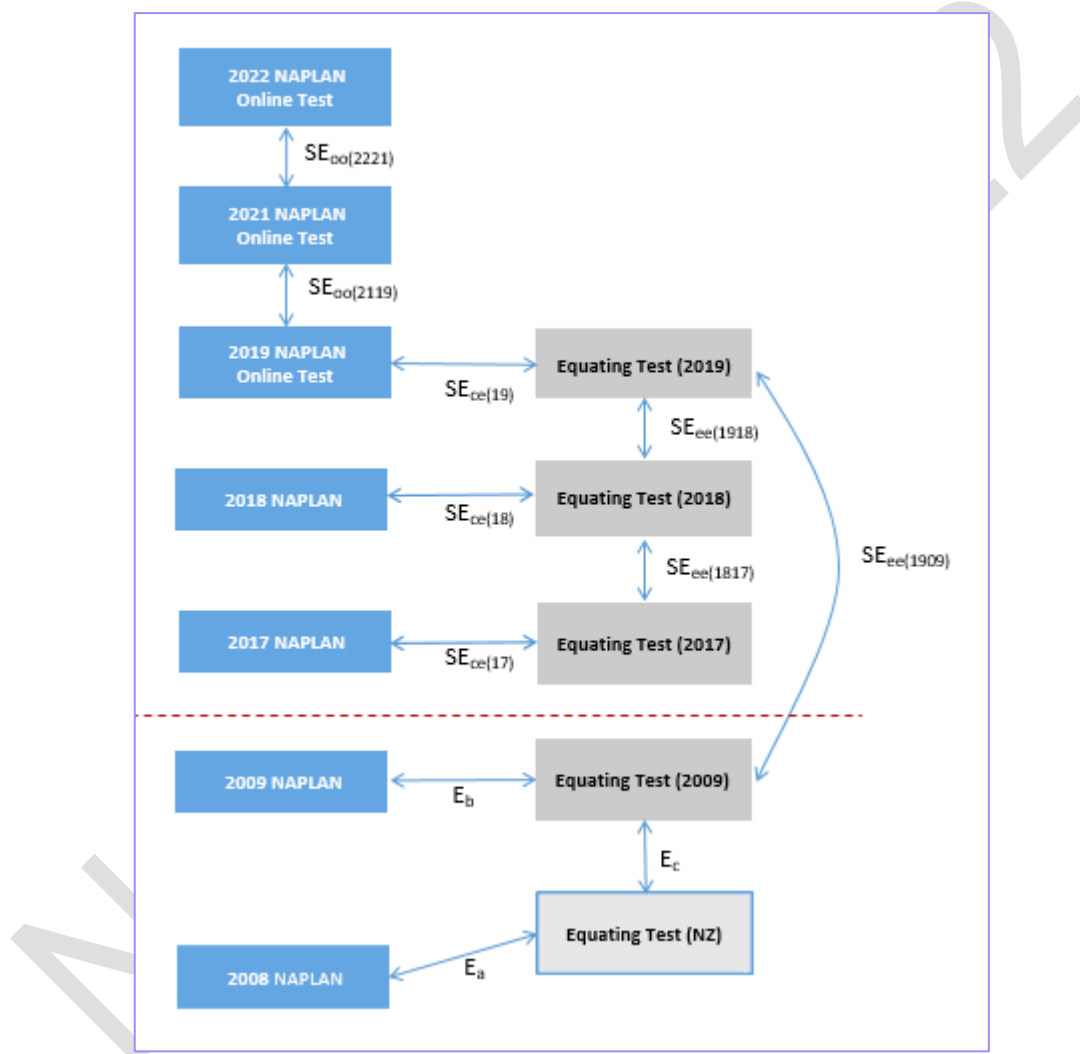


Figure 52: A schematic of the equating errors accumulated across NAPLAN administrations

For each domain and year level except writing:

- E_a is the standard error associated with equating the offshore equating test and the 2008 NAPLAN test
- E_b is the standard error associated with equating the onshore equating test and the 2009 NAPLAN test
- E_c is the standard error associated with equating the offshore and onshore equating tests; E_a , E_b and E_c were determined during 2009 equating process

- $SE_{ce(xx)}$ is the standard error associated with equating the NAPLAN 20xx test with the equating test (calibration to equating), xx stands for 17, 18 or 19
- $SE_{ee(1918)}$ is the standard error associated with equating the 2019 and 2018 administrations of the equating test (equating to equating), and so forth;
- $SE_{oo(2119)}$ is the standard error associated with equating the NAPLAN 2021 online test and the NAPLAN 2019 online test (equating to equating)
- $SE_{oo(2221)}$ is the standard error associated with equating the NAPLAN 2022 online test and the NAPLAN 2021 online test (equating to equating).

For reporting results of NAPLAN 2022, the equating errors for equating the 2022 scales to the base year (2008) scales were estimated by combining the relevant standard errors as follows:

$$SE_{2022tobase} = \sqrt{(SE_{2022to2021})^2 + (SE_{2021tobase}^{online})^2} \quad (122)$$

The equating errors between 2022 and 2021 were estimated taking the clustering of items in units into account as follows. Suppose we have a total of L score points in the link items in K modules. Use i to index items in a unit and j to index units so that $\hat{\delta}_u^y$ is the estimated difficulty of item i in unit j for year y , and let:

$$c_{ij} = \hat{\delta}_{ij}^{2022} - \hat{\delta}_{ij}^{2021} \quad (133)$$

The size (number of score points) of unit j is m_j so that:

$$\sum_{j=1}^K m_j = L \quad \text{and} \quad (144)$$

$$\bar{m} = \frac{1}{K} \sum_{j=1}^K m_j \quad (155)$$

Further let:

$$c_{\cdot j} = \frac{1}{m_j} \sum_{i=1}^{m_j} c_{ij} \quad \text{and} \quad (166)$$

$$\bar{c} = \frac{1}{N} \sum_{j=1}^K \sum_{i=1}^{m_j} c_{ij}$$

and then the link error, taking into account the clustering is as follows:

$$SE_{2022to2021}^{oo} = \sqrt{\frac{\sum_{j=1}^k m_j^2 (c_{.j} - \bar{c})^2}{K(K-1)\bar{m}^2}} = \sqrt{\frac{\sum_{j=1}^k m_j^2 (c_{.j} - \bar{c})^2}{L^2} \frac{K}{K-1}}$$

