NAPLAN 2023

Technical Report

May 2024

acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

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

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) 2024, 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

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The first National Assessment Program – Literacy and Numeracy (NAPLAN) tests took place in 2008. 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 gives federal and jurisdictional governments, schools and parents/carers information about whether young Australians are reaching important educational goals.

NAPLAN tests are the only Australian assessments that provide nationally comparable data on students' performance 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.

From 2008 to 2017, NAPLAN delivered only paper-based tests. From 2018, NAPLAN delivered both paper-based tests and online multistage tailored tests. The online tailored tests in reading, spelling, grammar and punctuation, and numeracy were delivered to students in participating schools. Online writing tests were delivered to students in Years 5, 7 and 9. Year 3 writing tests continue to be delivered on paper. Alternative-format tests (paper, large-print, Braille, electronic PDF) are made available for those students who require them. In 2023, almost all students completed online tests.

NAPLAN results are reported using 5 national achievement scales, one for each of the assessed aspects of literacy – reading, writing, spelling, and 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. In 2023, the NAPLAN achievement scales were reset. This meant that direct comparison between results in 2023 and earlier years was not possible.

One reason to reset the scales was that the timing of the NAPLAN tests changed. They were administered in March rather than May, so that results could be returned to schools earlier in the school year. The effect of this change on student achievement could not be predicted with certainty.

In addition, the adaptive tests allow the possibility of more precise measurement of student achievement, particularly for low- and high-performing students, who are presented with test items that better match their ability. However, this more precise measurement could not be fully realised while each year's results were equated to a historical scale, which had originally been based on fixed paper tests. The new scale established in 2023 better shows the distribution of achievement both within each year level and across year levels.

NAPLAN was also reported differently in 2023 with the introduction of proficiency levels. These replaced the 10 numerical achievement bands and national minimum standards, which were used in previous NAPLAN cycles.

Four outcome reports were produced for NAPLAN 2023:

- The Student and School Summary report (SSSR) is an interactive report produced for online schools, showing the achievement of their students in all NAPLAN tests.
- The Individual Student Report (ISR) provides information to parents/carers about their child's performance on the NAPLAN tests.
- The NAPLAN 2023 National Results show national performance data, as well as the performance of states, territories and subgroups. These results are available on the ACARA website for 2023 and all previous cycles.
- MySchool reports show NAPLAN results for each school, alongside a variety of other school information.

Resetting the measurement scales meant that direct comparisons from 2023 to previous assessment cycles could not be made. As a result, some features of the National Results and MySchool were not available in 2023, and will not be available until sufficient longitudinal data has accumulated.

The Australian Council for Educational Research (ACER) was appointed by ACARA to undertake the central analysis of test data for NAPLAN 2023. The key task in 2023 was to construct new measurement scales, through test calibration and a vertical equating exercise.

The aim of this technical report is to describe in detail the methodology used for NAPLAN 2023.

- Chapter 2 describes how items were developed and trialled in 2023 to establish a pool of "test-ready" items, reading texts and writing prompts, for use in future NAPLAN tests.
- Chapter 3 describes the test design and construction process, and the setting of branching rules.
- Chapter 4 describes the data preparation process.
- Chapter 5 describes the psychometric scaling methodology and outcomes.
- Chapter 6 describes the test equating processes used to establish the new NAPLAN measurement scales.
- Chapter 7 describes the processes through which the new proficiency levels were established.
- Chapter 8 describes the methodology used for reporting of NAPLAN 2023 performance.

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

Chapter 2: Item development and item trial

This chapter describes the processes through which NAPLAN items, prompts and texts were developed and trialled in 2023, to establish a pool of test material for use in future NAPLAN cycles. The first part of this chapter describes the processes by which items were developed and trial tests constructed, the second part describes the item trial administration, and the third part explains the psychometric analysis of trial data.

Item development

Numeracy, reading and conventions of language

Items and texts were developed to build and replenish pools of items available for use in item trial tests (so-called "trial-ready" items) in numeracy, reading, spelling, and grammar and punctuation. Trial tests were then constructed using items drawn from this pool, and administered to students in June 2023. The development and trialling of these items was guided by the need to develop a bank of items that meet specifications for difficulty, curriculum content and item type (so called "test-ready" items) that are available for use in the construction of future NAPLAN tests.

Items in each batch were reviewed by ACARA, the National Testing Working Group (NTWG) and independent domain experts. Further rounds of review were conducted as necessary by item writers, subject and language specialists, Indigenous reviewers, item development managers and editors.

Cultural reviews were also conducted for some reading texts. 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 by ACARA for intellectual property, Indigenous cultural and intellectual property, and moral rights.

All review feedback was synthesised by ACARA and the items or texts requiring modification were revised until acceptable.

Writing

In 2023, prompts for writing tests were developed and trialled according to the following process.

Education experts from all jurisdictions developed a large pool of writing tasks to engage students in Years 3 and 5, and Years 7 and 9. Each jurisdiction convened panels of experts with significant experience in writing assessment, and educators representing key special needs groups.

Expert panels undertook 4 stages of review of all writing tasks in the pool to ensure that they were accessible for students from a range of backgrounds. Panels considered what students might write about and whether the task would be fair for students. In early stages of the review, the panels made overall judgements of which writing tasks might be prioritised for administration in NAPLAN, providing feedback where necessary. In later stages of the review, they distilled the suitable tasks and suggested changes to wording and images. A shortlist of 10 topics was chosen and refined, for administration at trial.

All domains

Item developers in each domain complied with the following documents:

- NAPLAN Assessment Framework (ACARA 2017 link)
- NAPLAN Item development guidelines (ACARA internal document)
- Guidelines for the development of accessible NAPLAN online items (link).

Audio was recorded for all numeracy, audio dictation (spelling) items and writing prompts prior to trialling. This entailed marking up the text that needed to be recorded, followed by recording, editing, attaching audio, and quality assurance of all recordings.

Item trial

Each year, an assessment event is conducted to trial the performance of items. The item trial process produces critical item performance data used to identify items appropriate for use in future NAPLAN tests. These are stored in a "test-ready" item bank.

Item trial design: numeracy, reading and conventions of language

To support the placement of items on the NAPLAN scale, the trial tests were administered to a representative, stratified sample of schools and students. The trial tests included items from the previous year's NAPLAN tests so that the trial results could be equated to the NAPLAN scale using a common-item equating methodology.

As items presented at the end of a test could perform differently from those presented at the beginning (due to accumulated cognitive load or time pressure), the trial tests were designed so that items were presented at differing positions within the tests.

Items were incorporated into testlets, which were then rotationally allocated to students within each class, using functionality inbuilt within the national assessment platform. This ensured that items were administered to a set of students that was representative of the trial sample as a whole.

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

Table 1 to Table 4 below show the composition of the trial pools by domain, year level and item format: either multiple-choice(s) (MC), or other, which includes constructed response (CR) and technology-enhanced items (TEI). The conventions of language test is separated into its 2 component sections: grammar and punctuation, and spelling. All spelling items are constructed response, so are classified instead into audio dictation (AD) or proofreading (PR) formats.

Table 1: Composition of the 2023 numeracy item trial

	MC	Other	Total
Year 3	132	84	216
Year 5	183	119	302
Year 7	239	139	378
Year 9	245	185	430
Total	799	527	1326

Table 2: Composition of the 2023 reading item trial

	MC	Other	Total
Year 3	288	48	336
Year 5	289	47	336
Year 7	358	42	400
Year 9	347	53	400
Total	1282	190	1472

Table 3: Composition of the 2023 grammar and punctuation item trial

	MC	Other	Total
Year 3	109	143	252
Year 5	89	163	252
Year 7	84	168	252
Year 9	79	173	252
Total	361	647	1008

Table 4: Composition of the 2023 spelling item trial

	AD	PR	Total
Year 3	90	162	252
Year 5	90	162	252
Year 7	90	162	252
Year 9	90	162	252
Total	360	648	1008

Item trial design: writing

To support the placement of items on the NAPLAN scale, the trial tests were administered to a representative, stratified sample of schools and students. The trial tests included items from the previous year's NAPLAN tests so that the trial results could be equated to the NAPLAN scale using a common-item equating methodology.

Sample

Two samples were drawn for the item trial: primary students in Years 3 and 5, and secondary students in Years 7 and 9. For both primary and secondary samples, sample sizes of 250 schools each were chosen with probability proportional to school size. The sample size was based on the number of responses required for analysis of the items.

The following schools were excluded from selection for the item trial:

- remote and very remote schools
- schools with fewer than 20 students
- non-mainstream schools (such as schools for students with intellectual disabilities or hospital schools, Steiner, Montessori and Waldorf schools, distance education schools)
- schools without NAPLAN performance data
- schools that participated in the NAPLAN 2022 item trial.

Schools participating in major studies (NAP–Science Literacy Main Study 2023 and TIMSS Main Study 2023) within the same academic year were also excluded.

The sampling frame was based on schools' data supplied by ACARA and supplemented with additional information provided by the sampling contractor. It was stratified by state, sector, school size, NAPLAN performance and a school location-based measure of socio-economic background, the Australian Bureau of Statistics (ABS) Index of Education and Occupation, one of the ABS Socio-Economic Indexes for Areas

(SEIFA). For each sampled school, up to 2 schools with similar characteristics were identified as possible substitutes in case the sampled school did not participate. To improve the efficiency of the field operation, schools selected in outer regional Victoria, New South Wales and Queensland were adjusted to create hubs within a radius of 100km from a central point.

After sample selection, each school was systematically assigned one of the domain pair combinations supplied by ACARA (NR, RW, NC, NW, RC and WW) following a repeated sequence so that domain combinations were covered uniformly throughout the sampled list of schools. The allocation ensured that there were sufficient schools and students allocated to each domain to achieve the target responses from each domain for the Item Trial, while preserving the stratification structure across domains as far as possible. The school size variable was used to distinguish smaller and larger schools; some of the latter were requested to provide an additional class. At the primary level, a second domain pair was allocated to 50 larger schools. For secondary, 55 larger schools were allocated a second domain pair.

Table 5 shows the number of classes selected for each combination of domain pairs across the primary and secondary samples.

Table 5: Number of classes selected for each domain pair in primary and secondary year levels.

	Primary (Y3/Y5)			Secondary (Y7/Y9)		
Domain Pair	Count of 1st domain pair	Count of 2nd domain pair	Total domain pair class count	Count of 1st domain pair	Count of 2nd domain pair	Total domain pair class count
NR	42	7	49	42	8	50
NC	28	7	35	27	6	33
NW	41	7	48	69	5	74
RC	28	8	36	27	6	33
RW	42	7	49	29	18	47
CW	0	0	0	0	0	0
WW	69	14	83	56	12	68
Total	250	50	300	250	55	305

Survey

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

- gender
- device used
- whether students were used to typing stories or essays at school.

The responses to the gender item were used in the analysis of student performance to determine whether there was evidence of differential item functioning (DIF) by gender. Responses to the other items are used to monitor students' online experience over time.

Trial participation

A total of 446 schools across all states and territories participated. Note that while 250 primary schools and 250 secondary schools were sampled, the total number of schools reflects the fact that some schools provided both primary and secondary classes.

The number of students who completed the tests in each non-writing domain is presented in Table 6.

Table 6: Trial participation: reading, conventions of language and numeracy

Domain	Year 3	Year 5	Year 7	Year 9	Total
Reading	2,706	2,807	2,656	2,396	10,565
Conventions of language	2,989	3,082	3,002	2,738	11,811
Numeracy	2,940	3,023	3,288	3,059	12,310

The number of students who completed each writing task is presented in Table 7.

Table 7: Trial participation: writing

Prompt	Year 3	Year 5	Year 7	Year 9	Total
Task 1	526	621	490	447	2,084
Task 2	536	631	490	455	2,112
Task 3	529	616	498	469	2,112
Task 4	506	599	476	442	2,023
Task 5	493	603	488	466	2,050
Task 6	496	597	498	470	2,061
Task 7	501	591	512	471	2,075
Task 8	483	578	481	455	1,997
Task 9	392	416	419	412	1,639
Task 10	504	451	388	334	1,677
Task 1 paper	384	0	0	0	384
Task 5 paper	377	0	0	0	377
Total	5,727	5,703	4,740	4,421	20,591

Test administration

The National Assessment 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 in June 2023. The trial was supported by trained invigilators in all schools.

Marking of writing responses

A team of experienced NAPLAN markers was engaged by an external marking contractor to mark the writing responses. Writing responses were extracted from the platform, and sent along with the paper responses to the marking contractor. ACARA's writing test manager supported the training of the markers and remained in communication to oversee the marking process. Once the marking of each prompt was completed, a debriefing session was held with the test developers and amendments were made to the training materials as necessary. Qualitative feedback on the marking of each prompt was gathered to be used alongside the quantitative data when selecting prompts for the main study.

Psychometric analysis of item trial data: numeracy, reading and conventions of language

The following steps were taken to analyse the item trial data:

Data validation and recoding

In order to ensure the data was of high quality and could be used in the analysis, each data set was validated separately and anomalies were removed. Raw data was also recoded to suit the purposes of analysis: embedded missing responses were coded “9” and items not administered to a student were coded “8”.

Year level analysis

Data for each year level was analysed separately for each domain. The Rasch measurement model (Rasch 1960), using ACER Conquest (Adams, Wu, Cloney and Wilson 2020), was used for item calibration. The process allows for 2 rounds of item calibration, if it was necessary to correct item scoring or to omit misfitting items from analysis.

The calibrated items were then placed on the historical NAPLAN scale using a common-item equating methodology.

Key criteria for judging the performance of items were item fit statistics (measured by weighted mean square and point-biserial correlations) and item performance (illustrated by item characteristic curves and multiple-choice distractor curves).

Chapter 5 of this report provides more detail on how item performance was investigated using these measures. The procedures employed were very similar, whether they were undertaken at the time of trial or after the NAPLAN tests.

In addition to the fit of the items, items were tested for DIF. The Rasch model assumes that the probability of responding correctly to an item is only dependent on a person’s ability and not on any group membership. DIF is the violation of this assumption. For example, if a group of boys and a group of girls have the same mean ability, but the probability of success on an item for the girls is higher (or lower) than the probability of success for the boys, then the item displays gender DIF. DIF does not refer to the difference in raw percentages correct for the groups, since these differences could be due to the fact that the groups have varying abilities. In other words, DIF examines the performance of a group on an item relative to the group’s performance on other items. For the NAPLAN item trial, items were tested only for gender DIF, gender being ascertained through student responses to a survey item; other demographic data is not available for trial students.

Items were flagged as potentially exhibiting DIF if the interaction term was significantly different from zero at the 95% confidence level, and the difference in difficulty between genders was greater than 1.0 logits.

Content experts inspected these items to determine potential reasons for the observed bias. The items are not automatically removed based on statistical evidence. Items are discarded only where the psychometric evidence points to an item issue that is confirmed as actual bias by the content experts’ review.

The results emerging from the analysis provided a pool of psychometrically sound items to populate the “test-ready” item bank from which test managers can select items for inclusion in future NAPLAN tests. Of the items trialled, over 90% were found to be acceptable in each domain. This is a result of the robust item development, review and quality assurance processes.

Psychometric analysis of item trial data: writing

The marking data was analysed using the partial credit model (Masters 1982) to identify the difficulty of each task, and of each of the 10 writing criteria for each task. All year levels were analysed together, since all tasks were administered to all year levels.

This psychometric analysis provided evidence of which tasks were most suitable for administration at each year level.

The NTWG gave advice regarding the final sequence and allocation of writing tasks. This informed the design of the NAPLAN writing tests for 2024, as well as which tasks could be held in reserve for future cycles.

Chapter 3: Test construction

The aim of this chapter is to describe the design and construction of NAPLAN 2023 tests. 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 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 due to an imbalance between their ability and the difficulty of the test.
- A wider range of aspects of the curriculum can be tested. While each student answers approximately 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 1. This figure shows a design with 6 nodes: A, B, C, D, E and F. Each node comprises 3 testlets (for example, 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.

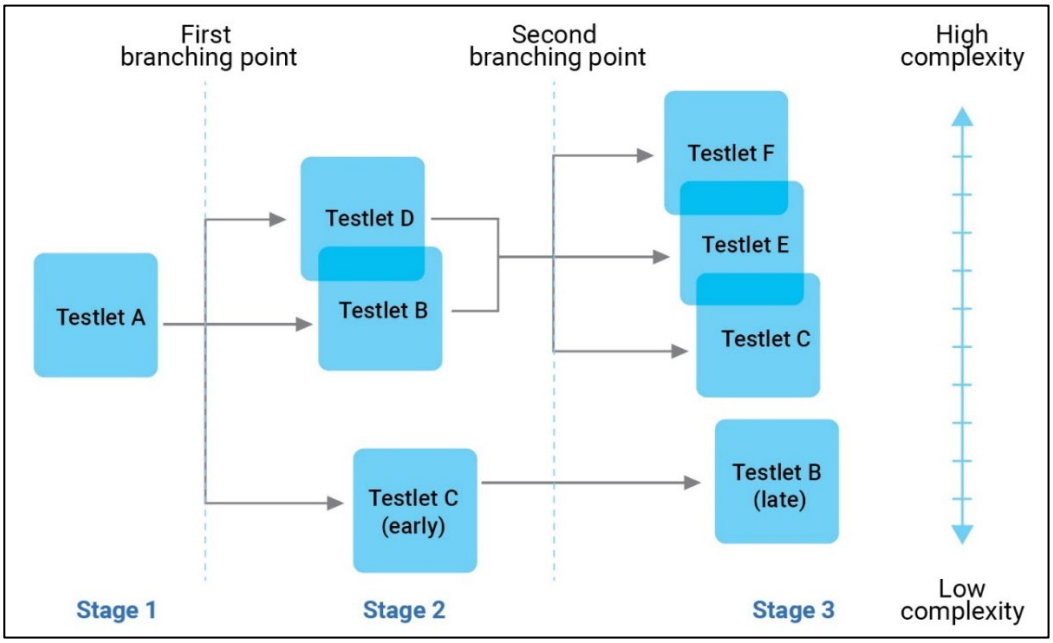


Figure 1: The multistage tailored test design for numeracy, reading, and grammar and punctuation

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 proficiency level), while a student who answers the same number of questions correctly, but follows a less complex pathway, is more likely to 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 189 different test pathways that each student could take, 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 tests include 2 sections in testlet A: a non-calculator section followed by a calculator-allowed section. An online calculator is available to students after completing the non-calculator section of the test. Students were advised that they could not return to the non-calculator section once they had moved to the calculator section.

The conventions of language test includes a spelling section and a grammar and punctuation section, each with 2 branching points. Students were advised that they could not return to the spelling section once they had moved to grammar and punctuation.

As noted above, the grammar and punctuation section of the conventions of language test has the same multistage 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 conventions of language test design is illustrated in Figure 2.

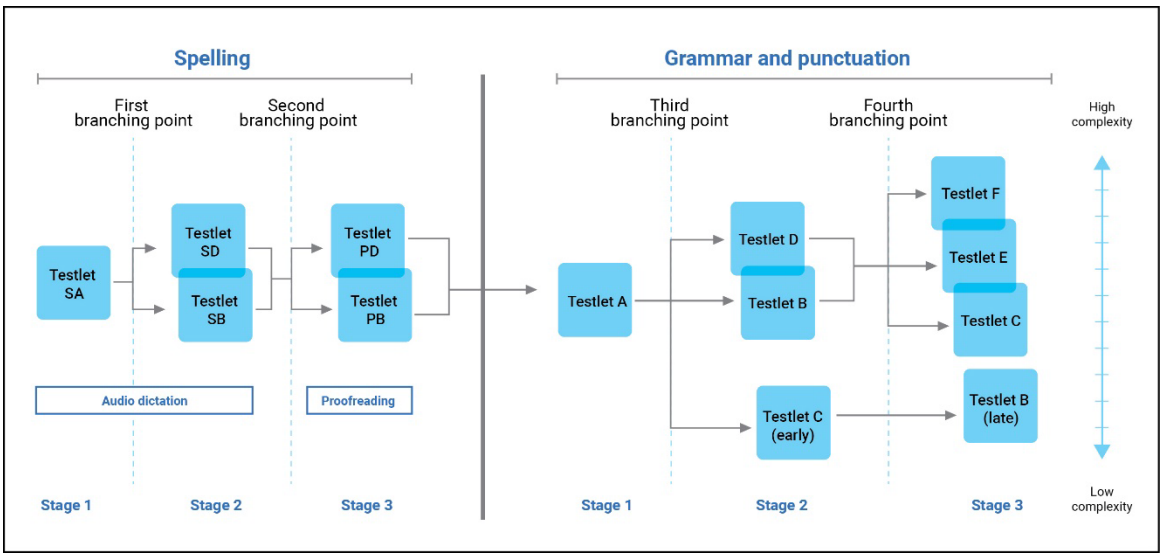


Figure 2: Online test design for conventions of language

As Figure 2 shows, the first 2 stages of the spelling section are focused on audio dictation 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

Items were selected for the 2023 NAPLAN tests based on their performance in past item trials or in the 2022 NAPLAN tests. Skills, curriculum strands and other aspects of curriculum content were balanced across nodes and testlets. When constructing tests, the choice and placement of link items were usually considered before other criteria. Link items are used to ensure that comparisons can be made between year levels, and in 2023 to construct the new measurement scales. Details of these processes are set out in Chapters 5 and 6.