(177) Table 82 shows the standard errors of equating associated with each test domain and year level in logits and in scale scores. The scale scores were transformed from the logit values, by applying the factors from formula (6); that is, the 2008 standard deviation and 100.

Table 82: Standard errors of equating

Domain	Year	Logit		Scale score	
		2022 to base*	2022 to 2021	2022 to base*	2022 to 2021
Numeracy	3	0.0761	0.0173	4.5697	1.0387
	5	0.0799	0.0170	4.7957	1.0189
	7	0.0597	0.0157	3.5854	0.9429
	9	0.0607	0.0141	3.6446	0.8476
Reading	3	0.0730	0.0190	4.9128	1.2806
	5	0.0753	0.0396	5.0683	2.6633
	7	0.0681	0.0305	4.5837	2.0509
	9	0.0661	0.0162	4.4486	1.0910
Spelling	3	0.1050	0.0204	4.0027	0.7762
	5	0.1133	0.0228	4.3193	0.8684
	7	0.1158	0.0314	4.4119	1.1967
	9	0.1149	0.0262	4.3772	0.9969
Grammar and punctuation	3	0.1047	0.0232	7.6922	1.7050
	5	0.1144	0.0218	8.4118	1.6014
	7	0.1027	0.0189	7.5497	1.3889
	9	0.0958	0.0240	7.0382	1.7608
Writing**	3579	0.1680	0.1684	4.9881	5.0006

* The base year for reading, spelling, grammar & punctuation, and numeracy is 2008; base year for writing is 2011.

** The writing equating error was calculated based on the pairwise equating data in a manner consistent with keeping the item parameters constant.

The equating errors were taken into account, together with sampling and measurement errors, in estimating the standard errors used to determine statistical significance in the comparisons between mean scores across years in NAPLAN reports. The equating errors are not included when estimating standard errors of estimates used to determine statistical significance in the

comparisons between mean scores of different subgroups within NAPLAN 2022. This is further explained in Chapter 9.

Estimates of standard errors of equating for percentages of students at or above minimum standards in different calendar years required a different estimation process and were not calculated as part of producing summary statistics in the central analysis process.

Further details regarding the application of standard errors to testing the statistical significance of performance differences are given in Chapter 9.

Estimating long-term trend errors

In 2022, the long-term trend was also estimated and tested for significance by domain and year level. As with other trend estimates, the standard errors for long-term trend estimates consist of error components due to sampling, measurement and equating. The long-term trend error was estimated by domain by grade using the following steps:

1. Build a variance-covariance matrix with sampling error, measurement error and equating error (the details on building the variance-covariance matrix can be found in Appendix O).
2. Build a vector with mean scores.
3. Draw 10,000 random vectors from a multivariate normal distribution with the mean score vector and the variance-covariance matrix built from step 1 and 2.
4. Fit regression models (either linear or quadratic model) for each simulated vector; the variance of slopes of regressions is the long-term trend error variance.

Chapter 8: NAPLAN proficiency bands

The main feature of the Rasch model is the placement of items and students on the same scale. A student with an achievement score equal to the difficulty of an item has 50% chance of responding correctly to that item. Consequently, a student has more than 50% chance of responding correctly to easier items and less than 50% to harder items. In other words, a student masters the skills that are needed to respond correctly to items with difficulties below their achievement scores. This scale has a response probability of 0.50 (RP50).

This feature enables construction of proficiency bands on the measurement scale in such a way that the items in a band describe the skills of the students in that same band. To be able to conclude that students master the skills within a band, however, the item difficulties need to be shifted up the scale so that every student within a band is likely to respond correctly to at least 50% of the items within the same band. The method to create these bands consists of 2 steps:

1. shift item difficulties upwards on the scale by changing the response probability
2. choose a width for the band so that students at the very bottom of a band are likely to respond correctly to 50% of the items in that band (and all other students to more than 50% of the items).

In 2008, a response probability of 0.62 (RP62) was chosen, which needs to be combined with a band width of 52 NAPLAN scale scores to satisfy the condition that all students in a band are expected to respond correctly to at least 50% of the items in the same band. It was decided to use the same cut scores between bands across all domains. Hence, the width of the bands in logits varies across domains. Table 83 shows the cut points between bands (lower bound) in scale scores and in logits.

Table 83: Lower bounds of proficiency bands in scale scores and in logits

Band	Scale score	Logits (RP50)				
	All domains	Numeracy	Reading	Writing	Spelling	Grammar
10	686	3.417	3.438	6.890	5.331	3.293
9	634	2.552	2.665	5.139	3.967	2.586
8	582	1.686	1.892	3.388	2.602	1.879
7	530	0.820	1.119	1.636	1.238	1.171
6	478	-0.046	0.346	-0.115	-0.127	0.464
5	426	-0.912	-0.427	-1.866	-1.491	-0.244
4	374	-1.778	-1.200	-3.618	-2.856	-0.951
3	322	-2.644	-1.973	-5.369	-4.220	-1.659
2	270	-3.510	-2.747	-7.120	-5.585	-2.366
Width	52	0.866	0.773	1.751	1.365	0.707

Once the proficiency bands were defined, the skills that students in each band mastered were described by reviewing the items with an RP62 difficulty located within each band. The descriptions of the bands are included in Table 84 to Table 87 for each domain.