In considering the selection of items from previous NAPLAN assessments, the guidelines shown below were followed:

- a weighted mean square fit between 0.8 and 1.2 (ideally between 0.9 and 1.1)
- balance of gender DIF across the set of link items as it is in the tests as a whole
- item difficulty between -2.5 and 2.5 logits (-4 and 4 logits for spelling, which has a wider scale)
- placement of items as close as possible to the same position in the previous NAPLAN administration (plus or minus 10, or ideally 5)
- placement of links between year levels as close as possible to the same position in both year levels (plus or minus 10, or ideally 5, adjusted for relative position where tests have different lengths)
- representativeness of items to the balance of Australian Curriculum strands in the tests
- even distribution of link items across nodes and testlets, unless constrained by test design.

Test length

Table 8 to Table 10 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 8: NAPLAN online numeracy test: number of items and time available

		Items per testlet	Total test items	Time available
Year 3		12	36	45 minutes
Year 5		14	42	50 minutes
Year 7	NC[1]	16 items x ½ testlet (8 items)	48	65 minutes
	CA[2]	16 items x 2 ½ testlets (40 items)		
Year 9	NC	16 items x ½ testlet (8 items)	48	65 minutes
	CA	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 9: NAPLAN online reading test: number of items and time available

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

Subdomain	Items per testlet	Items per section	Total test items	Time available
Spelling	7 items per stage 1 testlet (audio dictation) 9 items per stage 2 testlet (audio dictation) 9 items per stage 3 testlet (proofreading)	25	52	45 minutes
Grammar and punctuation	9 items per testlet	27		

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.

[1] NC – non-calculator
[2] CA – calculator-allowed

For reading, in general, the unit[3] with the lower average difficulty was presented first in each testlet and the unit with the higher average difficulty was presented last. Table 11 to Table 14 outline the predefined difficulty ranges in logits and average difficulty for the testlets in each test.

Table 11: NAPLAN online numeracy: predefined difficulty parameters for each testlet

Testlet	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 12: NAPLAN online reading: predefined difficulty parameters for each testlet

Testlet	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	3.0	1.3

Table 13: NAPLAN online spelling: predefined difficulty parameters for each testlet

Testlet	Lower bound	Upper bound	Average
SA	-3.0	2.0	-0.5
SB	-4.0	1.0	-1.0
SD	-1.0	4.0	1.0
PB	-5.0	1.0	-1.5
PD	-1.0	5.0	1.5

[3] A reading unit comprises one stimulus text with 4–7 items related to that stimulus text.

Table 14: NAPLAN online grammar and punctuation: predefined difficulty parameters for each testlet

Testlet	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	1.2

Item types for online tests

The numeracy tests contained items of the following formats: multiple-choice(s), text entry (constructed response) and technology-enhanced items.

The reading tests, and the grammar and punctuation section of the convention of language test included multiple-choice(s) and technology-enhanced items only.

In the spelling section of the conventions of language test, all items were text entry (constructed response).

Table 15 to Table 17 show the final distribution of item types in the suite of items at each year level.

Table 15: NAPLAN online numeracy: item types by year level

	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in suite
Year 3	118	38	60	216
Year 5	131	55	66	252
Year 7	154	56	78	288
Year 9	154	57	77	288

Table 16: NAPLAN online reading: item types by year level

	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in suite
Year 3	188	-	46	234
Year 5	201	-	33	234
Year 7	250	-	38	288
Year 9	249	-	39	288

Table 17: NAPLAN online conventions of language: item types 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	129	0	129
Spelling year 5	0	129	0	129
Spelling year 7	0	129	0	129
Spelling year 9	0	129	0	129
Grammar and punctuation year 3	80	0	82	162
Grammar and punctuation year 5	66	0	96	162
Grammar and punctuation year 7	67	0	95	162
Grammar and punctuation year 9	70	0	92	162

Numeracy test content

Items are written to cover the Australian Curriculum: Mathematics in 2 ways:

- maintaining a balance of items from each content strand (Number and Algebra, Measurement and Geometry, Statistics and Probability).
- maintaining a balance of proficiencies (Fluency, Understanding, Problem-solving, Reasoning).

Typically, the proportion of items assessing Problem-solving and Reasoning will be higher for the more complex test pathways than for the test as a whole, while the less complex test pathways will have higher proportions of items assessing Fluency and Understanding.

The test content proportions for numeracy are shown in Table 18 to Table 21. Target ranges refer to the overall test proportions; pathway proportions vary by complexity.

Table 18: NAPLAN numeracy Year 3 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Number and Algebra	50–60%	54%	55%	54%	53%	54%
Measurement and Geometry	25–35%	31%	31%	31%	31%	31%
Statistics and Probability	10–20%	15%	14%	16%	16%	16%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Fluency	15–25%	20%	26%	23%	19%	15%
Understanding	25–35%	32%	35%	30%	27%	26%
Problem-solving	25–35%	28%	21%	27%	35%	36%
Reasoning	15–25%	19%	18%	20%	19%	23%

Table 19: NAPLAN numeracy Year 5 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Number and Algebra	50–60%	53%	52%	52%	52%	52%
Measurement and Geometry	25–35%	30%	32%	31%	31%	31%
Statistics and Probability	10–20%	17%	17%	17%	17%	17%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Fluency	15–25%	19%	20%	16%	20%	17%
Understanding	25–35%	28%	33%	26%	26%	24%
Problem-solving	25–35%	33%	30%	36%	31%	33%
Reasoning	15–25%	21%	17%	22%	23%	26%

Table 20: NAPLAN numeracy Year 7 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Number and Algebra	50–60%	56%	56%	56%	56%	56%
Measurement and Geometry	25–35%	28%	29%	29%	30%	30%
Statistics and Probability	10–20%	16%	15%	15%	15%	15%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Fluency	15–25%	21%	23%	23%	24%	18%
Understanding	25–35%	31%	37%	30%	25%	25%
Problem-solving	25–35%	27%	18%	24%	31%	35%
Reasoning	15–25%	20%	22%	24%	21%	22%

Table 21: NAPLAN numeracy Year 9 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Number and Algebra	50–60%	56%	56%	56%	56%	56%
Measurement and Geometry	25–35%	31%	29%	29%	30%	30%
Statistics and Probability	10–20%	14%	15%	15%	15%	15%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Fluency	15–25%	23%	25%	23%	26%	22%
Understanding	25–35%	28%	37%	28%	26%	23%
Problem-solving	25–35%	30%	17%	26%	30%	38%
Reasoning	15–25%	19%	21%	23%	19%	17%

Reading test content

The reading tests primarily assess the Literacy strand of the Australian Curriculum: English, with a smaller focus on the Language and Literature strands.

They also contain a balance of items assessing the cognitive processes of Locating and identifying, Integrating and interpreting, and Analysing and evaluating. There is a greater focus on Analysing and evaluating in the secondary school years.

The more complex test pathways contain, on average, longer stimulus texts.

The test content proportions for reading are shown in Table 22 to Table 25. Target ranges refer to the overall test proportions; pathway proportions vary by complexity.

Table 22: NAPLAN reading Year 3 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Language	15–25%	23%	24%	22%	23%	21%
Literature	5–15%	6%	4%	6%	9%	9%
Literacy	60–80%	71%	72%	72%	68%	69%
Cognitive process	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Locating and identifying	30–60%	41%	48%	40%	37%	32%
Integrating and interpreting	35–60%	54%	52%	56%	56%	60%
Analysing and evaluating	0–15%	6%	0%	3%	8%	8%
Text content		**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Number of texts		-	7	6	6	6
Average word count		175	102	169	198	220

Table 23: NAPLAN reading Year 5 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Language	15–25%	18%	21%	18%	17%	20%
Literature	5–15%	11%	9%	12%	13%	11%
Literacy	60–80%	71%	71%	70%	70%	69%
Cognitive process	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Locating and identifying	30–60%	33%	40%	36%	26%	25%
Integrating and interpreting	35–60%	56%	51%	54%	61%	63%
Analysing and evaluating	0–15%	12%	9%	10%	13%	12%
Text content		**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Number of texts		-	6	6	6	6
Average word count		241	180	200	248	279

Table 24: NAPLAN reading Year 7 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Language	15–25%	18%	17%	17%	20%	22%
Literature	10–20%	12%	9%	13%	15%	13%
Literacy	55–75%	70%	74%	71%	65%	65%
Cognitive process	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Locating and identifying	15–45%	29%	39%	33%	25%	21%
Integrating and interpreting	40–65%	57%	53%	58%	60%	61%
Analysing and evaluating	5–30%	14%	8%	9%	15%	18%
Text content		**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Number of texts		-	9	9	9	9
Average word count		284	239	281	300	327

Table 25: NAPLAN reading Year 9 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Language	15–25%	25%	26%	26%	24%	22%
Literature	10–20%	11%	7%	10%	13%	13%
Literacy	55–75%	64%	67%	64%	63%	65%
Cognitive process	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Locating and identifying	15–45%	23%	32%	26%	19%	16%
Integrating and interpreting	40–65%	58%	57%	58%	56%	57%
Analysing and evaluating	5–30%	19%	11%	15%	24%	27%
Text content		**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Number of texts		-	9	9	9	9
Average word count		301	242	295	315	336

Conventions of language test content

The spelling section of the conventions of language test assesses spelling in 3 ways:

- audio dictation, where students hear a recording of the word, along with a sentence where the word is used in context, then students are asked to correctly spell the word
- proofreading (mistake identified), where a sentence contains a misspelled word that is highlighted for the student; students are asked to correctly spell the word
- proofreading (mistake not identified), where a sentence contains a misspelled word that is not highlighted for the student; students are asked to identify which word is misspelled and spell it correctly.

The grammar and punctuation section of the conventions of language test is divided in a ratio of approximately 70:30 between items assessing grammar and items assessing punctuation.

The conventions of language test assesses the Language strand of the Australian Curriculum: English almost exclusively.

The test content proportions for conventions of language are shown in Table 26 to Table 33, divided to show spelling separately from grammar and punctuation.

Table 26: NAPLAN spelling Year 3 test content by pathway

Item type	Target range	Overall	SASBPB	SASBPD	SASDPB	SASDPD
Audio dictation	55–65%	58%	64%	64%	64%	64%
Mistake identified	15–25%	23%	21%	19%	21%	19%
Mistake not identified	15–25%	19%	15%	17%	15%	17%

Table 27: NAPLAN grammar and punctuation Year 3 test content by pathway

Subdomain	Target range	Overall	ABC	ABE	ADE	ADF
Grammar	65–75%	66%	67%	67%	67%	65%
Punctuation	25–35%	34%	33%	33%	33%	35%

Table 28: NAPLAN spelling Year 5 test content by pathway

Item type	Target range	Overall	SASBPB	SASBPD	SASDPB	SASDPD
Audio dictation	55–65%	58%	64%	64%	64%	64%
Mistake identified	15–25%	21%	20%	16%	20%	16%
Mistake not identified	15–25%	21%	16%	20%	16%	20%

Table 29: NAPLAN grammar and punctuation Year 5 test content by pathway

Subdomain	Target range	Overall	ABC	ABE	ADE	ADF
Grammar	65–75%	67%	67%	67%	68%	67%
Punctuation	25–35%	33%	33%	33%	32%	33%

Table 30: NAPLAN spelling Year 7 test content by pathway

Item type	Target range	Overall	SASBPB	SASBPD	SASDPB	SASDPD
Audio dictation	55–65%	58%	64%	64%	64%	64%
Mistake identified	15–25%	21%	19%	17%	19%	17%
Mistake not identified	15–25%	21%	17%	19%	17%	19%

Table 31: NAPLAN grammar and punctuation Year 7 test content by pathway

Subdomain	Target range	Overall	ABC	ABE	ADE	ADF
Grammar	65–75%	66%	67%	67%	67%	68%
Punctuation	25–35%	34%	33%	33%	33%	32%

Table 32: NAPLAN spelling Year 9 test content by pathway

Item type	Target range	Overall	SASBPB	SASBPD	SASDPB	SASDPD
Audio dictation	55–65%	58%	64%	64%	64%	64%
Mistake identified	15–25%	19%	20%	12%	20%	12%
Mistake not identified	15–25%	23%	16%	24%	16%	24%

Table 33: NAPLAN grammar and punctuation Year 9 test content by pathway

Subdomain	Target range	Overall	ABC	ABE	ADE	ADF
Grammar	65–75%	65%	67%	68%	65%	64%
Punctuation	25–35%	35%	33%	32%	35%	36%

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, language conventions (spelling, grammar and punctuation), and writing. All students who sat paper-based tests completed the same set of test items.

All students in Year 3 complete writing tests on paper. For other domains, now that NAPLAN has transitioned to full online delivery, the paper tests are considered to be an alternative format, and administered only for an agreed subset of schools. Typically, only between 200 and 500 students sit each of these tests.

Items in all tests were distributed across approximately the same difficulty range as the online tests, except that the tailored test design allows slightly easier items to be administered in testlet C and harder items in testlet F.

Items were ordered approximately from easiest to hardest for numeracy, and within each section of the language conventions tests. For reading, the average of each unit (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 items preceded the non-calculator items.

The number of items and time available in the paper tests is shown in Table 34 to Table 36.

Table 34. 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	40	48	55 minutes	65 minutes
Year 7 NC	8		10 minutes	
Year 9 CA	40	48	55 minutes	65 minutes
Year 9 NC	8		10 minutes	

Table 35. 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 36. NAPLAN language conventions paper test number of items and time available

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

The content of each paper test has a similar balance to a single pathway of the corresponding online test. Specifications are shown in Table 37 to Table 39.

Table 37: Test content – numeracy paper tests

Strand	Target range	Year 3	Year 5	Year 7	Year 9
Number and Algebra	50–60%	56%	55%	56%	54%
Measurement and Geometry	25–35%	31%	29%	29%	31%
Statistics and Probability	10–20%	14%	17%	15%	15%
Proficiency	**Target range**	**Year 3**	**Year 5**	**Year 7**	**Year 9**
Fluency	15–25%	19%	19%	17%	21%
Understanding	25–35%	31%	31%	31%	31%
Problem-solving	25–35%	31%	31%	31%	29%
Reasoning	15–25%	19%	19%	21%	19%

Table 38: Test content – reading paper tests

Strand	Target range	Year 3	Year 5	Year 7	Year 9
Language	10–20%	22%	26%	17%	54%
Literature	10–20%	5%	10%	19%	31%
Literacy	50–70%	73%	64%	64%	15%
Cognitive Process	**Target range**	**Year 3**	**Year 5**	**Year 7**	**Year 9**
Locating and identifying	20–40%	32%	33%	21%	30%
Integrating and interpreting	40–60%	57%	51%	57%	49%
Analysing and evaluating	20–40%	11%	15%	21%	21%
Text Content		**Year 3**	**Year 5**	**Year 7**	**Year 9**
Stimulus texts		6	6	8	8
Average word count		195	243	313	309

Table 39: Test content – language conventions paper tests

Item type	Target range	Year 3	Year 5	Year 7	Year 9
Mistake identified	-	48%	48%	48%	48%
Mistake not identified	-	52%	52%	52%	52%
Subdomain	**Target range**	**Year 3**	**Year 5**	**Year 7**	**Year 9**
Grammar	65–75%	76%	76%	68%	68%
Punctuation	25–35%	24%	24%	32%	32%

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 2023, 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, while the vast majority of students in Years 5 to 9 completed an online test.

In 2023, 4 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 required in 2023. Two of the 4 prompts were assigned to the Years 3 and 5 tests, and 2 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 40). Each prompt has closely scripted scaffolding, or instructions. All prompts had been trialled and the prompts selected for the 2023 tests had been shown to function similarly at the allocated year levels.

Table 40. NAPLAN writing prompt designation schedule according to test day

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

All students were given 40 minutes to respond to the prompt. For the online tests, the timing commences before the students sees or hears 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 between the 3 stages of writing: planning, writing and editing, although students can use their time as they choose.

Table 41. 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. For each genre of writing, the same marking guide is used to assess students' writing at all year levels and

across calendar years, 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 using the partial credit model (Masters 1982). The 10 criteria with the associated number of score categories are shown in Table 42 and Table 43.

Table 42. 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 43. NAPLAN narrative marking criteria and score categories

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

Marking processes

Test administration authorities in each state and territory were responsible for marking student scripts from within their jurisdiction. Three jurisdictions – Queensland, South Australia and Western Australia – ran their own marking operations. The Australian Capital Territory scripts were marked through the New South Wales marking operation, and Victoria coordinated a marking operation for Victoria, Tasmania and the Northern Territory. In total, over 1 million student scripts were marked nationally across the 5 marking operations. In 2023, approximately 2000 markers were employed nationally. 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.

Training of markers

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.

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 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 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 for use in training new and experienced markers and leaders. The course was delivered through a Learning Management System. 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 Team[4] (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.

Markers scored training and practice scripts before commencing marking. Their scores were first checked to ensure comparability with the expert scores.

Quality assurance of marking

Daily control scripts were used throughout the marking period to monitor individual marker accuracy and collect data on the national consistency of marking. The first control script was issued when the first marking centre commenced marking, and the last control was issued on the final day of the last marking centre. However, as each jurisdiction has a slightly different marking window, not all controls were completed by all centres. Each day of the marking period, control script scores from each jurisdiction were provided to ACARA and aggregated. A summary marking performance report for each control script was provided to each jurisdiction so they could compare their own marking accuracy for that control script with that of other jurisdictions.

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 occurred for each marker and was done by a group leader, a centre leader, or other experienced, expert marker appointed by the test administration authority 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%).

Following administration of the national daily control scripts and implementation of local check-marking, jurisdictions used a variety of strategies and analytics 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 44 below).

Table 44. National marking protocols

	Monitor	Discuss/Re-train	Negotiate future marking
Total score	3–4 points discrepant	5–8 points discrepant	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 one occasion	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

[4] 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.

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 and imported to the platform 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 and Breithaupt 2006). In the AMI method, the intersection of the testlet information curves for the 2 adjacent testlets represents the branching cutoff. This approach is analogous to the maximum information item selection method in CAT (Breithaupt and 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 method was the primary guide for the development of the testlet targeting and boundaries. In addition, the following conditions were applied:

- The initial testlet (A or SA) should provide a sufficient number of easy entry items to engage students at the lower end of the ability scale.
- Where the tailored test design contains 2 nodes (B and D, SB and SD, or PB and PD), 50% of students should be directed to each node, plus or minus 10 percentage points.
- Where the tailored test design contains 3 nodes (C, E and F), 25% should be directed to each of C and F, plus or minus 5 percentage points, and 50% to E, plus or minus 10 percentage points.

While the AMI method is applied for most branching rules, there are 2 exceptions:

- Students are branched from A directly to C when they score between 0 and 2 (Years 3 and 5) or 0 and 3 (Years 7 and 9) on testlet A. This rule is imposed in order to preserve the ACB pathway for the students who are most likely to benefit from early delivery of the easiest items in the test.
- The branching rules to testlet F are set as equal to the AMI cut-score plus 1. Reports of student experience from the first few cycles of the NAPLAN adaptive tests indicated that the unadjusted AMI cut-scores required difficulty specifications, which were too onerous for students whose performance placed them near the boundary of testlets E and F.

There is an iterative process of developing tests that meet these conditions. The tests are built to the specifications set out earlier in this chapter, subject to constraints of content and item availability, and their performance is then verified by simulations.

Previous NAPLAN technical reports (2018 to 2022) provide worked examples of how branching rules are set in each of the NAPLAN multi-stage test designs (1 – 2 – 3 as in numeracy, reading, and grammar and punctuation, or 1 – 2 – 2 as in spelling).