Table 84: Described scale for numeracy

Proficiency band	Numeracy skills and knowledge
Band 10	Uses mathematical understanding to solve complex problems including those involving irrational numbers. Interprets and uses index notation. Evaluates algebraic expressions and solves equations and inequalities using a range of algebraic strategies. Solves surface area and volume problems using geometric reasoning or formulas. Calculates and compares numerical probabilities. Applies knowledge of line and angle properties to spatial problems.
Band 9	Solves complex reasoning problems. Uses square roots and powers. Evaluates algebraic expressions and solves equations and inequalities using substitution. Interprets simple linear graphs. Interrogates data and finds measures of centre. Calculates elapsed time across time zones. Determines angle size, area and volume of polygons and diameter and circumference of circles. Recognises congruence and uses similarity in regular shapes.
Band 8	Solves non-routine problems and compares common fractions, decimals and percentages. Continues linear patterns and identifies non-linear rules. Solves perimeter and area problems. Determines probabilities of outcomes of experiments. Classifies triangles and uses their properties. Identifies transformations of shapes and visualises changes to 3D objects. Determines direction using compass points and angles of turn.
Band 7	Solves multi-step problems involving relational reasoning. Calculates missing values in equations. Interprets rules and patterns and completes simple inequalities. Finds perimeters and areas of composite shapes. Calculates elapsed times across midday and midnight. Expresses probability as a fraction. Compares and classifies angles and solves problems involving nets. Uses scale to determine distance on maps.
Band 6	Applies appropriate strategies to solve multi-step problems, simple multiplication and division and patterning. Converts between familiar units of measure. Calculates durations of events. Interprets and uses data from a variety of displays. Recognises nets of familiar 3D objects and symmetry in irregular shapes. Uses simple legends and coordinate systems to interpret maps and grids.
Band 5	Solves routine problems using a range of strategies. Demonstrates knowledge of simple fractions and decimals. Continues number and spatial patterns. Uses familiar measures to estimate, calculate and compare area or volume. Reads graduated scales. Compares likelihood of outcomes in chance events. Recognises the effect of transformations on 2D shapes. Uses major compass points and follows directions to locate positions.
Band 4	Solves problems involving unit fractions, combinations of addition and subtraction of two-digit numbers and number facts to 10×10 . Identifies repeating parts of patterns. Interprets timetables and calendars and reads time on clocks to the quarter hour. Locates information in tables and graphs. Recognises familiar 2D shapes after a transformation and identifies a line of symmetry. Visualises 3D objects from different viewpoints.
Band 3	Solves single-step problems involving addition, subtraction or simple multiplication. Recognises representations of unit fractions and completes simple number sentences. Compares length and mass using familiar units of measure. Describes outcomes of simple chance events. Uses common features and properties to classify families of shapes and objects, and recognises symmetrical grid references.
Band 2	Compares and orders different representations of three-digit numbers. Applies addition and subtraction facts up to 20 to solve problems. Identifies equal groups of collections. Uses language of time and chance in familiar contexts. Visually compares area and locates information in simple tables. Recognises common features of positions on simple maps and plans by following directions.

Proficiency band	Numeracy skills and knowledge
Band 1	Uses counting strategies to solve problems and demonstrates knowledge of place value of three-digit numbers. Identifies the next term in a simple pattern. Interprets tally marks. Recognises and compares length and mass of familiar objects. Names common 2D shapes and familiar 3D objects and shows some understanding of spatial positioning.

Table 85: Described scale for reading

Proficiency band	Reading skills and knowledge
Band 10	Analyses and critically evaluates aspects of complex texts to recognise an author's purpose and stance, and to identify an underlying message, subtle character traits, tone and point of view.
Band 9	Evaluates and processes implicit ideas in a range of complex narrative and informative texts and interprets complex vocabulary. Analyses and evaluates key evidence in persuasive texts. Identifies language and text features to infer an author's intended purpose and audience.
Band 8	Interprets ideas and processes information in a range of complex texts. Analyses how characters' traits and behaviours are used to develop stereotypes. Analyses and interprets persuasive texts to identify bias and to infer a specific purpose and audience. Interprets vocabulary, including technical words, specific to an informative text or topic.
Band 7	Applies knowledge and understanding of different text types and features to enhance meaning and infer themes and purpose. Identifies details that connect implied ideas across and within texts to process information and form conclusions. Interprets character motivation in narrative texts, the writer's values in persuasive texts and the main ideas in informative texts.
Band 6	Makes meaning from a range of text types of increasing difficulty and understands different text structures. Recognises the purpose of general text features such as titles and subheadings. Makes inferences by connecting ideas across different parts of texts. Draws conclusions about the feelings and motivations of characters, and sequences events and information.
Band 5	Applies knowledge, makes inferences and processes information to infer the main idea in texts. Draws conclusions about a character in narrative texts. Connects and sequences ideas in informative texts and identifies opinions in persuasive texts.
Band 4	Makes inferences from clearly stated information in short informative texts and stories. Identifies the meaning of some unfamiliar words from their context. Finds specific information in longer stories and informative texts including those with tables and diagrams.
Band 3	Makes meaning from simple texts with familiar content and themes and finds directly stated information. Makes some connections between ideas that are not clearly stated and identifies simple cause and effect. Makes some inferences and draws conclusions, such as identifying the main idea of a text.
Band 2	Makes some meaning from short texts, such as simple reports and stories, that have some visual support. Makes connections between pieces of clearly stated information.