Results of branching

This section describes how different pathways were used in NAPLAN 2023 online tests, taking 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 is shown in Figure 3. The total percentage of students directed to testlet B was 52.9% and to testlet D was 47.1%. The total percentage of students directed to testlet C was 24.2%, to testlet E was 51.8% and to testlet F was 24.0%. These percentages are all within the tolerances set out above. The fact that the achieved percentages remain close to the simulated percentages is an indication that the performance of most items in the 2023 tests was very similar to their performance at trial or in previous cycles.

Note that very low proportions of students are directed to the ADC and ABF pathways. These are designed as corrective pathways and are needed only if students demonstrate a very different level of performance in their second testlet to their first.

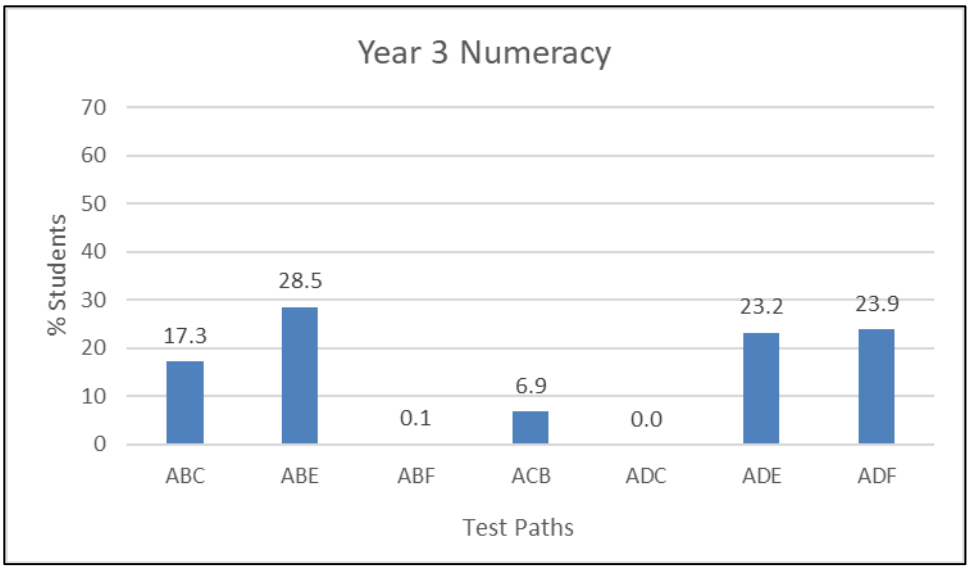


Figure 3. Percentage of students assigned to each pathway in Year 3 numeracy

Ability distributions by pathway are illustrated in Figure 4. 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 4 also shows that pathways overlapped in abilities.

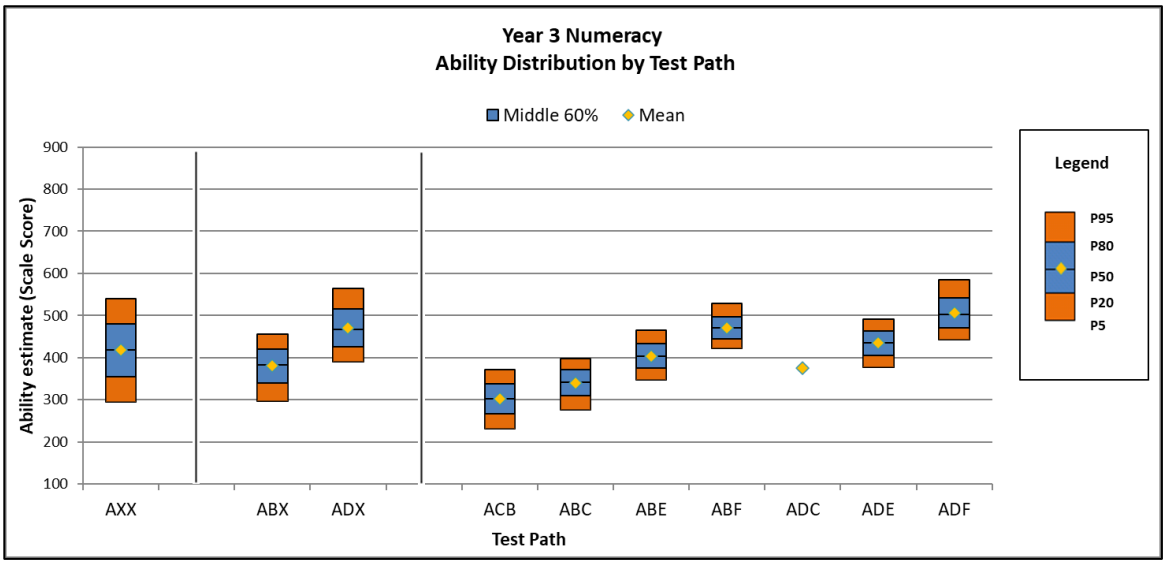


Figure 4. Ability distribution by pathway for Year 3 numeracy

Chapter 4: Data collection and preparation

This chapter describes data collection and delivery, data validation and data preparation for NAPLAN 2023. The chapter focuses on how data for online and paper tests are collected by test administration authorities (TAAs) from each jurisdiction and delivered to ACARA, as well as describing how data is validated and prepared by the contractor before performing the analysis.

Data collection, cleaning and validation

TAAs are responsible for:

1. implementing and administering the NAPLAN tests in their jurisdiction, following the National Protocols for Test Administration provided by ACARA
2. collecting NAPLAN test and student background data in their jurisdiction and 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 Characteristics*[5] is to guide 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 intended 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
- Aboriginal and/or Torres Strait Islander status
- Parental school education
- Parental non-school education
- Parental occupation group
- Language other than English spoken at home.

Test response data was delivered to the contractor in 5 main batches:

- delivery of online test data, sequentially by test domain, including both scored and raw response data, which is used for item calibration
- delivery of the NAP Analysis Extracts (NAE) for preliminary analysis and to generate initial Student and School Summary Reports (SSSRs)
- delivery of the Calibration extracts to calibrate writing criteria
- delivery of the Student Master File (SMF-2b) and Item Response File (IRF-1b), referred to as stage 1 data, and the NAE extracts, to generate the Individual Student Reports (ISRs) and final SSSRs.
- delivery of the Student Master File (SMF-3b) and Item Response File (IRF-2b), referred to as stage 2 data, and the NAE extracts to produce the NAPLAN 2023 National Results.

Online tests

Education Services Australia (ESA) managed the online national assessment platform (the platform) through which the NAPLAN 2023 online tests were delivered. The Australian Council for Educational

[5] www.acara.edu.au/reporting/data-standards-manual-student-background-characteristics

Research (ACER), as the analysis contractor in 2023, received the online test data extracted from the platform. Data was provided directly from ACARA, by domain, as each became available.

Paper tests

Data collection for paper tests was undertaken by the TAAs in each of the jurisdictions. Paper Item Response Files (IRF) were used to deliver paper data to ACARA.

Data cleaning and validation

ACARA used a systematic process of data validation to ensure that each dataset was consistent with national code frames and data dictionaries. There were several types of exception rules implemented in the NAPLAN Quality Assurance (QA) scripts to identify issues categorised according to their structural, inconsistent, advisory and statistical impact. A list of the exception rules is included in Appendix N.

The tight timeline between the online assessments and the delivery of School and Student Summary Reports (SSSRs) necessitated quality assurance checks of online data extracted from the platform, along with the SMF and IRF, commencing after the first week of testing. Preparation for data checking and management, and for the analysis of online data, followed the quality assurance measures. Data integrity checking involved verifying 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.

Any concerns raised during this process 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 contractor prior to data analysis. The recoding rules depend on participation status, and are shown in Table 45.

Table 45: Rules for data coding

Participation code	Data recoding rules
P – present	**Data received** • A data string of responses to all items in the test (whether administered to students or not) was expected from the TAA. • In this data string, any embedded missing responses were indicated with a 9. • For items in testlets that were not administered to the student, responses were coded as 8. • For paper tests only, invalid responses such as selection of an incorrect number of multiple-choice options were indicated with a 7. **Data treatment** • Trailing missing responses were 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* for the purpose of item calibration. These not-reached responses were treated as *incorrect* for the final estimation of student abilities. Any embedded missing responses within the data string were kept as a 9. • Invalid paper test responses were recoded from 7 to 0 (incorrect).

	• For the online test data, responses for items in those testlets that were not administered to the students were recoded from 8 to R. • Students who were present but did not attempt any question ("non-attempts") can be identified by having a string of 9s for administered testlets and 8s elsewhere. Their item responses were recoded to a string of Rs.
A – absent	**Data received** • A data string of all 8s for that test was expected from the TAA. See National Protocols for Test Administration, section 5.4. **Data treatment** • Item response data was recoded to a string of Rs and excluded from the item calibration.
S – sanctioned abandonment	**Data received** • This participation code is specifically used to indicate students who unexpectedly abandon the test due to illness or injury. Since some responses may have been provided before abandonment, the TAA may have supplied a response string containing codes other than 8. See National Protocols for Test Administration, section 5.5. **Data treatment** • Item response data was recoded to a string of Rs and excluded from the item calibration.
W – withdrawn	**Data received** • A data string of all 8s for that test was expected from the TAA. See National Protocols for Test Administration, section 5.3. **Data treatment** • Item response data is recoded to a string of Rs and excluded from the item calibration.
E – exempt C – cancelled N – no longer enrolled	**Data received** • A data string of all 8s for that test was expected from the TAA. See National Protocols for Test Administration, section 5.2. **Data treatment** • Item response data is recoded to a string of Rs and excluded from the item calibration.

After recoding, the data for unscored items can be summarised as follows:

9	embedded missing
M	not-reached
R	not administered/not attempted.

Responses to scored items are generally coded as 0 (incorrect) or 1 (correct). The exception to this is during the item calibration phase, for multiple-choice items only, where responses are coded, for example, as 1–5 for a 5-option item. This allows analysis of each option by comparison with the item keys.

Data for partial-credit items (each of the 10 writing criteria) was indicated by ordered categories starting with 0 up to the maximum possible value.

Students who did not attempt all 3 testlets of the online test 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 2. 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 46: Pathway assignment rules to incomplete online tests

Domain	Last item attempted		Assigned pathway
Numeracy, reading, grammar and punctuation	Stage 1	A	ABE
Numeracy, reading, grammar and punctuation	Stage 2	B	ABE
Numeracy, reading, grammar and punctuation	Stage 2	C	ACB
Numeracy, reading, grammar and punctuation	Stage 2	D	ADE
Spelling	Stage 1	SA	SASBPB
Spelling	Stage 2	SB	SASBPB
Spelling	Stage 2	SD	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 the percentage of trailing missing responses across all students for a given online test pathway.

Figure 5 to Figure 8 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 only 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). However, similar patterns of trailing missing responses were found in other pathways.

Grammar and punctuation had the lowest trailing missing rates of any domain. Across test pathways, the most difficult test pathway (A1-D1-F1) and the test pathway for the lowest-performing students (A1-C1-B1) tended to have the highest trailing missing rates. Patterns of trailing missing differed across year levels in each domain: Year 3 or Year 9 commonly showed higher rates, but in Numeracy it was Year 5.

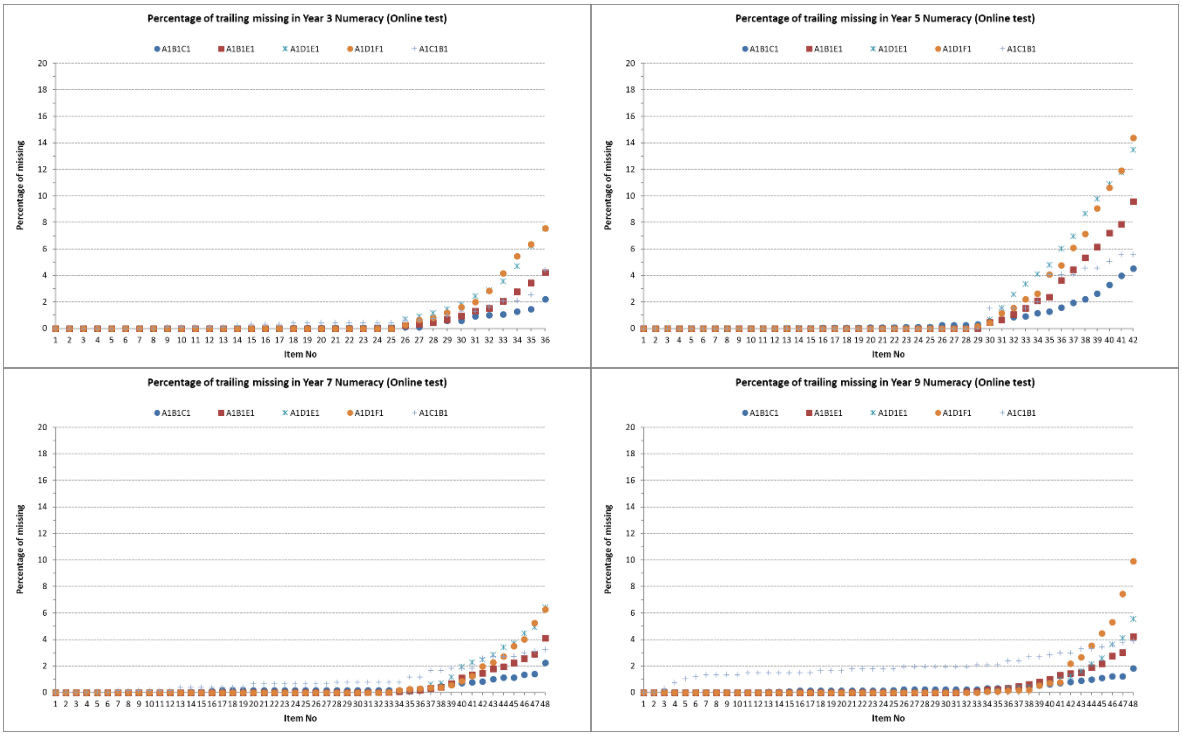


Figure 5: Trailing missing percentage in numeracy

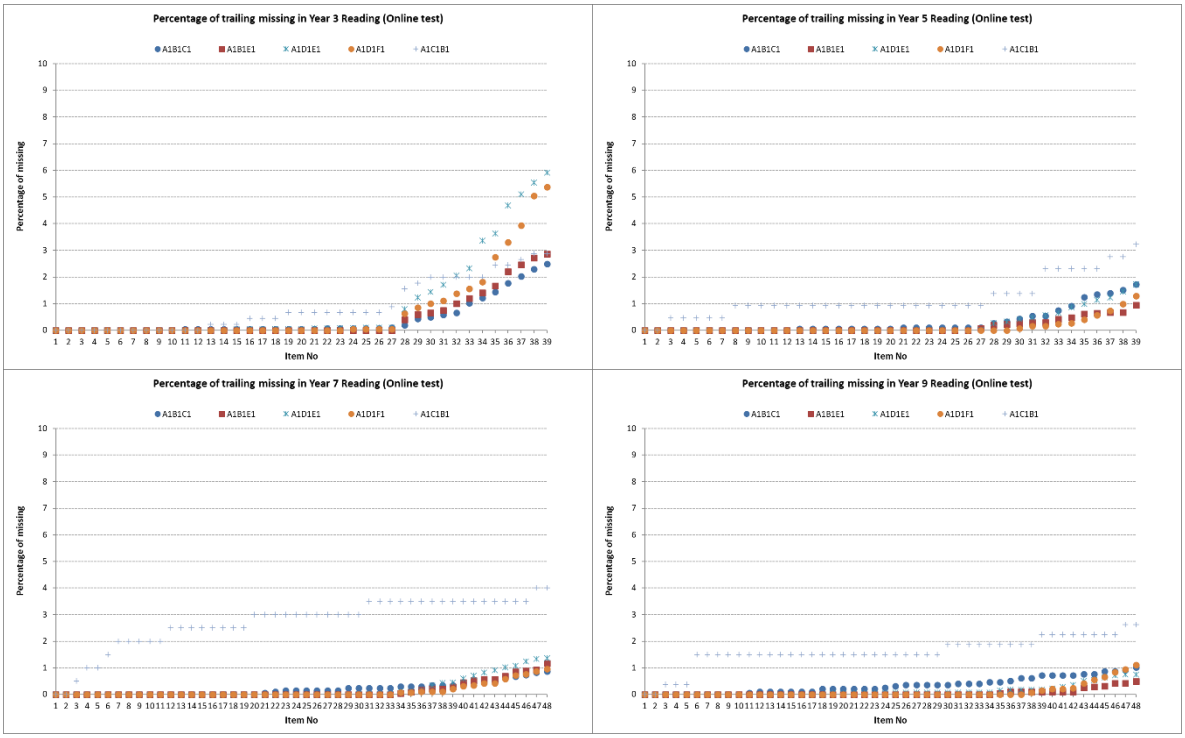


Figure 6: Trailing missing percentage in reading

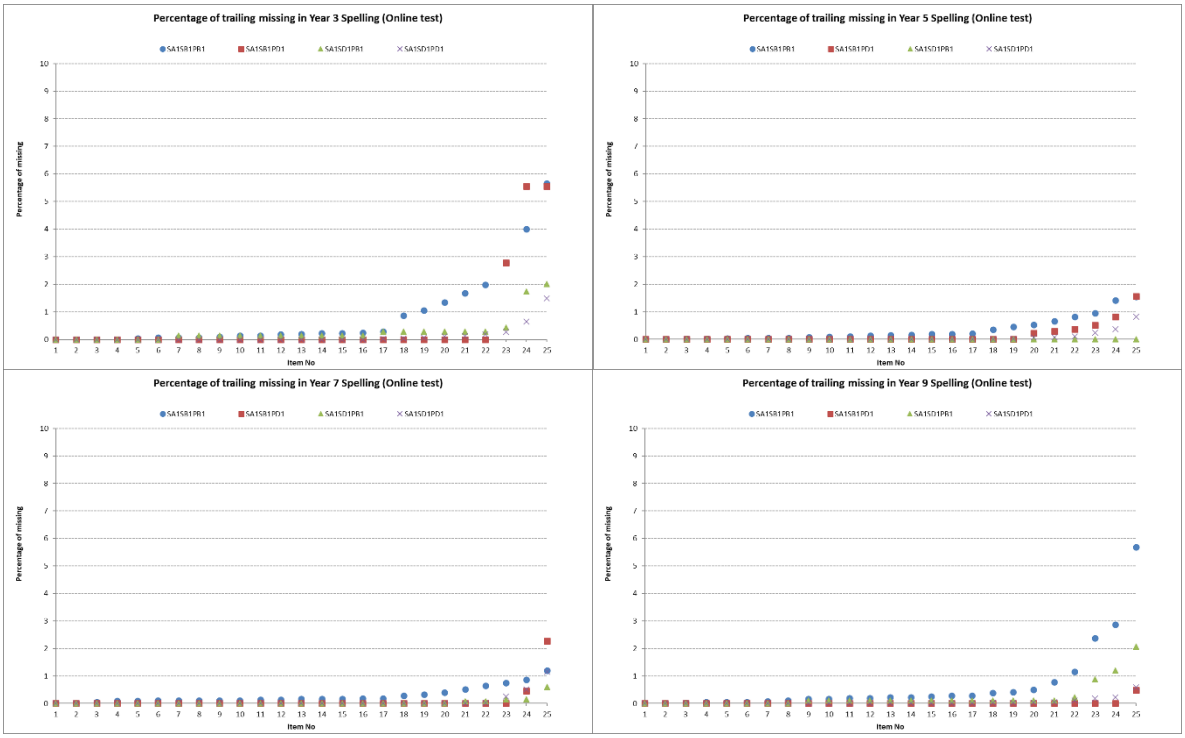


Figure 7: Trailing missing percentage in spelling

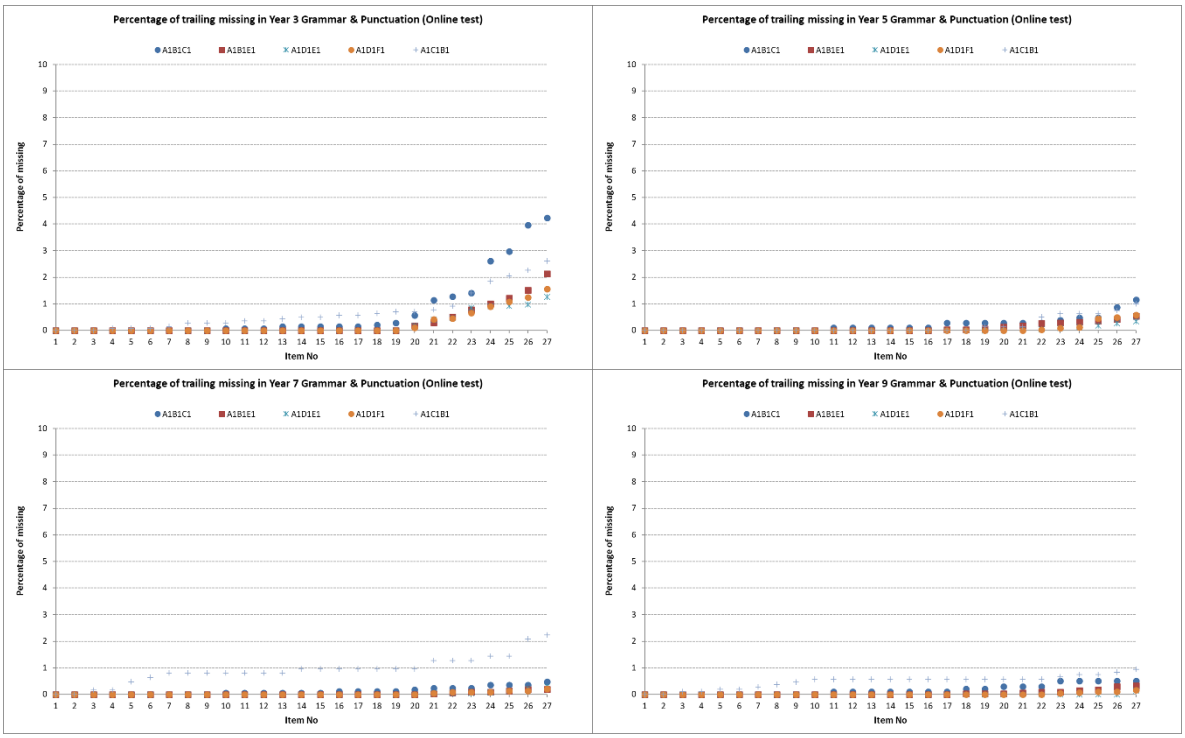


Figure 8: Trailing missing percentage in grammar and punctuation

Final student participation rates

The participation category diagram for NAPLAN 2023, with the data file participation codes shown in parentheses, is shown in Figure 9. Participating students include present (assessed, non-attempts) and not present (exempt) students. Final student participation rates for NAPLAN 2023 are recorded in Table 47 by TAA, year level and domain. The participation rate technical standard was 90% participation in at least one test at national and jurisdictional level to ensure unbiased population statistics. Results in the National Report were annotated if the participation rate technical standard was not met. These percentages, shown in the "At least one test (%)" column, are coloured red in Table 47.

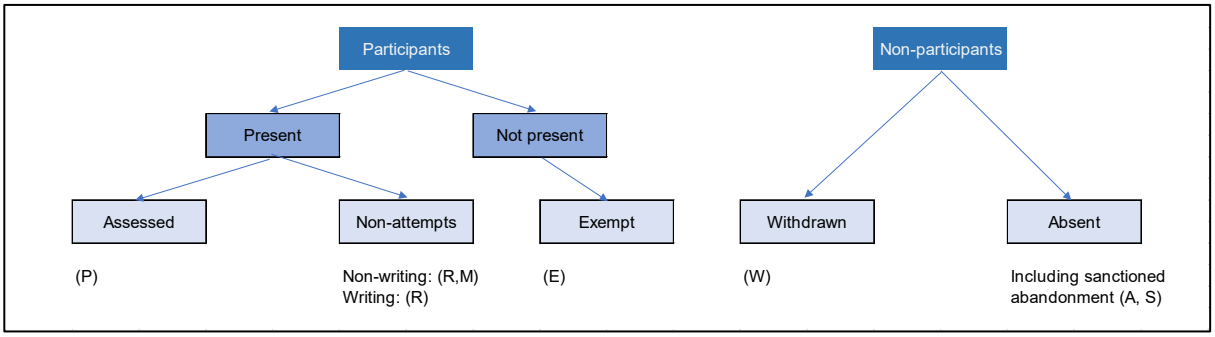


Figure 9: NAPLAN 2023: Participation categories

Table 47: Student participation rates

TAA	Year level	Numeracy (%)	Reading (%)	Writing (%)	Spelling (%)	Grammar and punctuation (%)	At least one test (%)
NSW	3	96.4	97.0	96.1	96.6	96.6	97.5
Vic.	3	95.1	95.5	94.6	94.9	94.9	96.5
Qld	3	92.1	93.0	92.4	92.3	92.3	94.2
WA	3	95.0	95.7	95.0	95.2	95.2	96.5
SA	3	94.6	95.2	94.0	94.7	94.7	96.0
Tas.	3	96.0	97.2	95.6	96.6	96.6	97.8
ACT	3	94.0	94.5	92.8	93.9	93.9	95.5
NT	3	80.4	82.3	80.5	80.4	80.4	86.1
Aus.	3	94.7	95.3	94.5	94.7	94.7	96.2
NSW	5	96.9	97.5	97.0	97.1	97.1	97.9
Vic.	5	95.6	96.1	95.8	95.6	95.6	97.0
Qld	5	92.4	93.5	93.2	92.7	92.7	94.5
WA	5	96.0	96.6	96.5	96.2	96.2	97.4
SA	5	95.3	96.0	95.4	95.5	95.5	96.7
Tas.	5	95.5	96.5	96.0	95.8	95.8	97.3
ACT	5	94.5	94.9	94.8	94.5	94.5	96.1
NT	5	81.7	83.2	83.7	82.5	82.5	87.6
Aus.	5	95.2	95.9	95.5	95.3	95.3	96.7
NSW	7	95.6	96.5	96.7	96.0	96.0	97.9
Vic.	7	94.9	95.8	95.8	95.1	95.1	97.3
Qld	7	88.8	89.8	90.5	88.9	88.9	92.5
WA	7	95.1	96.2	96.5	95.3	95.3	98.0
SA	7	93.6	94.8	94.9	94.1	94.1	96.4
Tas.	7	93.9	95.1	94.3	93.8	93.8	97.4
ACT	7	93.7	94.7	94.9	93.7	93.7	96.4
NT	7	78.8	80.7	81.3	78.2	78.2	85.4
Aus.	7	93.5	94.5	94.7	93.7	93.7	96.3
NSW	9	91.6	92.8	93.2	92.3	92.3	95.1
Vic.	9	90.6	91.6	91.7	90.5	90.5	94.1
Qld	9	80.0	81.3	82.2	80.2	80.2	85.0
WA	9	92.0	93.3	93.6	92.2	92.2	95.6
SA	9	88.9	90.4	90.6	89.4	89.4	92.9
Tas.	9	87.8	89.6	89.9	87.4	87.4	93.5
ACT	9	89.5	91.0	91.1	89.6	89.6	93.5
NT	9	73.1	75.1	75.7	73.4	73.4	79.9
Aus.	9	88.4	89.6	90.0	88.7	88.7	92.3

Chapter 5: Scaling methodology and outcomes

This chapter describes the processes and methodologies used in the NAPLAN 2023 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 2023 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_i). 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, such as for the NAPLAN writing assessment in this context, 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, τ_{ij} is a step parameter of score j on item i, and m_i is the maximum possible score for 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 (*Adams et al. 2020) was used. *ACER ConQuest* provides tools for the estimation of a variety of item response models and latent 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 item calibration of non-writing domains in each year level was extracted as soon as sufficient data was collected overall and in each jurisdiction. Typically the critical threshold was obtaining data from 1,000 students in the Northern Territory. For non-writing domains, the calibration sample contains student response data from the online tests only, for students who completed a full test pathway with no trailing missing responses. In total, the number of students included in the estimation of each domain was between 179,802 and 235,884 by year level.

For the 2023 NAPLAN online tests, the numeracy, reading, spelling, and grammar and punctuation tests were calibrated separately by domain and year level, resulting in 16 separate calibrations. For each of the

4 non-writing online tests, items from all testlets within a domain and a year level were calibrated in a concurrent analysis. In 2023, there was only a small number of students who participated in NAPLAN paper tests, and it was not possible to construct a representative national calibration sample, hence no paper test calibration was carried out. Since all questions in the paper tests are included in the online test, paper test item parameters were anchored to their values from the online test.

For 2023 writing, the resulting scripts from students who responded on paper (predominantly Year 3 students, with a small number of alternative-format tests delivered to students in other year levels) or online (all except those on paper) from different tasks were scored for each criterion using the same marking rubric based on 10 criteria. The scored writing data from Years 3, 5, 7 and 9 was calibrated concurrently based on the partial credit model (Eq. 2) with the latent distribution conditioned on year level and test mode. The vertical writing scale was constructed with this concurrent calibration across the 4 year levels. The reason for applying the concurrent calibration was that some scores did not occur for some year levels. Writing is calibrated only when all jurisdictions have completed marking; effectively, the whole population is available for calibration.

In the estimation of parameters, only students with complete test pathways were included in the non-writing calibration data. Students with an incomplete test pathway or with trailing missing responses (identified by 2 or more consecutive response codes of 9 at the end of the test) were excluded from the calibration data. Online items that were not included in a student's pathway and therefore not presented to students (responses were coded as R) were treated as *not administered* in all analyses, and embedded-missing responses (9) were treated as *incorrect* responses.

Senate weights were used for calibrating the online numeracy, reading, spelling, and grammar and punctuation tests to ensure each jurisdiction contributed equally to the calibration.

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 that 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, the senate weight was calculated by year level according to the equation above, thus equal representation of each jurisdiction in the calibration was achieved.

Review of test and item characteristics

The *ACER ConQuest* item analysis results for the NAPLAN 2023 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 classical summary statistics (for example, Mean) shown at the end of the item analysis results for the 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, and summary statistics based on item response theory (IRT) should be considered instead, as detailed in the following sections.

The Rasch item parameter estimates and fit statistics are summarised in Appendix C for the items in each of the 16 item pools for the numeracy, reading, spelling, and grammar and punctuation tests across 4 year levels. The item parameters shown in these tables are delta-centred for each test (that is, the mean of item difficulties is set to zero). The 95% confidence interval from *ACER ConQuest* output for the expected value of the infit mean square is also provided for each item.

Item Characteristic Curves (ICCs) for all items are shown in Appendix D. The ICC plot shows a comparison of the empirical ICC based on observations from ability groupings (broken line joining each dot) and the

expected model-based ICC (smooth line). The distance shown on each plot was constrained to be equal for each test node (generic testlet) to display the appropriate 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 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 48 shows the IRT-based reliabilities, calculated using weighted likelihood estimates (WLEs) or plausible values (EAP/PVs) for each 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 0.91 and 0.93 for the spelling tests, and between 0.81 and 0.84 for the grammar and punctuation tests. The EAP/PV reliabilities were between 0.89 and 0.94 for the numeracy tests, between 0.86 and 0.88 for the reading tests, between 0.87 and 0.93 for the spelling tests, and between 0.79 and 0.84 for the grammar and punctuation tests. The reliabilities for the writing test were 0.95 and 0.90 for WLE reliability and EAP/PV reliability, respectively. In general, the WLE reliability is equal to or higher than the EAP/PV reliability.

Table 48: Reliability (EAP/PV, WLE) for NAPLAN 2023 tests

	Numeracy		Reading		Spelling		Grammar and punctuation		Writing*	
Year level	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV
3	0.91	0.89	0.91	0.88	0.93	0.93	0.84	0.84	0.95	0.90
5	0.92	0.92	0.88	0.88	0.92	0.92	0.82	0.81		
7	0.94	0.90	0.91	0.86	0.92	0.87	0.81	0.79		
9	0.94	0.94	0.91	0.87	0.91	0.88	0.84	0.83		

*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 locations / 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 maps for each domain at each year level. It is important to note that the 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 tests, the 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 map, the mean of the case (student) 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 10 displays the map for the Year 3 numeracy test. This map indicates that the test was well-targeted to the average numeracy achievement level of the student group. The distribution of student abilities (each X represents approximately 326 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 11 with the latent ability distribution on the left-hand side. Figure 12 shows locations of the Thurstonian thresholds of each item, 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.

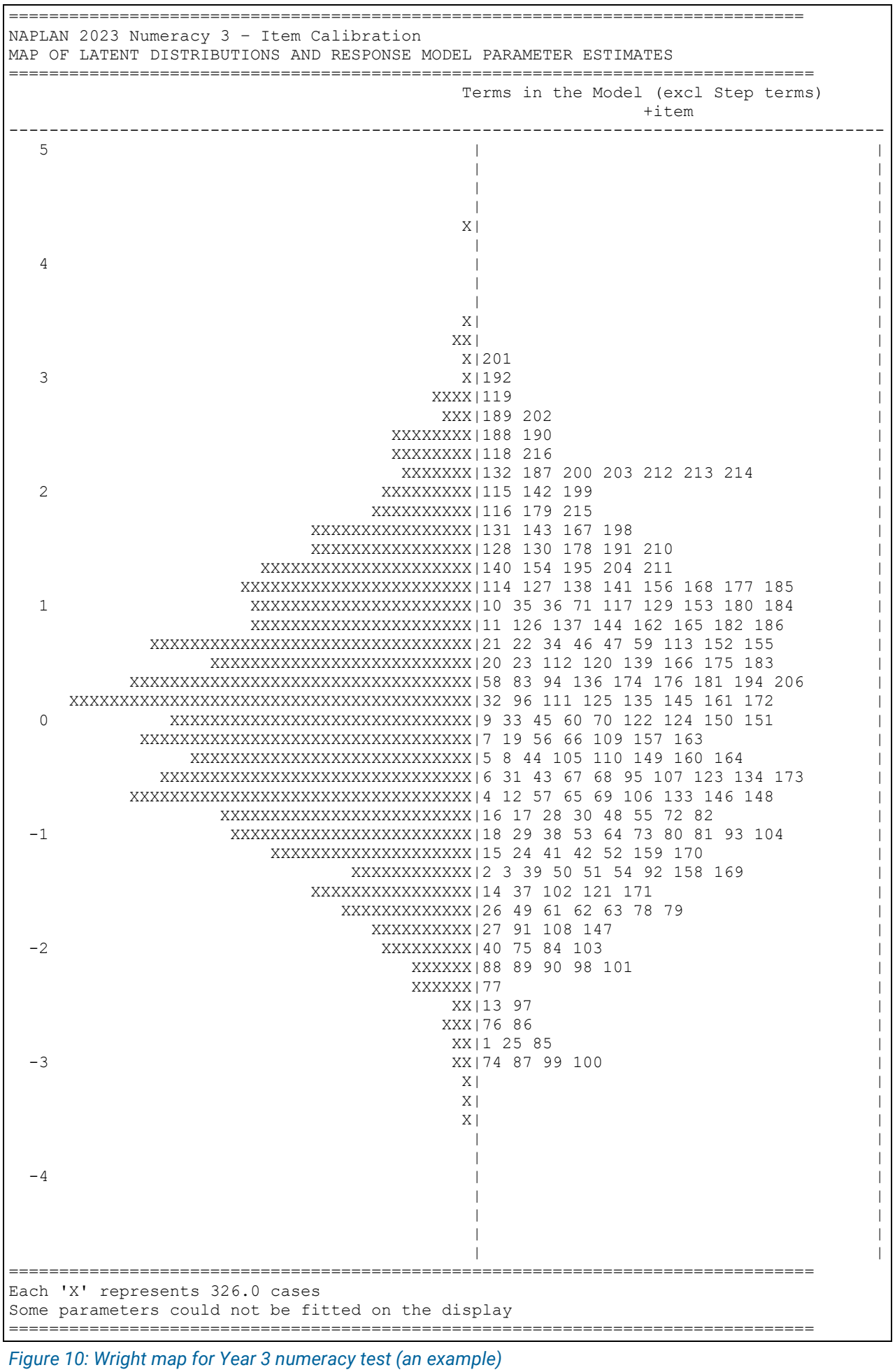


Figure 10: Wright map for Year 3 numeracy test (an example)

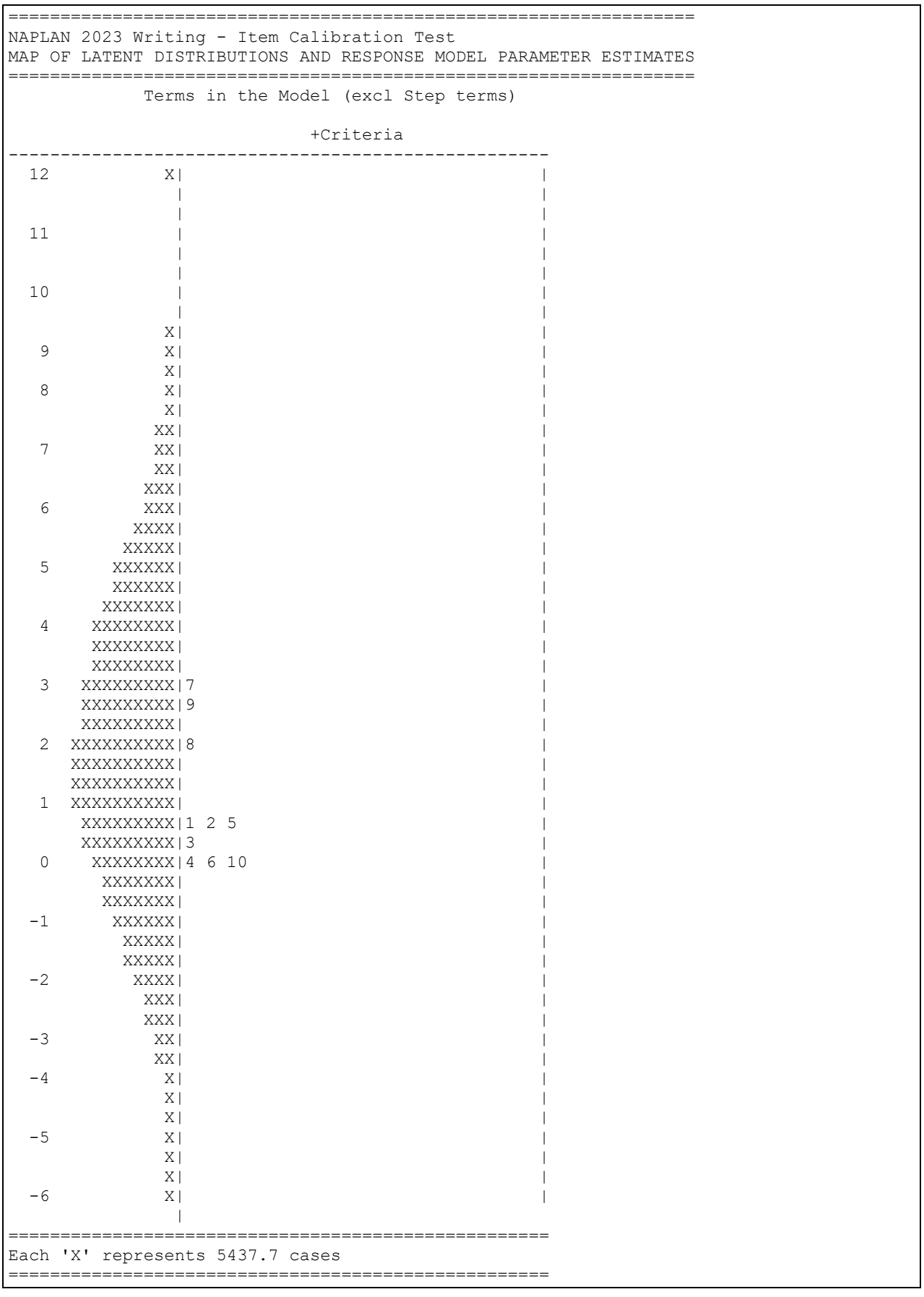


Figure 11: Wright map for writing test (a polytomous example)

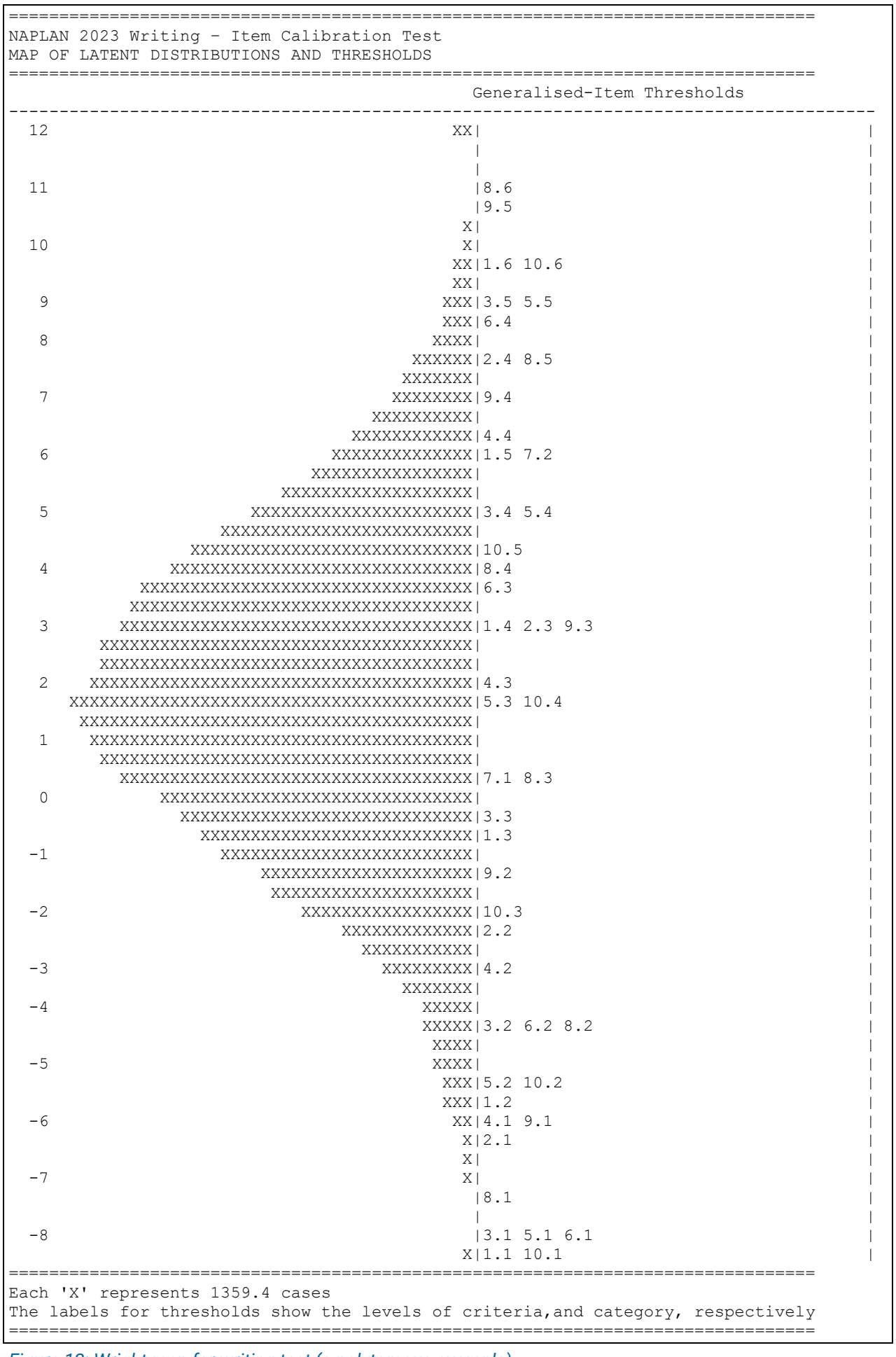


Figure 12: Wright map for writing test (a polytomous example)

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 to which 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). An infit value of 1.20 was used 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 an item that fails to discriminate). Classical item statistics such as item facility were also calculated. Values of the infit mean square and classical item statistics for each item can be found in Appendix B.