Proficiency band	Reading skills and knowledge
Band 1	Makes some meaning from simple texts with familiar content. Texts have short sentences, common words and pictures to support the reader. Finds clearly stated information.

Table 86: Described scale for writing

Proficiency band	Writing skills and knowledge
Band 10	Writes a cohesive, engaging text that explores universal issues and influences the reader. Creates a complete, well-structured and well-sequenced text that effectively presents the writer's point of view. Effectively controls a variety of correct sentence structures. Uses punctuation correctly, including complex punctuation. Spells all words correctly, including many difficult and challenging words.
Band 9	Incorporates elaborated ideas that reflect a worldwide view of the topic. Makes consistently precise word choices that engage or persuade the reader and enhance the writer's point of view. Punctuates sentence beginnings and endings correctly and uses other complex punctuation correctly most of the time. Shows control and variety in paragraph construction to pace and direct the reader's attention.
Band 8	Writes a cohesive text that begins to engage or persuade the reader. Makes deliberate and appropriate word choices to create a rational or emotional response. Attempts to reveal attitudes and values and to develop a relationship with the reader. Constructs most complex sentences correctly. Spells most words, including many difficult words, correctly.
Band 7	Develops ideas through language choices and effective textual features. Joins and orders ideas using connecting words and maintains clear meaning throughout the text. Correctly spells most common words and some difficult words, including words with less common spelling patterns and silent letters.
Band 6	Organises a text using paragraphs with related ideas. Uses some effective text features and accurate words or groups of words when developing ideas. Punctuates nearly all sentences correctly with capitals, full stops, exclamation marks and question marks. Correctly uses more complex punctuation markers some of the time.
Band 5	Structures a text with a beginning, complication and resolution, or with an introduction, body and conclusion. Includes enough supporting detail for the text to be easily understood by the reader, although the conclusion or resolution may be weak or simple. Correctly structures most simple and compound sentences and some complex sentences.
Band 4	Writes a text in which characters or setting are briefly described, or in which ideas on topics are briefly elaborated. Correctly punctuates some sentences with both capital letters and full stops. May demonstrate correct use of capitals for names and some other punctuation. Correctly spells most common words.

Proficiency band	Writing skills and knowledge
Band 3	Attempts to write a text containing a few related events or ideas on topics, although these are usually not elaborated. Correctly orders the words in most simple sentences. May experiment with using compound and complex sentences but with little success. Orders and joins ideas using a few connecting words but the links are not always clear or correct.
Band 2	Shows audience awareness by using common text elements, for example, begins writing with <i>Once upon a time</i> ; or <i>I think ... because ...</i> . Uses some capital letters and full stops correctly. Correctly spells most simple words used in the writing. Some other one- and two-syllable words may also be correct.
Band 1	Writes a small amount of simple content that can be read. May name characters or a setting; or write a few content words on a topic. May write some simple sentences with correct word order but full stops and capital letters are usually missing or incorrect. Correctly spells a few simple words used in the writing.

Table 87: Described scale for conventions of language

Proficiency band	Conventions of language skills and knowledge
Band 10	Identifies errors and correctly spells difficult words and challenging words (<i>interrupt, camouflaged, instantaneous</i>). Demonstrates knowledge of the correct use of a wide range of grammar and punctuation conventions in complex texts.
Band 9	Identifies errors and correctly spells words with difficult spelling patterns (<i>rehearsals, deliberately, consistently</i>). Demonstrates knowledge of grammar and punctuation conventions in more complex texts, such as the correct use of possessive pronouns (<i>its</i>) and rhetorical questions.
Band 8	Identifies errors and correctly spells most words with difficult spelling patterns (<i>angrily, substantial, performance</i>). Demonstrates knowledge of grammar and punctuation conventions in more complex texts, such as the correct use of adverbs, pairs of conjunctions (<i>neither, nor</i>), cause and effect structures, quotation marks for effect and for speech and apostrophes for plural possession (<i>parents'</i>).
Band 7	Identifies errors and correctly spells words with common spelling patterns and some words with difficult spelling patterns (<i>applauded, received, achievement</i>). Demonstrates knowledge of grammar and punctuation conventions in more complex texts, such as appropriate and consistent sentence structure and the correct use of italics, apostrophes and commas to separate phrases.
Band 6	Identifies errors and correctly spells most words with common spelling patterns (<i>gloves, collect, hungry, comfortable</i>). Demonstrates knowledge of grammar and punctuation conventions in longer sentences and speech, such as the correct use of commas to separate phrases and apostrophes for contractions (<i>we'll</i>).
Band 5	Identifies errors and correctly spells one- and two-syllable words with common spelling patterns (<i>spill, locked, pleasing, benches</i>). Recognises grammar and punctuation conventions in standard sentences and speech, such as the correct use of adjectives, compound verbs (<i>could have</i>), capital letters for compound proper nouns and commas in lists.

Proficiency band	Conventions of language skills and knowledge
Band 4	Identifies errors and correctly spells most one- and two-syllable words with common spelling patterns (<i>clear, mail, brick, won</i>). Recognises grammar and punctuation conventions in short sentences and speech, such as the correct use of groups of adjectives, referring pronouns (<i>those</i>) and capital letters for simple proper nouns.
Band 3	Identifies errors and correctly spells one-syllable words with simple spelling patterns (<i>out, feet, rain, hose, would</i>). Recognises grammar and punctuation conventions in short sentences, such as the correct use of linking and coordinating words (<i>that, but</i>), describing words, capital letters to begin a sentence, full stops and question marks.
Band 2	Identifies errors and correctly spells some words with simple spelling patterns. Recognises grammar and punctuation conventions in short sentences, such as the correct use of pronouns (<i>herself</i>).
Band 1	Identifies errors and correctly spells a few words with simple spelling patterns. Recognises a small range of grammar and punctuation conventions in short sentences, such as the correct use of simple conjunctions (<i>because</i>) and common verbs (<i>will go</i>).

Out of the 10 bands, only 6 bands were reported for each year level. Bands 1 to 6 were reported for Year 3, bands 3 to 8 for Year 5, bands 4 to 9 for Year 7 and bands 5 to 10 for Year 9. Students in the lowest band for each year level were regarded as achieving below the National Minimum Standard (NMS), while students in the second lowest band for each year level were regarded as being at the NMS.

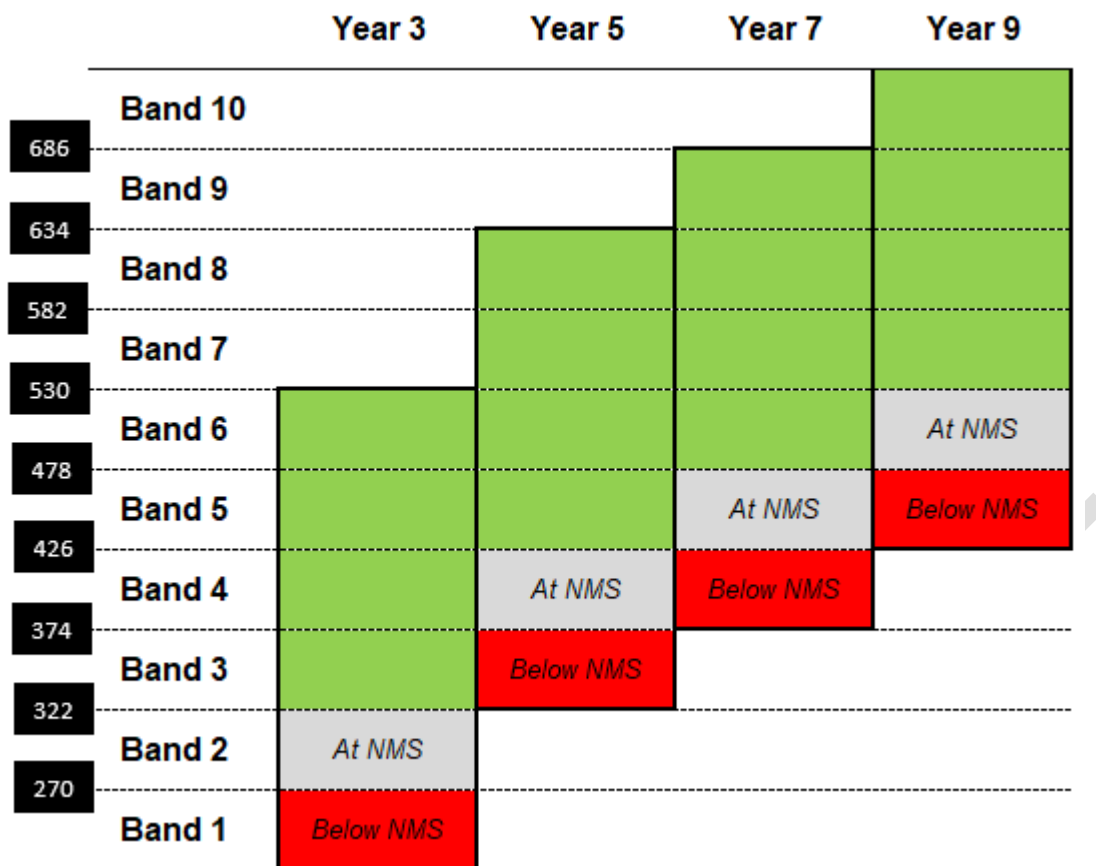


Figure 53: Schematic picture of proficiency bands by year levels

Illustrations

One Year 5 student received a NAPLAN score of 480 for numeracy. A score of 480 is near the lower bound of band 6. This student is expected to respond correctly to 50% of the items that have an RP62 difficulty between 478 and 530 and therefore is regarded as mastering the skills that are described for band 6 (see Table 84). This student is ready to be introduced to some of the skills and concepts described for band 7.

Another Year 5 student received a NAPLAN score of 530 for numeracy. This student achieves at the very top of band 6 and is expected to respond correctly to about 70% of the items in this band. The student, therefore, has mastered most skills within band 6 (see Table 84) and is ready to learn the skills and concepts described for band 7.

Chapter 9: Reporting of national results

NAPLAN produces several reports for a variety of audiences each year. The student and school summary report (SSSR)¹ is a preliminary report with student and school level results for school staff. The individual student report (ISR)² is a report for parents/carers about their child's NAPLAN achievement. The national report includes final national statistics to inform policymakers and researchers. Additional reporting, with results for individual schools, is also provided on the website My School³, which is accessible to the general public. This chapter describes analysis for the national report.

Calculation of statistics using plausible values

All statistics included in the national report were based on plausible values. Plausible values are student-level achievement score that result in unbiased population statistics. For each student, 5 plausible values were drawn. When performing secondary analyses, each analysis needed to be run 5 times, once for each plausible value. The final statistic was the average of the 5 results. Plausible values should never be averaged at the student level. The formal notation for this is:

$$\theta = \frac{1}{5} \sum_{i=1}^5 \theta_i \quad (18)$$

Where θ_i is a population parameter estimate from the i^{th} plausible value, with θ being any type of population statistic (mean, standard deviation, percentage).

Computation of standard errors

All statistics are associated with a level of uncertainty. This uncertainty is expressed as a standard error. Appropriate standard errors are crucial for ensuring that conclusions drawn based on observed score or performance differences are accurate. More precisely, appropriate standard errors need to be used as part of statistically testing the likelihood that certain observed performance differences could have arisen by chance alone before concluding that a statistically meaningful difference exists.