As mentioned above, the ICC of each item shows a comparison of the empirical ICC based on observations from ability groupings (broken line joining each dot) and the expected model-based ICC (smooth line). 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 year levels. As all items had previously been trialled and examined, few items were 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 49 presents summaries of item statistics in the NAPLAN 2023 tests. They present the number of items having infit mean square greater than 1.20. They also present the number of items with facility outside the range of 0.10 to 0.90, although it is acknowledged that these facility rates must be interpreted in the context of a branching test where items are seen by only a subset of the student population.

As seen from Table 49, 38 out of 3,232 items from 16 non-writing online tests had infit greater than 1.20. There were 88 items with facility higher than 0.90 and 31 items with facility less than 0.10. Figure 13 shows the ICC of one numeracy Year 3 item (item x00131075) with an infit statistic equal to 1.00. In contrast, Figure 14 shows the ICC of one Year 3 reading item (item x00077642) with an infit statistic (1.36) higher than the criterion value (1.20) for evaluating the goodness of fit of each item. The item parameter estimates and statistics are included in Appendix C for each of the 16 online tests calibration 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. Two criteria (paragraphing and punctuation) exhibited misfit to the Rasch partial credit model. Their infit values were 1.41 and 1.60 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 any of the 10 criteria, with small discrepancies visible for the criteria with the highest infit. The ICCs of the 10 writing criteria for writing are included in Appendix D.

Table 49: Summary of item statistics in NAPLAN 2023 tests

Domain	Year level	Total number of items	Number of items with infit > 1.20	Number of items with	
				Facility > 0.90	Facility < 0.10
Numeracy	3	216	2	3	0
	5	252	1	5	1
	7	288	7	9	0
	9	286	10	3	0
Reading	3	234	2	0	0
	5	234	1	9	0
	7	288	2	9	0
	9	288	1	8	1
Spelling	3	125	4	1	4
	5	126	2	6	7
	7	129	2	11	4
	9	125	2	10	9
Grammar and punctuation	3	160	1	2	1
	5	161	0	4	1
	7	160	0	5	3
	9	160	1	3	0
Writing	3	10*	2	n/a	n/a
	5				
	7				
	9				

* Item in writing is criterion.

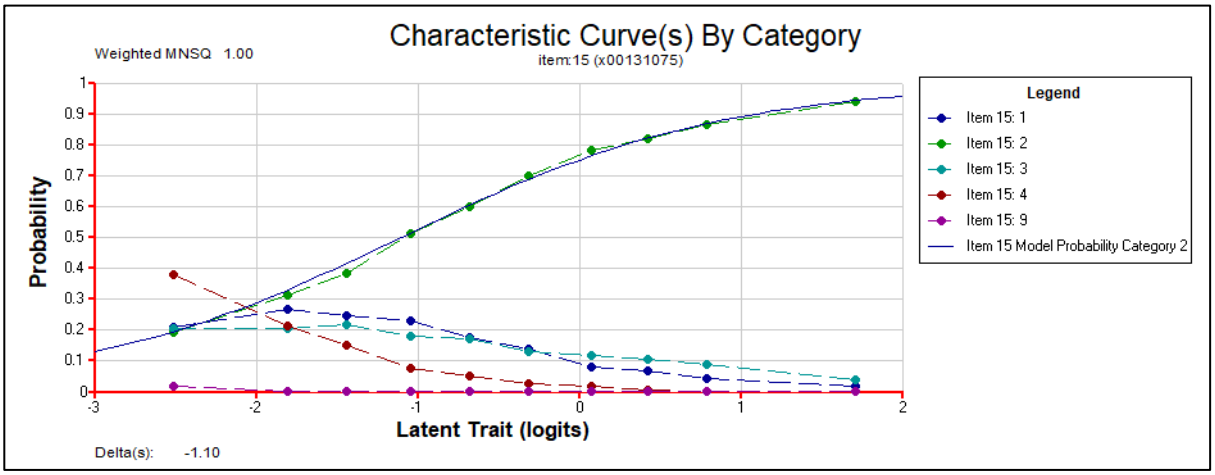


Figure 13: Item characteristic curves for an item with infit = 1.00

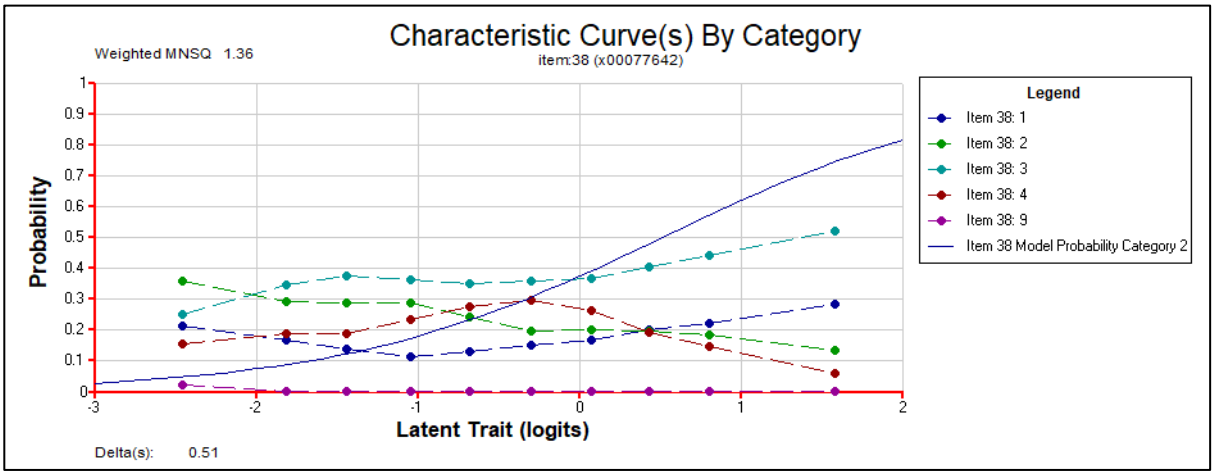


Figure 14: Item characteristic curves for an item with infit = 1.36

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 analyses by subgroup (gender[6], language background and Indigenous status), jurisdiction and device were performed for the NAPLAN 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*(θ), of a correct response differs for two groups. (Camilli and Shepard 1994)

[6] As per the *Data Standards Manual: Student Background Characteristics*, "gender" is considered a social and cultural concept. It is about social and cultural differences in identity, expression and experience as a male, female or non-binary person. Non-binary is an umbrella term describing gender identities that are not exclusively male or female. Due to the small number of individuals identifying by categories other than male and female, the analysis of gender DIF was limited to comparisons between males and females.

In the analysis for NAPLAN, subgroups were arbitrarily categorised as either reference or focal groups. While males, LBOTE students and Indigenous students were assigned to the reference group, females, non-LBOTE students and non-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 in order 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, or DIF.

For visual depiction of DIF, item locations of the reference group are plotted against those of the focal group as seen from appendices F, G and H (that is, gender, language background 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.

Gender DIF

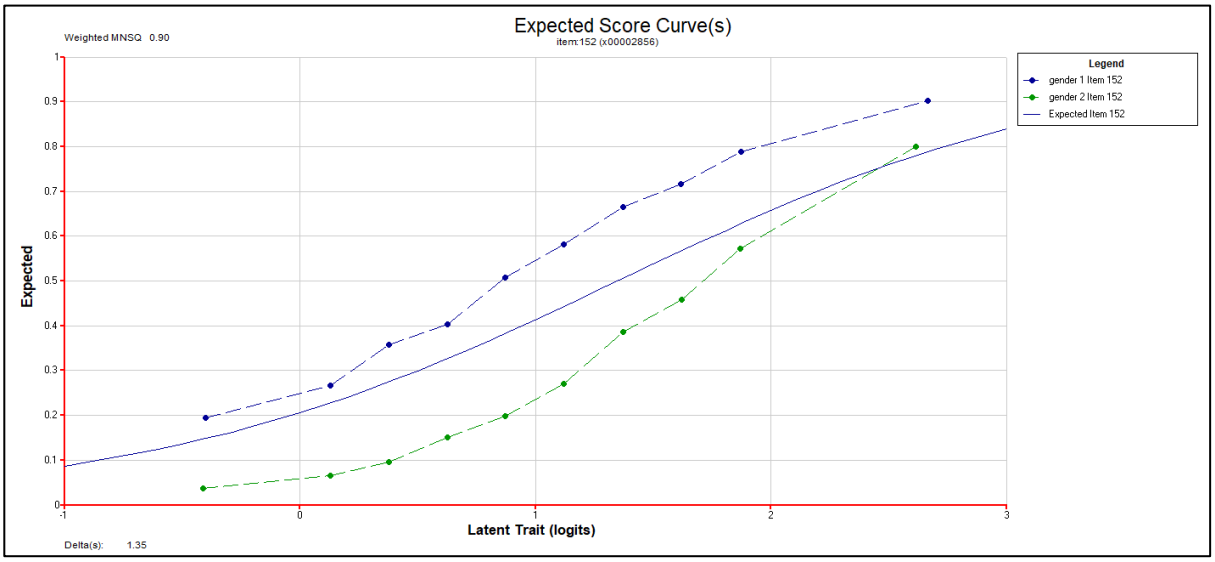
Appendix G presents the scatter plots for examining gender DIF in the 5 domains. The plots for numeracy, reading, spelling, and grammar and punctuation are presented by year levels. The writing gender DIF was performed by combining all 4 year levels together. Overall, the plots indicate that there are few items that exhibit gender differences in the adjusted item estimates, and that any differences are not large and thus are not of great concern.

Table 50 identifies the number of items (out of the total number of items) that show gender DIF with an absolute difference of 0.80 or greater for numeracy, reading, spelling, grammar and punctuation, and writing[7]. Figure 15 shows, as an example, one Year 5 numeracy item (Item x00002856) with an absolute difference of 0.80 or greater (in this case -1.29). A positive difference indicates that the reference group found the item harder than the focal group, and a negative difference means the opposite. In this case, there is a negative difference, meaning that that reference group (males) found the item easier than the focal group (females). Appendix G includes DIF plots that show for each of the items the observed curves by gender group compared with the expected ICC.

Table 50: Number of items showing gender DIF by domain by year level

Year level	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
3	3/216	0/234	0/125	0/160	0/10
5	2/252	0/234	1/126	0/161	
7	4/288	0/288	6/129	0/160	
9	3/286	1/288	12/125	0/160	

[7] For writing, 'item' refers to a marking criterion. This is applied throughout the report.



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

Figure 15: Example of item characteristic curves displaying gender DIF†

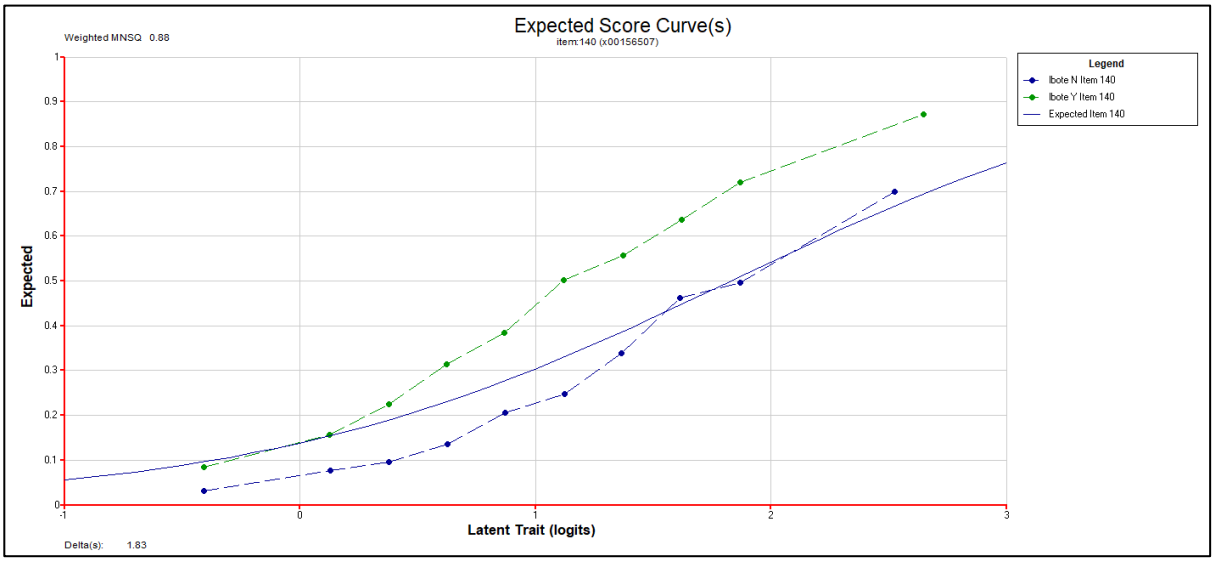
Language background DIF

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

Table 51 indicates the number of items that show DIF with an absolute adjusted difference of 0.80 or greater for numeracy, reading, spelling, grammar and punctuation, and writing. Figure 16 depicts one Year 5 numeracy online test item (item x00156507) with an absolute mean difference of 0.80 or greater. This item was relatively easy (mean difference = -1.00) for LBOTE students.

Table 51: Number of items showing LBOTE DIF by domain by year level

Year level	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
3	0/216	0/234	1/125	1/160	0/10
5	2/252	0/234	1/126	0/161	
7	0/288	0/288	0/129	1/160	
9	1/286	0/288	0/125	3/160	



† "Ibote Y" indicates "LBOTE group" and "Ibote N" indicates "non-LBOTE group".

Figure 16: Example of item characteristic curves displaying language background DIF †

Indigenous status DIF

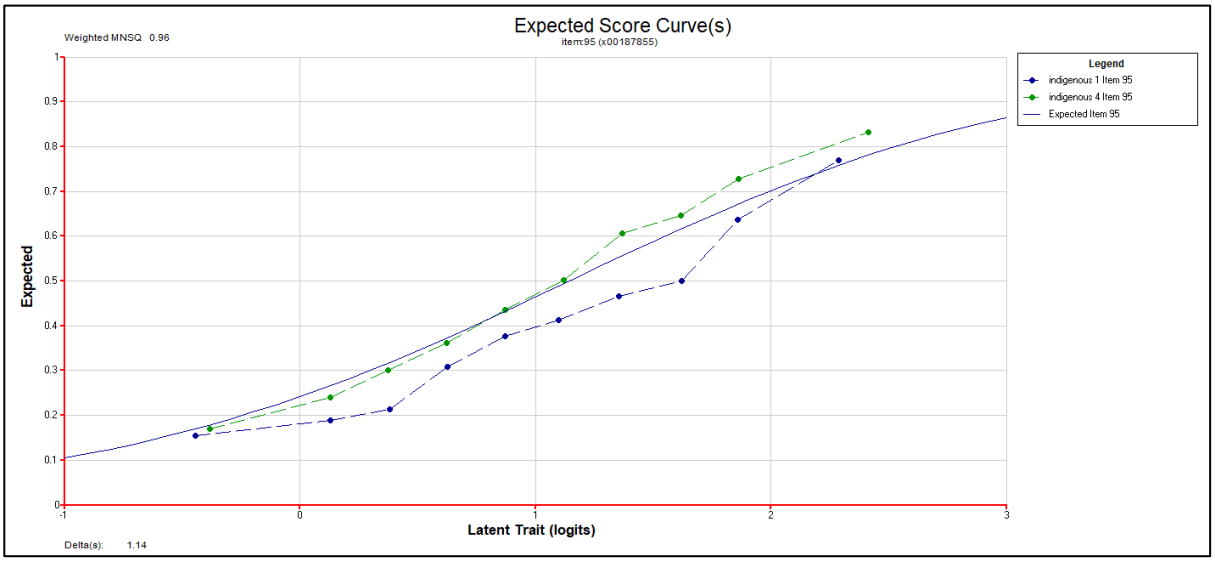
Appendix I includes scatter plots for examining Indigenous DIF in the 5 domains for both paper and online tests. Writing Indigenous DIF was performed by combining all 4 grades. These plots showed that there were not many items that showed notable differences in the relative item difficulties for tests.

Table 52 lists the number of items that show Indigenous DIF with an absolute adjusted difference of 0.80 or greater for numeracy, reading, spelling, grammar and punctuation, and writing. Figure 17 depicts one Year 9 numeracy online test item (item x00175090) with an absolute mean difference of 0.80 or greater. This item was relatively easy (mean difference = 0.87) for non-Indigenous students.

Appendix I provides the item DIF plots for items listed in Table 52. 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 52: Number of items showing Indigenous DIF by domain by year level

Year level	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
3	0/216	0/234	0/125	0/160	0/10
5	0/252	0/234	0/126	1/161	
7	2/288	0/288	1/129	1/160	
9	3/286	0/288	0/125	0/160	



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

Figure 17: Example of item characteristic curves displaying Indigenous status DIF†

DIF values of individual items for gender, language background and Indigenous status, as well as for jurisdiction and device are presented in Appendix J.

Jurisdictional DIF

To determine whether jurisdictional 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 of the calibration sample. 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:

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

If the difference for an item between a state/territory and the national average was greater than 0.40 logit, then the item was deemed harder for the state/territory. If the difference was less than -0.40 logit, then the item was deemed easier for the state/territory.

The number of items showing jurisdictional DIF in numeracy, reading, spelling, grammar and punctuation, and writing is shown in Table 53. In the headings of Table 53, "E" indicates that the item is relatively easy for the jurisdiction, and "H" indicates that the item is relatively hard for the jurisdiction. Note that, due to the smaller sample size, more items are shown as displaying DIF for smaller jurisdictions. Table 53 can be read in conjunction with Appendix K, which contains item DIF plots for items showing jurisdictional DIF for items listed in Table 53. The plots show, for each of these items, the observed curves by state/territory compared with the expected ICC. Figure 18 depicts one Year 3 numeracy test item (item x00075150) showing jurisdictional DIF. This item was relatively easy for Qld students.