Three types of errors were estimated and different combinations of the standard errors were used for different types of comparisons. The first type of error was the uncertainty caused by the selection of students participating in the study: the sampling error. The second type of error was uncertainty caused by the measurement tool (the tests): the measurement error. The third type was uncertainty caused by the equating design: the equating error. Estimation of the equating error was explained in Chapter 7. The other 2 types of errors are explained in this chapter.

Sampling error

The inclusion of sampling error might be considered surprising in that all students in the target year levels were included in the assessment. However, the aim of NAPLAN is to make inferences about trends in the educational systems over time and not about the specific student cohorts in 2022. In addition, even in census assessments, there is a certain amount of non-response that must be taken into account. Sampling error was considered at both the student and the school level. At the student level, there is a random element from one assessment year to another with respect to different age cohorts at each year level. At the school level, it needs to be considered that schools may be closed from one year to another or new schools may be opened.

The Taylor Series Linearization method (Wolter 1985, Levy and Lemeshow 2013) was used to construct an approximation to the functional form of the estimated population characteristic that is a

¹ www.nap.edu.au/docs/default-source/default-document-library/how-to-interpret-the-sssr.pdf?sfvrsn=10

² www.nap.edu.au/results-and-reports/student-reports

³ www.myschool.edu.au/

linear function of the original observations and hence is amenable to construction of a variance estimator.

The process of *linearization* or *Taylor series variance estimation* involves several steps. To look at a simple case, consider a population characteristic θ and assume that an estimator $\hat{\theta} = f(x, y)$ exists such that the variables x and y are linear functions of the sample observations, but that $f(x, y)$ is *not* a linear function of the sample observations. The next step is to use a first-order Taylor series to approximate $f(x, y)$. This results in an approximation that is linear in the variables x and y , and hence, linear in the sample observations. The final step is to take this linear approximation, identify the sample design, and apply the design-based formula to estimate the variance (Levy and Lemeshow 2013).

Taylor series variance estimation can be done using commercially available statistical software. For NAPLAN 2022, the Complex Samples module implemented in the SPSS software package and the procedure *Proc Surveymeans* in the SAS software package were used in parallel processing for checking. Example of these procedures are included in Figure 54. The sampling error is equal to the square root of the sampling variance.

SPSS	SAS
<pre> Compute WGT=1. Exe. * Analysis Preparation Wizard. CSPLAN ANALYSIS /PLAN FILE='directory\report\calibration.csaplan' /PLANVARS ANALYSISWEIGHT=WGT /SRSESTIMATOR TYPE=WOR /PRINT PLAN /DESIGN CLUSTER=school_id /ESTIMATOR TYPE=WR. </pre>	<pre> proc surveymeans data=temp; cluster school_id; by grade <subgroups>; var PV1-PV5; ods output statistics=PVout; run; </pre>

Figure 54: Examples in SPSS and SAS for estimating sampling variance

Measurement error

Plausible values methodology enables the computation of the uncertainty in the estimate of θ due to the lack of precision in the test. This is not possible if point estimates for student achievement, such as WLEs, are used in secondary analysis for reporting. If a perfect test could be developed, then the measurement error would be equal to zero and the 5 statistics from the plausible values would be identical. Since no test is perfectly reliable, the 5 sets of statistics will not be identical. The measurement variance is estimated as:

$$B_M = \frac{1}{M-1} \sum_{i=1}^M (\theta_i - \theta)^2 \quad (19)$$

It corresponds to the variance of the 5 plausible value statistics of interest. The measurement error is equal to the square root of the measurement variance.

The measurement variance is combined with the sampling variance to express the uncertainty in population statistics:

$$V = U + \left(1 + \frac{1}{M}\right) B_M \quad (20)$$

$$SE = \sqrt{V} \quad (21)$$

with U being the sampling variance.

Macros were written in both SPSS and SAS to combine the estimates of sampling error with the estimates of measurement error to obtain final standard errors for the performance statistics reported for the census data. The standard errors were used to determine statistical significance in mean differences in NAPLAN 2022 performance in the reports.

Testing for differences

Two types of differences could be computed and tested for significance. The first type of comparison was between subgroups within the NAPLAN 2022 data; for example, between male and female students or between jurisdictions. The second type of comparison was between 2022 results and results from earlier assessment years. Differences of the first type were tested for significance using the standard errors estimated from the sampling variance and the measurement variance. For testing the second type of differences, the equating errors needed to be taken into account as well.

To illustrate how statistical testing of the 2 types of performance differences was carried out in the NAPLAN context, 2 hypothetical examples – focusing on differences in mean scores – are provided.

The first example shows the comparison of 2 hypothetical mean scale scores – θ_A and θ_B – for 2 subgroups (for example, gender) A and B, within the same calendar year. As these hypothetical means can be regarded as independent (that is, zero covariance), a standard error for the difference between them can be computed using the following formula:

$$SE_{DIFF} = \sqrt{SE_A^2 + SE_B^2} \quad (22)$$

where SE_{DIFF} is the standard error of the difference and SE_A and SE_B are the standard errors of the respective means θ_A and θ_B for groups A and B. The test statistic t is calculated by dividing the difference between the 2 means by the standard error of the difference. A significance level of 0.05 was used for all statistical tests, with corresponding critical values of ± 1.96 . This illustrative example can be taken further by setting θ_A and θ_B to 500 and 515, respectively, and setting SE_A and SE_B to 3 and 4, respectively. Then, θ_B minus θ_A equals 15 and the standard error for this difference is equal to the square root of the sum of 9 and 16, thus SE_{DIFF} is equal to 5. The t statistic is therefore equal to 15 divided by 5, which equals 3, exceeding the critical value of 1.96, and thus representing a statistically significant difference at the 0.05 significance level.

The second example involves statistical testing of performance differences between calendar years. This requires inclusion of the equating error in the calculation of SE_{DIFF} . Drawing on the previous example, if we now consider the difference between group A's mean score in 2022 and 2021, we need to add the equating error between these 2 years, $SE_{2022to2021}$, to the calculation in the following way:

$$SE_{DIFF} = \sqrt{SE_{A21}^2 + SE_{A22}^2 + SE_{2022to2021}^2} \quad (6)$$

The same procedure as shown in the previous example can then be applied to evaluate the statistical significance of the difference. Actual equating errors for comparisons of mean scale scores involving 2022 NAPLAN with 2021, and with the base year for each domain and year level, are included in Chapter 7. No NAPLAN tests were administered in 2020 due to the pandemic, hence 2020 was skipped from reporting of the NAPLAN long-term trend and in NAPLAN growth results.

Only when differences between subgroups are compared between calendar years – for example, the gap between Indigenous and non-Indigenous students over time – does the equating error not need to be taken into account. This is because both group statistics are equally affected by uncertainty due to equating, which is therefore cancelled out. This type of comparison, however, is not included in the NAPLAN 2022 National Report.

Effect sizes

All significance testing in NAPLAN is accompanied by an effect size measure, which indicates the magnitude of any difference as opposed to indicating the likelihood that the difference could have arisen through chance alone. The incorporation of a measure of effect size can usefully aid the interpretation of differences, because under conditions of relatively small standard errors (as can often arise with large sample sizes), statistical testing alone can flag small differences as being significant when such differences could be inconsequential from a practical point of view.

Effect size for comparing means

In previous years, effect sizes for pairwise comparisons between assessment years (current year versus previous year and current year versus base year) were calculated by *Hedge's g* (ACARA 2022). The effect size was then compared to the criterion of 0.2 for a small effect and 0.5 for a large effect. For 2022, a different method was used for calculating the effect size for each pairwise comparison and it was based on estimated growth as described below.

First, a logarithmic model was fitted to regress mean achievement on year level by domain. Mean achievement is the year level average over *all previous assessment years* (2008 to 2021 for non-writing domains; 2011 to 2021 for writing domains). The logarithmic regression function is

$$\bar{Y} = a * LN(X) + b \quad (23)$$

Where X is the year level (i.e. 3, 5, 7 or 9) and \bar{Y} is the estimated mean score. The fit of the logarithmic function was over 0.99 for all domains. Figure 31 shows the logarithmic regression function for numeracy as an example.

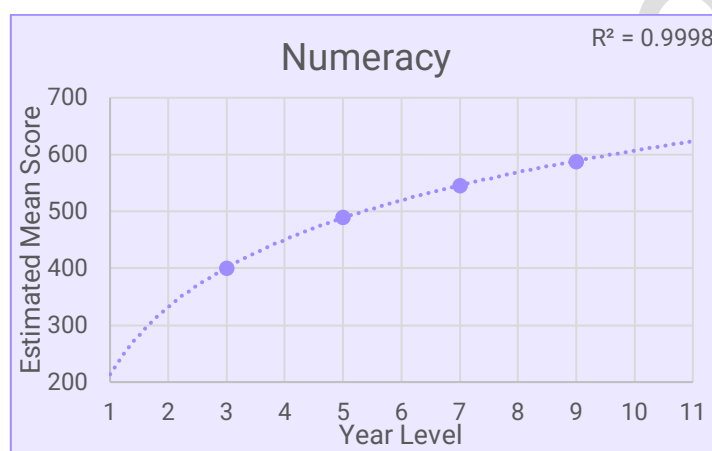


Figure 55: Logarithmic regression function for numeracy

The intercepts and slopes of the logarithmic function for each domain are included in Table 88.

Table 88: Intercept and slope of growth regression by domain

	Numeracy	Reading	Spelling	Grammar and Punctuation	Writing
Intercept (b)	213.536	266.519	243.870	283.983	279.442
Slope (a)	170.861	142.522	154.027	131.820	122.554

Second, the growth-based effect size (d) in months of learning for pairwise comparisons of the current year with the previous year was calculated as follows:

$$d = 12 * (\bar{X}_{22} - \bar{X}_{21}) = 12 * \left(e^{\left(\frac{Y_{22}-b}{a}\right)} - e^{\left(\frac{Y_{21}-b}{a}\right)} \right) \quad (24)$$

For pairwise comparisons of the current year with the base year, the following equation was used to calculate the growth-based effect size (d) in months of learning:

$$d = 12 * (\bar{X}_{22} - \bar{X}_{base}) = 12 * \left(e^{\left(\frac{Y_{22}-b}{a}\right)} - e^{\left(\frac{Y_{base}-b}{a}\right)} \right) \quad (25)$$

Third, the effect size (d) was compared to the criterion of 2 months of growth for a small effect and 3 months of growth for a large effect.

Effect size for differences in percentages

The effect size for differences in percentages has not changed from previous years and is given by Cox's d , the formula for which is:

$$OR = \frac{p_E q_C}{q_E p_C} \quad (26)$$

$$d_{Cox} = \frac{L(OR)}{1.65} \quad (27)$$

Where p_E and p_C are the percentages of comparison, and $q_E=100-p_E$, $q_C=100-p_C$.

Three effect sizes were reported for differences in percentages as follows:

- “substantially above/below” refers to an effect size of greater than 0.5 / less than -0.5
- “above/below” refers to an effect size between 0.2 and 0.5 / between -0.2 and -0.5
- “close to” refers to an effect size of less than 0.2 but greater than -0.2.