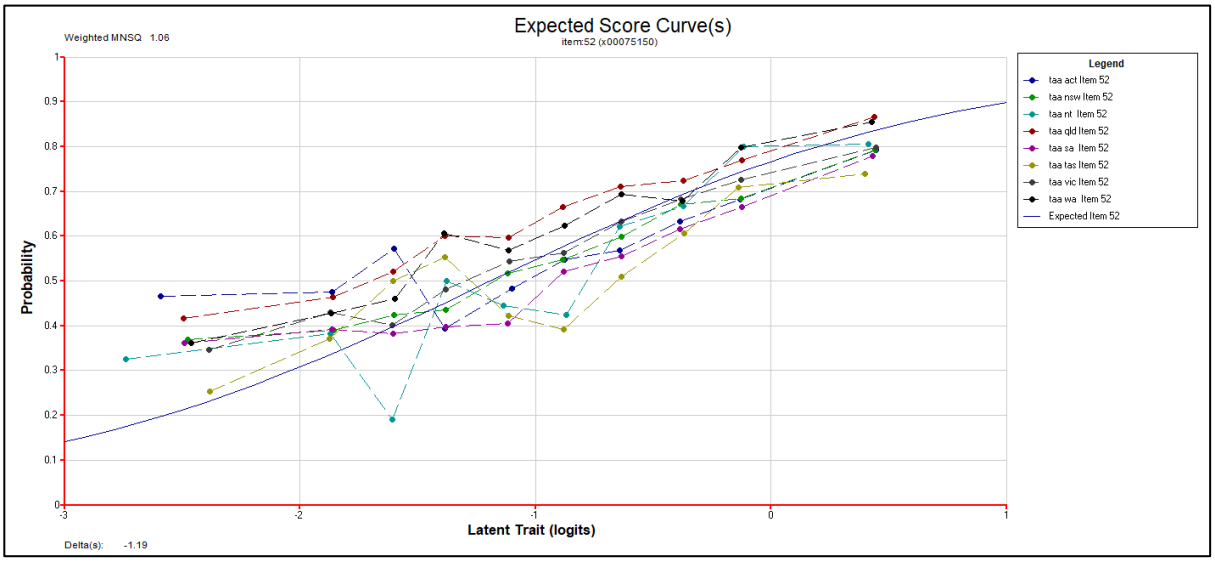


Figure 18: Example of item characteristic curves displaying jurisdictional DIF

Table 53: Number of items showing jurisdictional DIF by domain by year level

Domain	Year level	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	-	-	-	9	5	1	-	-	1	-	4	-	-	-	-
	5	1	-	1	1	13	8	-	-	1	-	2	5	1	-	-	-
	7	3	1	2	1	21	19	-	-	1	-	1	7	3	-	1	1
	9	1	-	1	1	14	9	-	-	-	1	5	5	-	-	3	1
Reading	3	3	1	-	-	5	6	-	-	-	-	1	1	-	-	-	-
	5	2	3	-	-	5	7	-	-	-	-	-	-	1	-	-	-
	7	1	-	-	-	5	17	-	-	-	-	-	1	1	-	-	-
	9	1	-	-	-	3	8	-	-	-	-	4	1	1	-	6	-
Spelling	3	1	1	-	-	1	6	-	-	-	1	-	1	-	1	2	-
	5	1	-	-	-	4	9	-	-	-	1	3	2	1	-	-	1
	7	1	1	-	-	4	2	-	1	-	-	4	2	-	-	1	-
	9	3	4	1	-	2	8	-	-	1	-	3	2	1	-	1	-
Grammar and punctuation	3	2	1	-	-	8	3	-	-	-	-	1	2	-	-	-	-
	5	8	2	1	-	4	6	2	1	-	-	3	1	3	-	-	-
	7	4	-	-	-	6	8	2	1	2	-	3	2	1	-	-	-
	9	-	2	-	-	3	5	3	1	-	-	-	2	-	-	1	-
Writing	3 5 7 9	-	1	-	-	-	3	2	-	-	-	1	2	1	2	-	-

Note. “E” indicates that the item was relatively easy for the jurisdiction, and “H” indicates that the item was relatively hard for the jurisdiction.

Device DIF

For online tests, a device DIF analysis was also carried out for non-writing domains[8] as there were different devices used by different students. There were 4 different types of device used: Chromebook, iOS, Mac and Windows. The same method used to determine jurisdictional DIF was used for determining device DIF. Table 54 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 the national item parameters. An item parameter demonstrating an absolute value of the difference greater than 0.40 logits was deemed as exhibiting DIF. A summary of device DIF is shown in Table 55. Table 55 shows that Mac devices had the most items demonstrating DIF, especially in numeracy, reading, and grammar and punctuation. Appendix L includes scatter plots for examining device DIF in the 4 non-writing domains.

Table 54: Number of students by device

Domain	Year level	Chromebook	iOS	Mac	Windows
Numeracy	3	37,072	73,320	2,183	88,178
	5	36,867	51,696	5,770	97,949
	7	19,960	17,578	31,697	135,714
	9	16,762	13,944	35,392	125,485
Reading	3	40,317	79,562	2,591	100,189
	5	43,650	61,480	7,188	123,563
	7	21,721	18,508	33,544	158,251
	9	18,850	14,805	38,328	141,105
Spelling	3	36,669	68,925	2,119	83,776
	5	38,979	53,293	5,907	99,837
	7	19,740	16,732	30,382	135,592
	9	15,827	12,371	33,701	117,903
Grammar and punctuation	3	35,617	68,077	2,046	81,220
	5	38,868	53,341	5,905	99,734
	7	19,914	16,902	30,514	136,707
	9	16,126	12,697	34,260	120,478

[8] Device DIF was not investigated for writing as all Year 3 students completed the test on paper and some students in Year 5, 7 and 9 completed the test on paper while others completed the test online.

Table 55: Number of items showing device DIF by domain by year level

Domain	Year level	Chromebook		iOS		Mac		Windows	
		E	H	E	H	E	H	E	H
Numeracy	3	-	-	-	-	1	1	-	-
	5	-	-	-	-	-	-	-	-
	7	-	-	-	-	4	-	-	-
	9	-	-	-	-	4	-	-	-
Reading	3	-	-	-	-	-	-	-	-
	5	-	-	-	-	1	-	-	-
	7	-	-	-	-	1	-	-	-
	9	-	-	-	-	3	-	-	-
Spelling	3	-	-	-	-	-	3	-	-
	5	1	-	-	-	-	-	-	-
	7	-	-	-	-	-	-	-	-
	9	-	-	-	-	-	-	-	-
Grammar and punctuation	3	-	-	-	-	2	-	-	-
	5	-	-	-	-	2	1	-	-
	7	-	-	-	-	3	-	-	-
	9	-	-	-	-	-	-	-	-

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 (testlet) 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 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 pathway of the online tests by domain by year level based on delta-centred item parameters. 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 to convert them to the new NAPLAN reporting scales.

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*. They were based on imputation techniques and a multidimensional item response model (partial credit model) with latent regression

(Adams et al. 2020) 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 tested students together. 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[9]. A diagrammatic representation of the multidimensional model is shown in Figure 19.

The categorical conditioning variables (gender, LBOTE status, Indigenous status, parental education, parental occupation, interaction dummy variables of school sector by school geolocation, 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 Indigenous status was designated as having 3 categories, namely, *non-Indigenous*, *Indigenous* and *not stated/unknown*. The categories of Indigenous status were recoded for each student using one indicator variable to denote *Indigenous*, and a second indicator variable to denote *not stated/unknown*. If the pair of indicator variables had the values 1 and 0 respectively, this meant that the Indigenous status category for the student was *Indigenous*; when the indicator variables had the values of 0 and 1, then the Indigenous status category was *not stated/unknown*. When both indicators were 0, this indicated that the Indigenous status category for the student was *non-Indigenous*. 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 student. Test mode was included in the conditioning model for all jurisdictions and year levels where there were sufficient paper tested students.

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 and year level for all students (including absent students and withdrawn students) except for students who were exempt from NAPLAN testing.

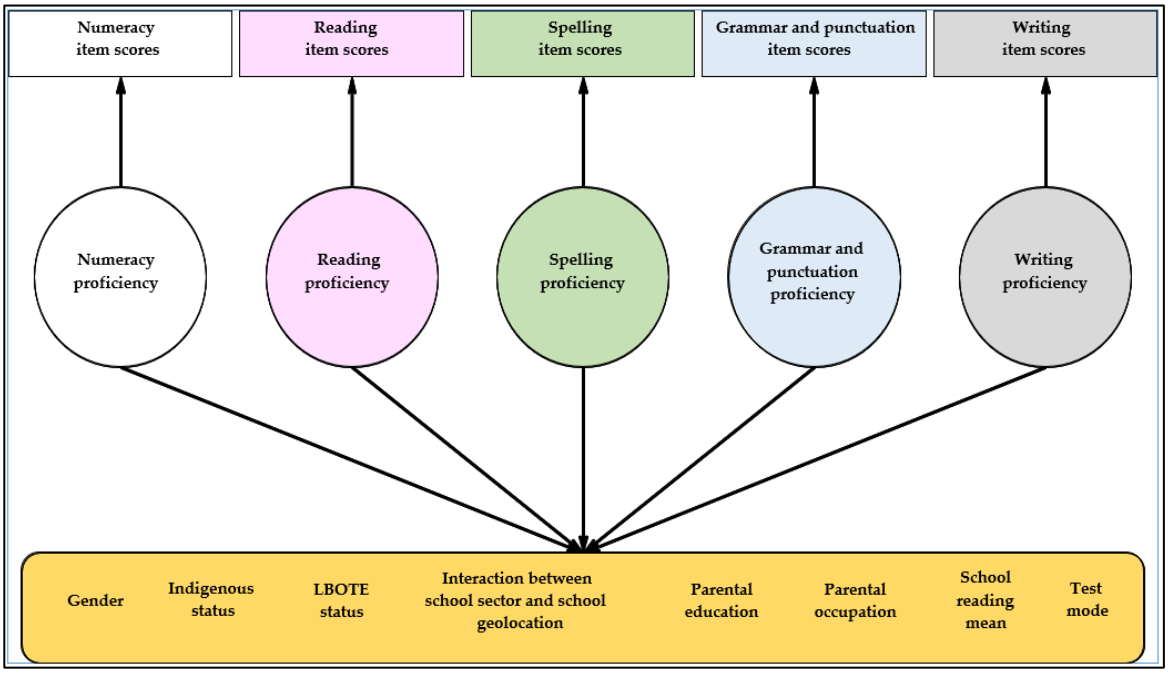


Figure 19: Conditioning variables for the multidimensional item response model with latent regression

[9] the inclusion of test mode as a regressor varied by jurisdiction.

Chapter 6: Equating procedures

In 2023, the NAPLAN scale was reset, and the time series discontinued, because of the full transition to the adaptive online assessment and the change in testing window from May to March.

This chapter describes the process of equating the 2023 tests to construct new NAPLAN scales. The first section describes the vertical equating of each year level test on to common scales for each of the 2023 numeracy, reading, spelling, and grammar and punctuation domains, followed by a description of the equating procedures for writing, for which a different equating design and methodology was applied. The chapter finishes with a summary of equating parameters.

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. Each of the reading, spelling, grammar and punctuation, and numeracy scales across Years 3, 5, 7 and 9 was constructed through test equating processes using link items embedded in adjacent tests (common items in Year 3 and Year 5 tests; in Year 5 and Year 7 tests; and in Year 7 and Year 9 tests). The vertical equating design for the 2023 tests is represented schematically in the data matrix in Table 56.

Table 56: Equating design

	NAPLAN test items – vertical links						
Students	**Y3**	**Y3&5**	**Y5**	**Y5&7**	**Y7**	**Y7&9**	**Y9**
Y3 population							
Y5 population							
Y7 population							
Y9 population							

For each of the 4 domains, before calculating the vertical equating shifts, the quality of the 3 sets of link items (Y3/Y5, Y5/Y7, Y7/Y9) was systematically reviewed. Only items that showed satisfactory and similar psychometric properties across adjacent 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 to the Rasch model and evidence of differential item functioning (DIF) between test forms. Review of the vertical 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 relative item difficulties (for example, of 2023 Year 3 link items) was then plotted against the other set of relative item difficulties (for example, of 2023 Year 5 link items). Two plots are 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 reviewing and selecting good link items. On the plots, each dot represents a common item. Links were broken in 3 steps:

1. Any items that were more than 10 positions apart between test forms were broken first[10].
2. Outliers (items with an absolute difference larger than 0.9 of a logit between their relative difficulties) were then broken, and the process was repeated if necessary.
3. Any other items with an absolute difference of more than 0.5 logits between their relative difficulties were broken in the third step, and the process was repeated if necessary.

- For each set of linked test scales, the mean difficulty of the remaining link items was calculated for each of the 2 test forms. The equating shift is the difference between the 2 means.
- In addition to relative item difficulties of the link items, (average) position of the item in the pathway, infit mean square 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 test form (for example, 2023 Year 3 SD / 2023 Year 5 SD). The ideal ratios between equated tests should fall between 0.85 and 1.15. The actual ratios of retained link items for all vertical equating were between 0.93 and 1.09.

The outcome of the review of vertical link items is summarised in Figure 20 to Figure 31. These plots show the comparisons of item difficulty estimates between 2 adjacent test forms for each of the 4 domains. 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.

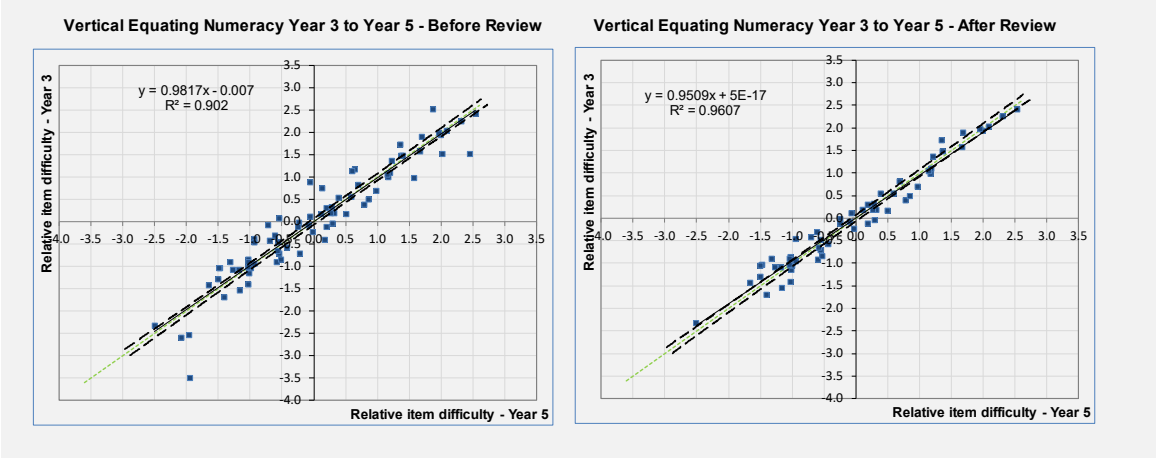


Figure 20: Scatterplot of numeracy, vertical link items between Year 3 and Year 5 online tests

[10] Even before this, items common to the calculator-allowed section of the Year 7 Numeracy test and the (non-calculator) Year 5 test were removed from consideration whenever they included a calculation demand, due to possible differences in the performance of these items. Each of the subsequent steps was then applied.

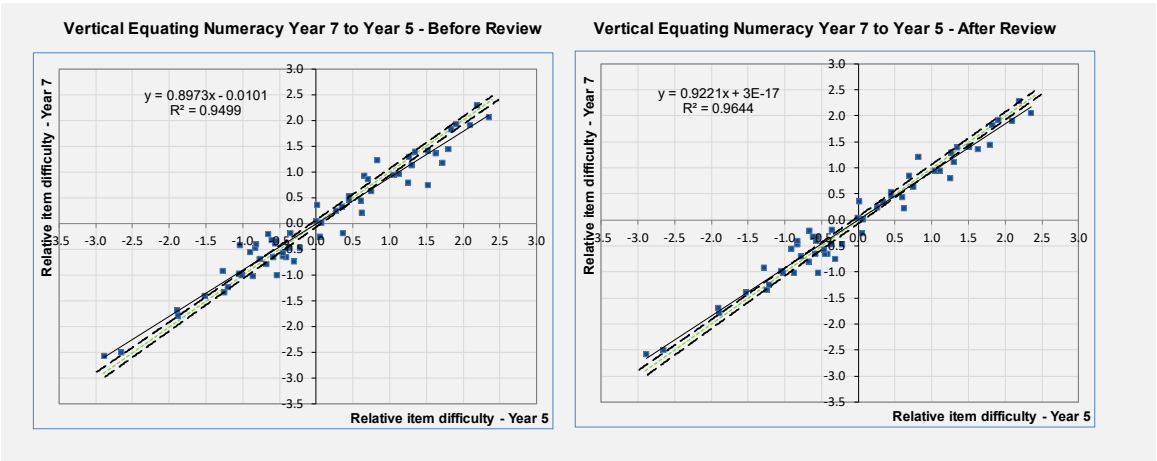


Figure 21: Scatterplot of numeracy, vertical link items between Year 7 and Year 5 online tests

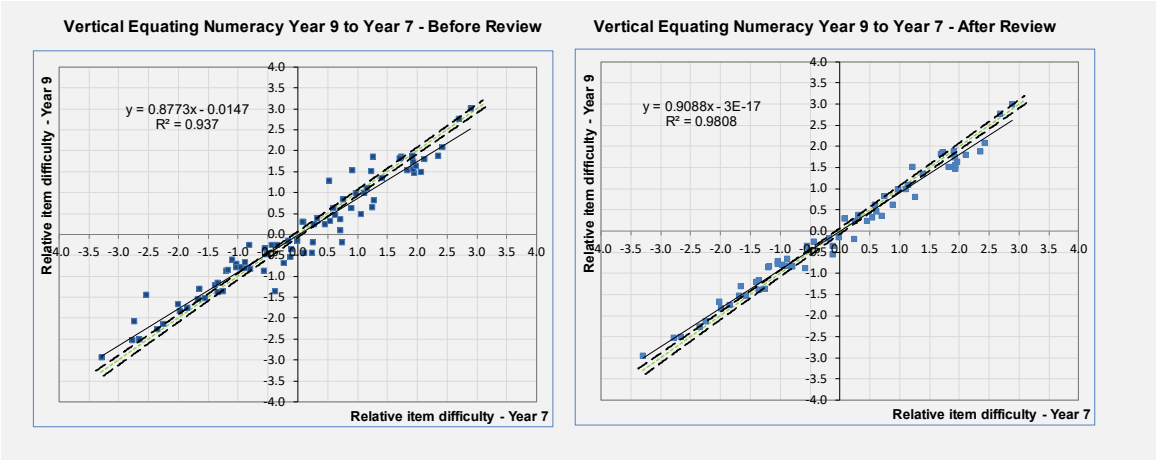


Figure 22: Scatterplot of numeracy, vertical link items between Year 9 and Year 7 online tests

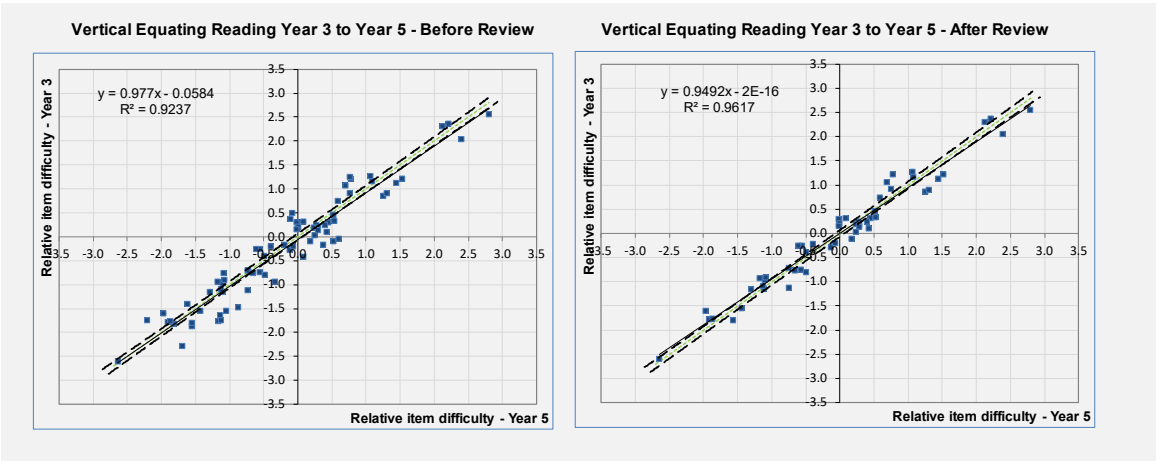


Figure 23: Scatterplot of reading, vertical link items between Year 3 and Year 5 online tests