Effect size for long-term trends

As mentioned in Chapter 7, for 2022, long-term trends and their significance were determined by domain and year level. The following steps were applied by domain and year level for each subgroup:

1. Fit regression model with mean scores as Y and calendar years indicators as X (long-term trend).
2. Calculate predicted Y_{22} based on the regression coefficients from step 12.
3. Calculate predicted Y_{21} by subtracting the slope of the long-term trend, e.g. $Y_{21} = Y_{22} - \text{Slope}$.
4. Use formula 18 to estimate effect size (d_L) which is the average annual trend.
5. The criterion for effect size (d_L) is 0.25, which is a quarter of a month, or about one week.

References

- Adams RJ, Wu ML, Cloney D and Wilson MR (2022) *ACER ConQuest: generalised item response modelling software* [computer software], version 5. Camberwell, Victoria: Australian Council for Educational Research.
- Adams JR and Lazendic G (2013) *Observations on the Feasibility of a Multistage Test Design for NAPLAN*, unpublished technical report.ACARA (Australian Assessment, Curriculum and Reporting Authority) (2022) *The Australian National Assessment Program Literacy and Numeracy (NAPLAN): 2021 Technical Report*, ACARA: Sydney.
- Breithaupt K and Hare DR (2007) Automated Simultaneous Assembly of Multistage Testlets for a High-Stakes Licensing Examination, *Educational and Psychological Measurement*, 67(1), pp 5–20.
- Camilli G and Shepard LA (1994) *Methods for identifying biased test items* (Vol. 4), Thousand Oaks: Sage.
- Eggen TJ and Verhelst ND (2011) Item calibration in incomplete testing designs, *Psicológica*, 32(1), pp 107–132.
- Hendrickson A (2007) An NCME Instructional Module on Multistage Testing, *Educational Measurement: Issues and Practice*, 26, 2.
- Humphry SM and McGrane JA (2015) Equating a large-scale writing assessment using pairwise comparisons of performances, *The Australian Educational Researcher*, 42, pp 443–60.
- Levy PS and Lemeshow S (2013) *Sampling of populations: methods and applications* (4th edition), New York: John Wiley & Sons.
- Lord FM and Novick MR (1968) *Statistical Theories of Mental Test Scores*, Addison-Wesley: Menlo Park.
- Luecht RM, Brumfield T and Breithaupt K (2006) A testlet assembly design for adaptive multistage tests, *Applied Measurement in Education*, 19(3), pp 189–202.
- Masters GN (1982) A rasch model for partial credit scoring, *Psychometrika* 47, pp 149–174.
- Mislevy RJ and Sheehan KM (1987) Marginal estimation procedures, in Beaton AE, Editor (1987) *The NAEP 1983–84 technical report, National Assessment of Educational Progress*, Educational Testing Service, Princeton, pp 293–360.
- Rasch G (1960) *Probabilistic models for some intelligence and attainment tests*, Copenhagen: Danmark Paedagogiske Institut.
- Rasch G (1980) *Probabilistic models for some intelligence and attainment tests* (expanded ed.), Chicago: University of Chicago Press.
- Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.
- Rubin DB (1991) EM and beyond, *Psychometrika*, 39, 111–21.
- Warm TA (1989) Weighted Likelihood Estimation of Ability in Item Response Theory, *Psychometrika*, 54 (3), pp 427–50.
- Wolter KM (1985) *Introduction to Variance Estimation*, New York: Springer-Verlag.

Appendices

Appendix A

<https://nap.edu.au/docs/default-source/default-document-library/appendix-a-percentages-and-ability-distribution-by-pathway-2022.pdf>

Appendix B

<https://nap.edu.au/docs/default-source/default-document-library/appendix-b-item-analysis-details-2022.pdf>

Appendix C

<https://nap.edu.au/docs/default-source/default-document-library/appendix-c-item-summary-tables-2022.pdf>

Appendix D

<https://nap.edu.au/docs/default-source/default-document-library/appendix-d-item-characteristic-curves-2022.pdf>

Appendix E

[https://nap.edu.au/docs/default-source/default-document-library/appendix-e-expected-score-curves-\(writing\)-2022.pdf](https://nap.edu.au/docs/default-source/default-document-library/appendix-e-expected-score-curves-(writing)-2022.pdf)

Appendix F

<https://nap.edu.au/docs/default-source/default-document-library/appendix-f-item-person-maps-2022.pdf>

Appendix G

<https://nap.edu.au/docs/default-source/default-document-library/appendix-g-gender-dif-analysis-2022.pdf>

Appendix H

<https://nap.edu.au/docs/default-source/default-document-library/appendix-h-lbote-dif-analysis-2022.pdf>

Appendix I

<https://nap.edu.au/docs/default-source/default-document-library/appendix-i-indigenous-status-dif-analysis-2022.pdf>

Appendix J

<https://nap.edu.au/docs/default-source/default-document-library/appendix-j-dif-summary-tables-2022.pdf>

Appendix K

<https://nap.edu.au/docs/default-source/default-document-library/appendix-k-jurisdictional-dif-2022.pdf>

Appendix L

<https://nap.edu.au/docs/default-source/default-document-library/appendix-l-device-dif-2022.pdf>

Appendix M

<https://nap.edu.au/docs/default-source/default-document-library/appendix-m-vertical-link-item-comparisons-2022.pdf>

Appendix N

<https://nap.edu.au/docs/default-source/default-document-library/appendix-n-horizontal-link-item-comparison-2022.pdf>

Appendix O

<https://nap.edu.au/docs/default-source/default-document-library/appendix-o-construct-variance-covariance-matrix-2022.pdf>

Appendix P

<https://nap.edu.au/docs/default-source/default-document-library/appendix-p-naplan-2022-exception-report-2022.pdf>