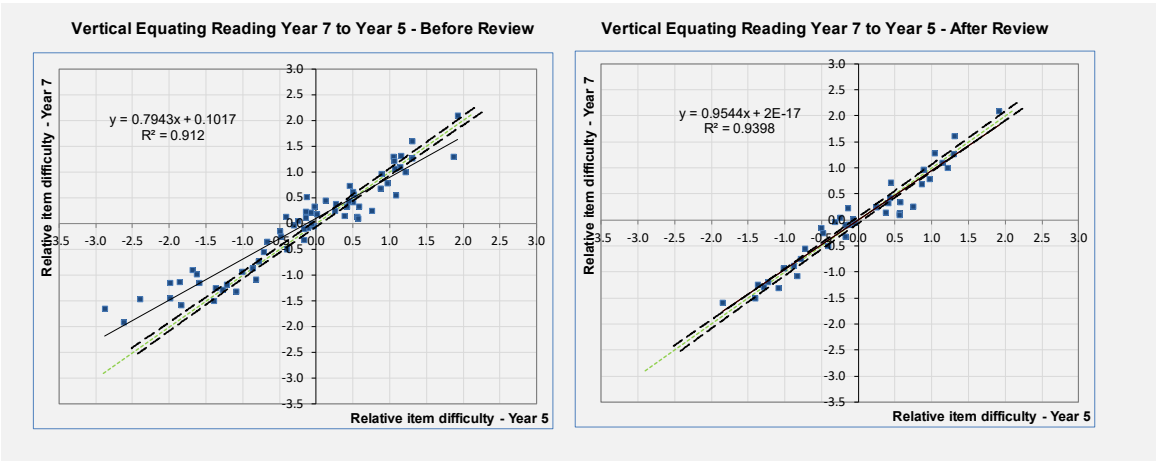


Figure 24: Scatterplot of reading, vertical link items between Year 7 and Year 5 online tests

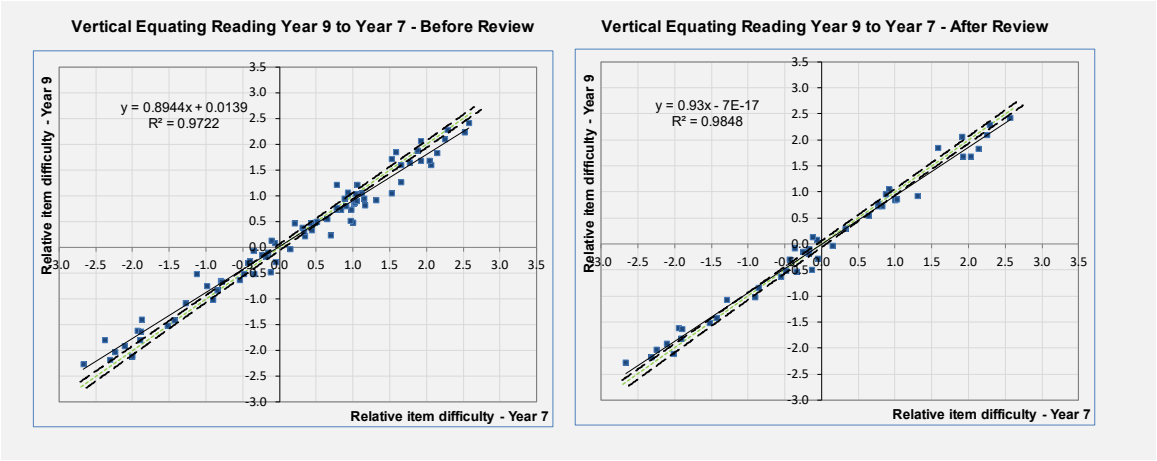


Figure 25: Scatterplot of reading, vertical link items between Year 9 and Year 7 online tests

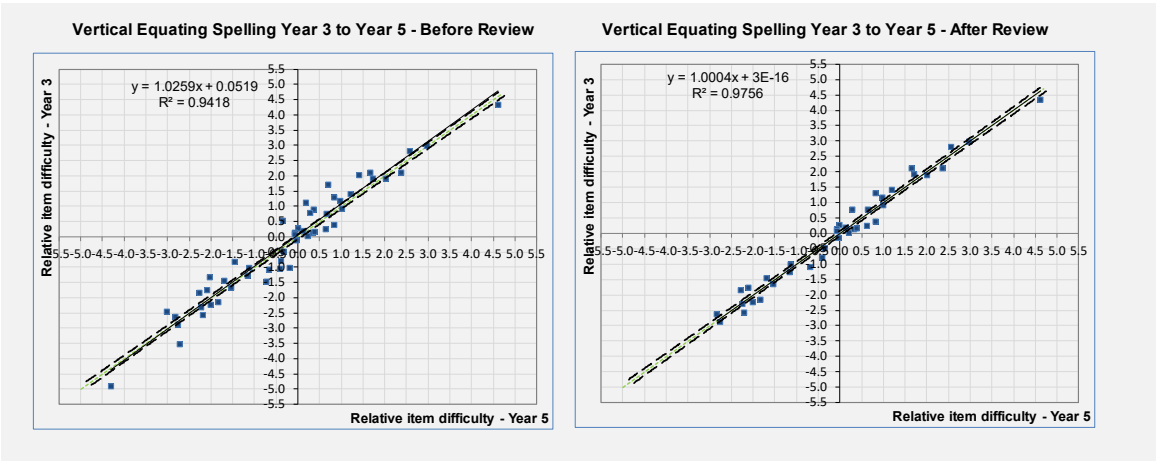


Figure 26: Scatterplot of spelling, vertical link items between Year 3 and Year 5 online tests

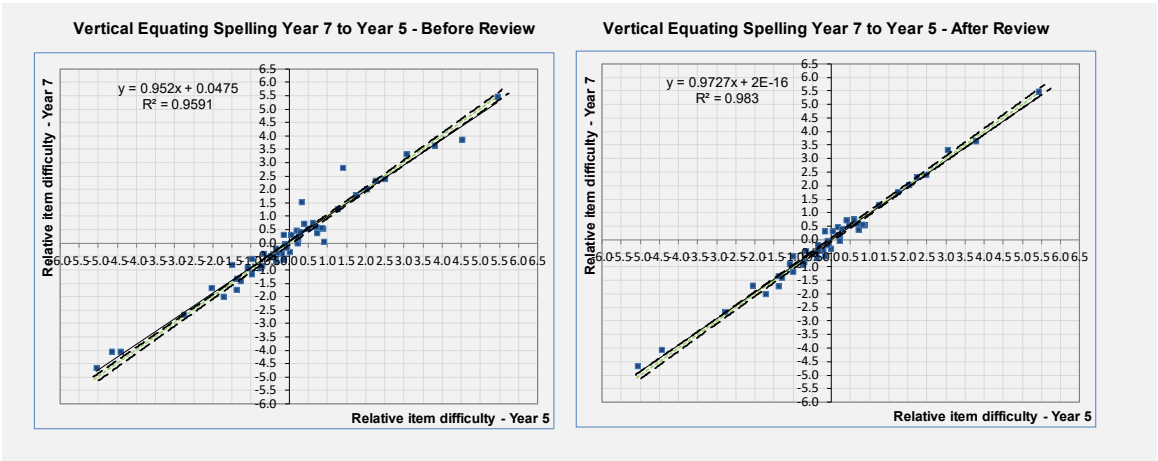


Figure 27: Scatterplot of spelling, vertical link items between Year 7 and Year 5 online tests

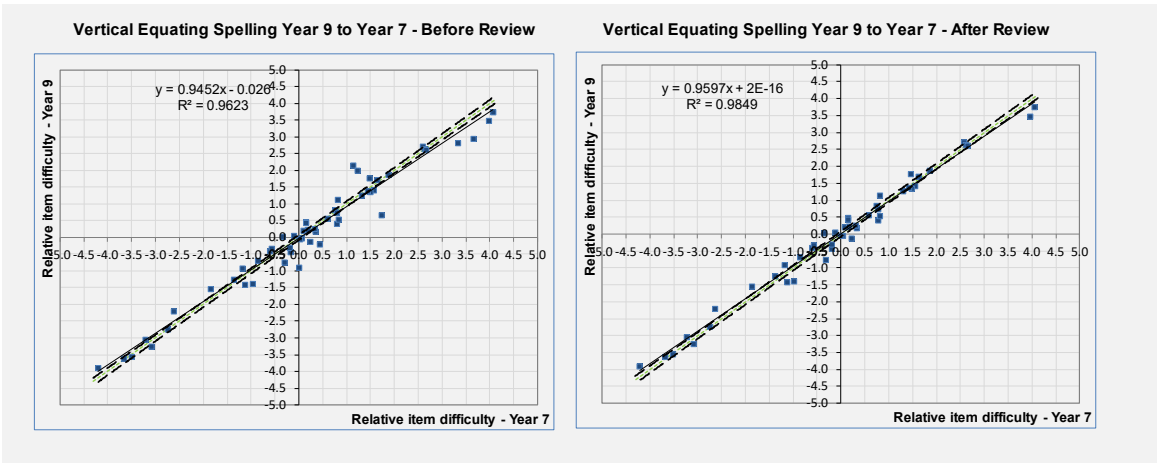


Figure 28: Scatterplot of spelling, vertical link items between Year 9 and Year 7 online tests

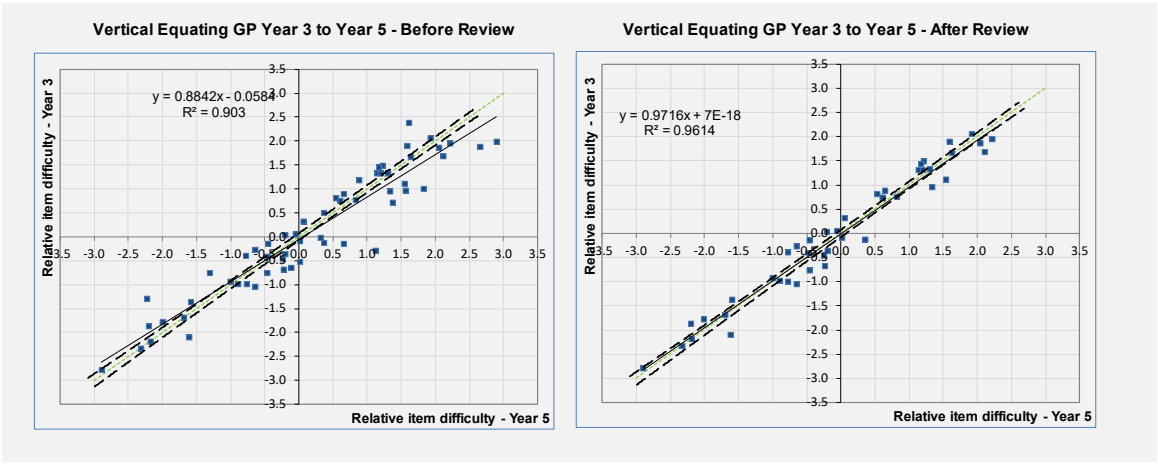


Figure 29: Scatterplot of grammar and punctuation, vertical link items between Year 3 and Year 5 online tests

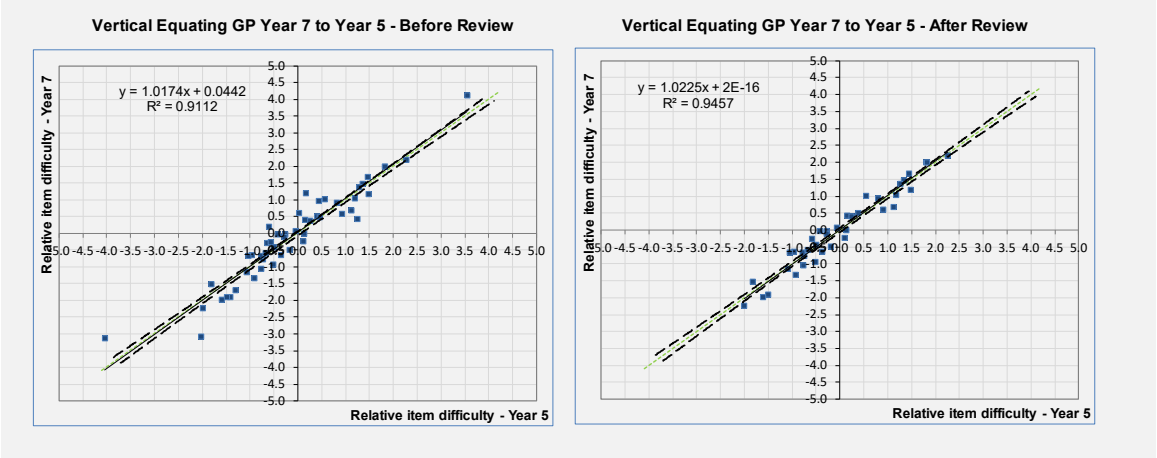


Figure 30: Scatterplot of grammar and punctuation, vertical link items between Year 7 and Year 5 online tests

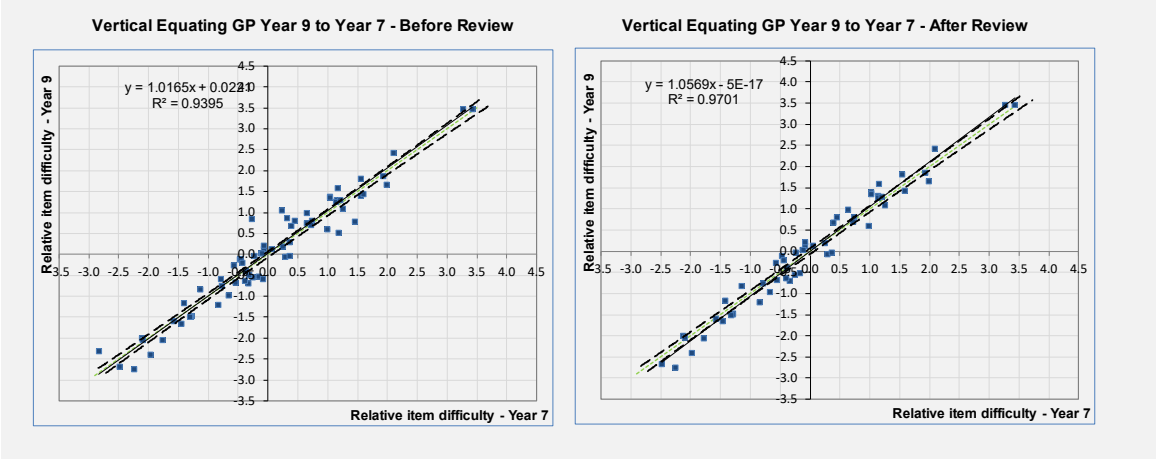


Figure 31: Scatterplot of grammar and punctuation, vertical link items between Year 9 and Year 7 online tests

The final sets of link items were used to calculate the vertical shifts between year levels. With the Year 5 mean item difficulties centred at zero logits, the shift constants were applied to the Year 3, Year 7 and Year 9 item difficulties to construct a vertical scale for the domain. This vertical scale by domain is the reset 2023 NAPLAN scale. The outcome of the review of vertical link items is summarised in Table 57 and Table 58. Details of the parameters used in the calculations can be found in Appendix M.

Appendix M presents the 2023 vertical link item locations (Rasch difficulties), standard errors, and differences in the item locations by domain and year level.

Table 57: Vertical link review summary for online tests (Number of used links/Number of common items)

Adjacent year levels	Numeracy	Reading	Spelling	Grammar and punctuation
3/5	58/75	52/72	38/51	42/57
5/7	56/61	40/66	46/52	38/50
7/9	62/80	46/80	47/54	54/67

Table 58: Vertical equating shifts between adjacent year level item locations and their associated equating errors by domain

Adjacent year levels	Numeracy		Reading		Spelling		Grammar and punctuation	
	Shift	Error	Shift	Error	Shift	Error	Shift	Error
3/5	-1.33551	0.0304	-1.01874	0.0319	-2.12529	0.0441	-0.79962	0.0413
5/7	0.65079	0.0313	0.60641	0.0351	1.32911	0.0349	0.73657	0.0435
7/9	0.70227	0.0305	0.44173	0.0281	0.69276	0.0343	0.12775	0.0354

Table 59: Final vertical equating shifts applied for each test by year level by domain

Year level	Numeracy	Reading	Spelling	Grammar and punctuation
3	-1.33551	-1.01874	-2.12529	-0.79962
5	0	0	0	0
7	0.65079	0.60641	1.32911	0.73657
9	1.35306	1.04814	2.02187	0.86432

The pertinent shifts in logits listed in Table 59 were applied to the item difficulties of each year level test to equate the tests onto the Year 5 scale for each domain. This resulted in vertical scales in reading, spelling, grammar and punctuation, and numeracy that spanned across achievement levels in Years 3, 5, 7 and 9. The vertical logit scales were ready to be transformed into the reset 2023 NAPLAN reporting scales.

Equating of writing results

As described in Chapter 6, the writing data from all 4 year levels was concurrently calibrated to construct the vertical writing scale. Because this process placed the 4 year levels on the same scale, there was no separate equating process required.

Standardisation of scales from logits to reporting scales

For each domain, estimates in logits were transformed to the NAPLAN reporting scale scores. To establish scale transformation equations, the overall preliminary mean and standard deviation across the 4 year levels were calculated for each domain based on plausible values drawn from the stage 1 census data. Stage 1 data contains data for all domains and is available at the end of the marking operations for writing and for paper scripts. The estimated mean and standard deviations in logits are shown in Table 60. These were used to standardise each domain scale to have an overall mean of 500 and standard deviation of 100 as follows:

$$Score_{NAPLANScale} = 100 \cdot \frac{Score_{logit} - DomainMean_{2023}}{DomainStdDeviation_{2023}} + 500 \quad (4)$$

where $DomainMean_{2023}$ and $DomainStdDeviation_{2023}$ were the estimated overall domain mean and domain standard deviation calculated using the 2023 stage 1 data.

It should be noted that for each domain, the standard error (SE) in logits associated with each individual student WLE estimate was transformed to the NAPLAN scale metric as follows:

$$SE_{NAPLANScale} = 100 \cdot \frac{SE_{logit}}{DomainStdDeviation_{2023}} \quad (5)$$

Table 60: Domain mean and standard deviation for transforming logits to NAPLAN scale scores

Domain	Domain mean overall	Domain SD overall
Reading	0.21845	1.41868
Writing	0.62741	3.16266
Spelling	0.24156	2.77813
Grammar and punctuation	0.26014	1.29412
Numeracy	0.24273	1.67176

Summary of equating parameter estimates for NAPLAN 2023

In 2023, the NAPLAN scales for each domain were reset using a vertical equating methodology to place all 4 year levels together on the same scale. For each domain, a student's ability estimate on the delta-centred scale for their year level was first shifted to the delta-centred Year 5 logit scale and then transformed to a NAPLAN scale score as below:

$$\theta^{x}_{2023onYear5} = \theta^{xy}_{2023} + VerticalShift^{x}_{yto5} \quad (6)$$

$$Score^{x}_{NAPLANScale} = \frac{(\theta^{x}_{2023onYear5} - Mean^{x}_{2023})}{StdDeviation^{x}_{2023}} * 100 + 500 \qquad (7)$$

where θ^{xy}_{2023} is the 2023 achievement score in logits on the Year *y* delta-centred scale for domain *x*, $\theta^{x}_{2023onYear5}$ is the vertically equated 2023 achievement score in logits on the Year 5 delta-centred scale for domain *x*, and $VerticalShift^{x}_{yto5}$ is the vertical shift from Year y to Year 5 for domain x, listed in Table 59. For writing, $VerticalShift^{x}_{yto5}$ equals zero. $Score^{x}_{NAPLANScale}$ is the scale score for domain x on the new NAPLAN scale, $Mean^{x}_{2023}$ is the average achievement score across all 4 year levels in logits for domain x, and $StdDeviation^{x}_{2023}$ is the standard *deviation* across all 4 year levels for domain x, listed in Table 60.

Together, the shifts and transformations are collated and shown in Table 61.

Table 61: Summary of parameters for transforming the 2023 logit scores to the NAPLAN reporting scales

Domain	Year level	Vertical shift	Mean	Standard deviation
Reading	3	-1.01874	0.21845	1.41868
	5	0	0.21845	1.41868
	7	0.60641	0.21845	1.41868
	9	1.04814	0.21845	1.41868
Writing	3	0	0.62741	3.16266
	5	0	0.62741	3.16266
	7	0	0.62741	3.16266
	9	0	0.62741	3.16266
Spelling	3	-2.12529	0.24156	2.77813
	5	0	0.24156	2.77813
	7	1.32911	0.24156	2.77813
	9	2.02187	0.24156	2.77813
Grammar and punctuation	3	-0.79962	0.26014	1.29412
	5	0	0.26014	1.29412
	7	0.73657	0.26014	1.29412
	9	0.86432	0.26014	1.29412
Numeracy	3	-1.33551	0.24273	1.67176
	5	0	0.24273	1.67176
	7	0.65079	0.24273	1.67176
	9	1.35306	0.24273	1.67176

Chapter 7: Proficiency levels

In 2023, proficiency levels were introduced for NAPLAN. These replaced the numerical achievement bands and national minimum standard that were in place until 2022.

Four levels of proficiency were defined for each domain and year level:

- Exceeding
- Strong
- Developing
- Needs additional support.

Standard setting

The lower boundaries for the Exceeding and Strong levels were set in 2022 by panels of experienced and expert teachers, along with curriculum and assessment specialists from states and territories. The panels had expertise working in a wide variety of educational settings, and were able to bring diverse perspectives to the panels.

Panels were convened in each year level for each of the 5 NAPLAN domains: numeracy, reading, writing, spelling, and grammar and punctuation. This made 20 panels in all, with 11 participants per panel on average. Some panellists participated in more than one panel.

Panellists in domains other than writing set cut-points for these levels by analysing a set of NAPLAN items of known difficulty, then judging whether a student who had only just achieved that level of proficiency would be able to answer the item correctly.

Writing panels followed a parallel process, analysing a set of previously scored NAPLAN writing responses, then judging whether a student who had only just achieved that level of proficiency would be able to produce that piece of writing.

Panels convened twice: once to work through the process, after which they independently made their judgements of the items or responses; then again to discuss borderline decisions and arrive at consensus judgements.

Panels worked to the following policy descriptors for the levels:

- Strong: The student's result meets challenging but reasonable expectations at the time of testing.
- Exceeding: The student's result exceeds expectations at the time of testing.

The lower boundaries set by the panels for these levels are shown in Table 62. The scores shown here are on the scales that were in place until 2022, since the scales had not been reset at the time of the panels.

Table 62: Panel proficiency judgements on historical NAPLAN scale

Domain	Year level	Developing/Strong	Strong/Exceeding
Numeracy	3	364	479
	5	452	565
	7	511	618
	9	554	669
Reading	3	379	503
	5	464	560
	7	505	603
	9	548	640
Writing	3	365	493
	5	462	572
	7	500	584
	9	541	638
Spelling	3	390	505
	5	459	553
	7	498	599
	9	551	640
Grammar and punctuation	3	397	518
	5	475	593
	7	513	627
	9	553	639

Transformation to new scale

Once the scales were reset after the 2023 NAPLAN tests, the boundaries between levels were translated from the historical NAPLAN scale to the reset NAPLAN scale.

This translation was achieved in non-writing domains by drawing a set of plausible values, then transforming them on to both old (up till 2022) and new (2023) measurement scales. In writing, 2 sets of plausible values were drawn, one using the 2022 item and step parameters, and one using the 2023 parameters.

It should be noted that the translations are by their nature imprecise, in that the new scales capture the distribution of achievement revealed by the adaptive tests in a way that the old scales cannot.

The transformed boundaries are shown in Table 63.

Table 63: Panel proficiency judgements for Exceeding and Strong levels on reset NAPLAN scale

Domain	Year level	Developing/Strong	Strong/Exceeding
Numeracy	3	377	492
	5	452	585
	7	503	618
	9	532	680
Reading	3	365	479
	5	454	558
	7	501	607
	9	536	635
Writing	3	364	500
	5	467	584
	7	507	596
	9	551	654
Spelling	3	386	496
	5	443	542
	7	489	594
	9	542	633
Grammar and punctuation	3	397	511
	5	478	598
	7	522	638
	9	534	629

An anomaly appears in that the lower boundaries of the Exceeding level in Years 7 and 9 of grammar and punctuation become disordered on the new scale. This suggests that the scale transformations between Year 7 and Year 9 in grammar and punctuation were different on old and new scales. As noted above, resetting the scale reflects the added information about student performance gained by the adaptive NAPLAN test design, and helps to resolve such historical anomalies.

Logarithmic regression

The judgements of the 4 year level panels in each domain were consolidated by applying logarithmic regression, so that the cut-points in each year level followed a smooth growth curve. The regressions in each domain are shown in Figure 33 to Figure 36.

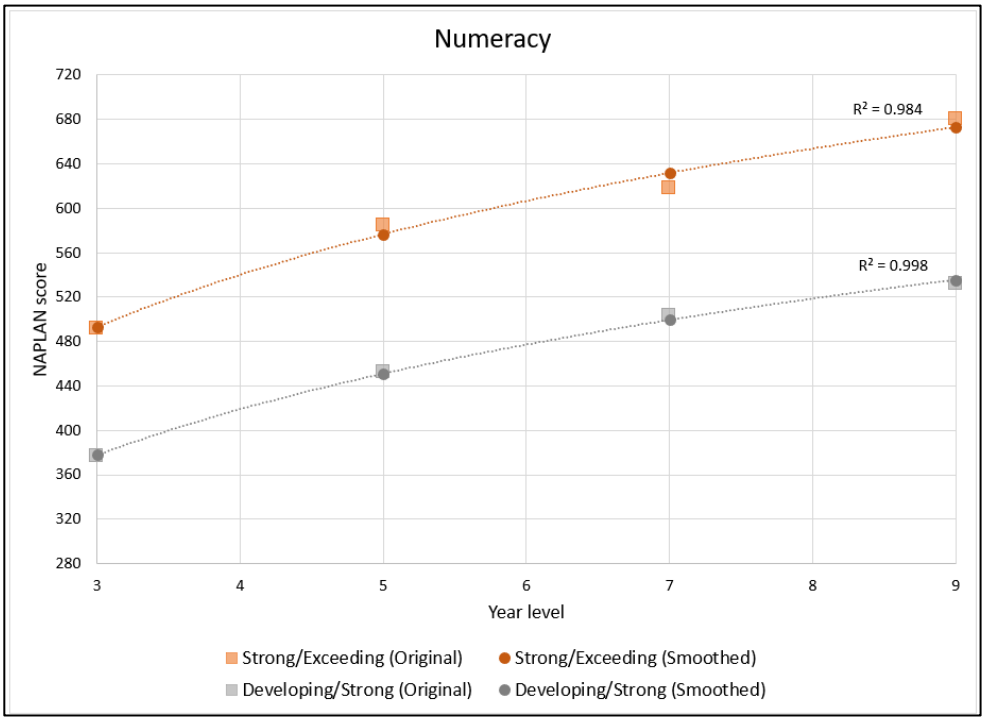


Figure 32: Logarithmic regression of proficiency cut-points (numeracy)

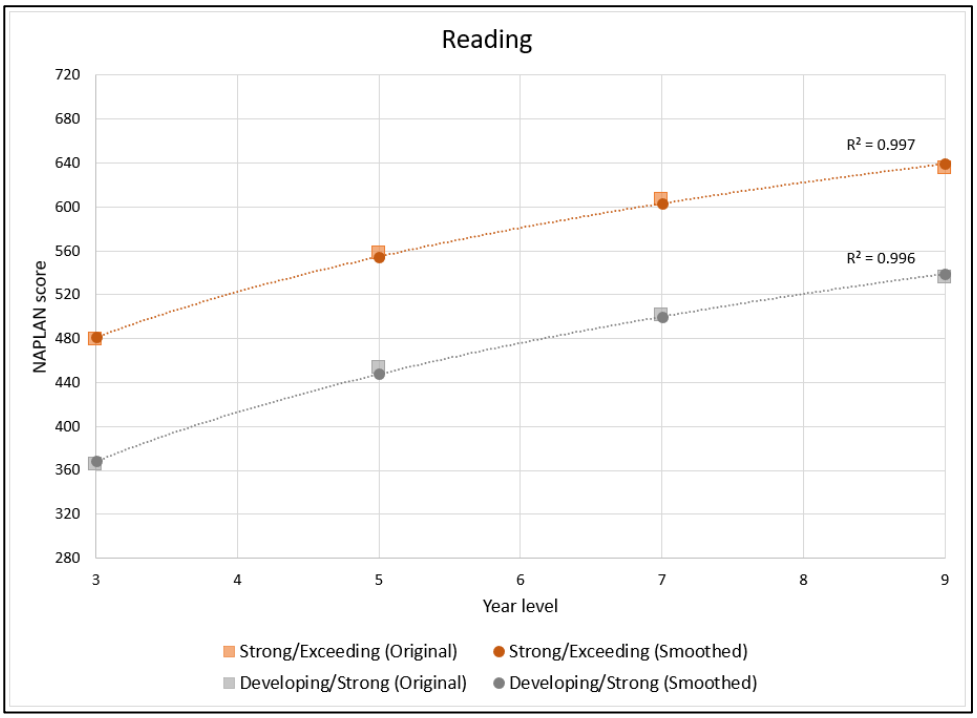


Figure 33: Logarithmic regression of proficiency cut-points (reading)

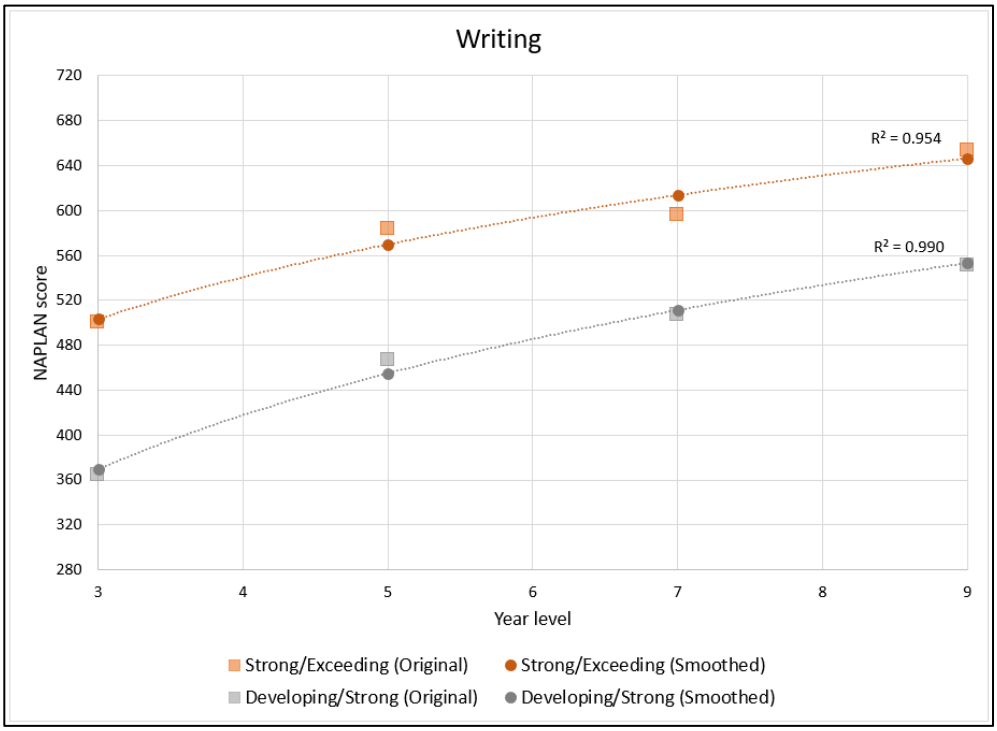


Figure 34: Logarithmic regression of proficiency cut-points (writing)

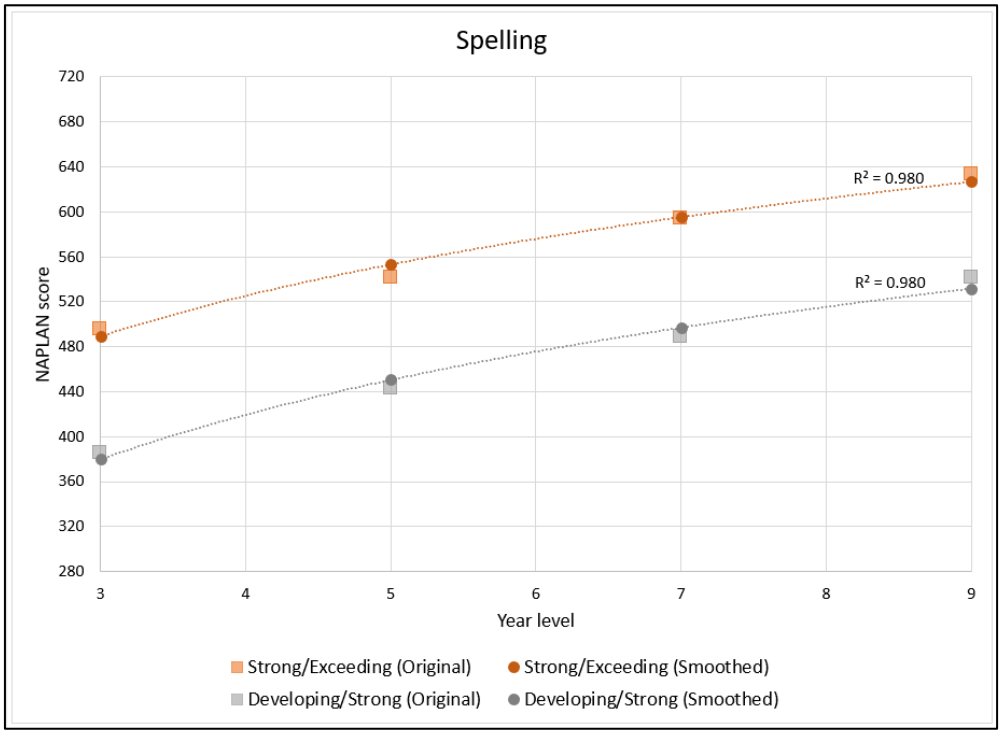


Figure 35: Logarithmic regression of proficiency cut-points (spelling)

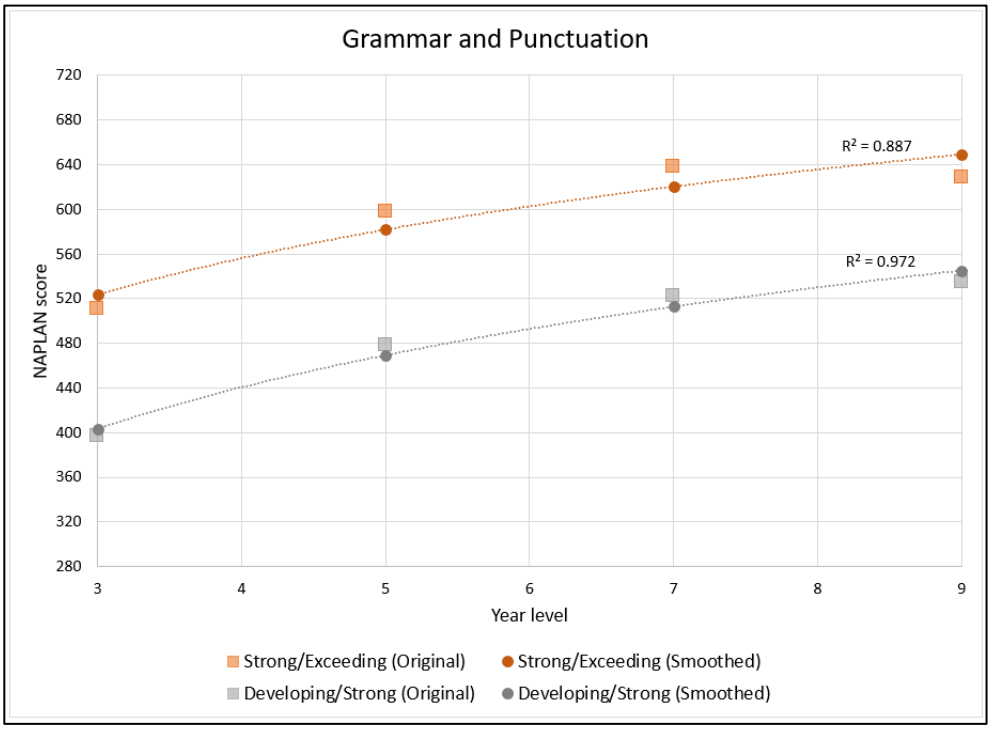


Figure 36: Logarithmic regression of proficiency cut-points (grammar and punctuation)

The proficiency level cut-points after logarithmic regression are shown in Table 64.

Table 64: Proficiency level cut-points after logarithmic regression

Domain	Year level	Developing/Strong	Strong/Exceeding
Numeracy	3	378	493
	5	451	577
	7	500	632
	9	536	673
Reading	3	368	481
	5	448	555
	7	500	603
	9	539	639
Writing	3	370	503
	5	455	570
	7	511	614
	9	553	647
Spelling	3	380	489
	5	451	553
	7	497	595
	9	532	627
Grammar and punctuation	3	404	523
	5	470	582
	7	513	620
	9	545	649

Cut-points between Needs additional support and Developing

Education ministers agreed in February 2023 that the policy intent of the Needs additional support level, which superseded the national minimum standard, would best be served if it were defined using the judgements made by the proficiency standards panels. One-third of the students below the lower cut-point would be assigned to Needs additional support, and the remaining two-thirds to Developing.

Taking into account exempt students, the formulation was:

n (Developing + Needs additional support + exempt) = 3 x n (Needs additional support + exempt)

Validation of cut-points

Based on the panel judgements, ACARA literacy and numeracy experts compiled descriptions of the skills associated with each proficiency level, using the Australian Curriculum and NAPLAN assessments as key reference documents.

Panels reconvened in May 2023 to review, refine and validate these descriptions. There were a few differences in the constitution of the panels, with some of the panellists who had set the standards being replaced due to unavailability.

Panels reviewed the proficiency level descriptions against the following complete set of policy descriptors:

- **Needs additional support**: The student's result indicates that they are not achieving the learning outcomes expected at the time of testing. They are likely to need additional support to progress satisfactorily.
- **Developing**: The student's result indicates that they are working towards expectations at the time of testing
- **Strong**: The student's result meets challenging but reasonable expectations at the time of testing.
- **Exceeding**: The student's result exceeds expectations at the time of testing.

A key task of the validation panels was to ensure that the descriptions for Needs additional support (NAS) and Developing did accurately reflect the policy descriptors for those levels, given that the cut-point between those levels had been set since the panels last convened.

After finalisation of the level descriptions, panellists were surveyed to confirm that they considered the descriptions to accurately reflect all levels. Agreement was at a very high level in all domains and year levels. The finalised and validated proficiency level descriptions can be accessed at https://www.nap.edu.au/naplan/results-and-reports/proficiency-level-descriptions.

These online proficiency level descriptions replace the example items that were presented in NAPLAN technical reports, until 2022, to exemplify the difficulty of each achievement band.

Final cut-points for NAPLAN 2023

The complete set of proficiency level cut-points for NAPLAN 2023 derived from this process is shown in Table 65.

Table 65: Proficiency level cut-points for NAPLAN 2023

Domain	Year level	NAS/Developing	Developing/Strong	Strong/Exceeding
Numeracy	3	311	378	493
	5	386	451	577
	7	431	500	632
	9	463	536	673
Reading	3	282	368	481
	5	377	448	555
	7	430	500	603
	9	464	539	639
Writing	3	296	370	503
	5	385	455	570
	7	439	511	614
	9	469	553	647
Spelling	3	294	380	489
	5	378	451	553
	7	430	497	595
	9	470	532	627
Grammar and punctuation	3	312	404	523
	5	397	470	582
	7	444	513	620
	9	460	545	649

These cut-points will stay in place for future years as a benchmark of the proficiency levels. Changes in performance will be visible by noting changes in the percentages of students at each level.

Chapter 8: Reporting of national results

NAPLAN produces several reports for a variety of audiences each year. The Student and School Summary Report (SSSR)[11] is a preliminary report for school staff with student and school level results. The Individual Student Report (ISR)[12] is a report for parents/carers about their child's NAPLAN achievement. The national results include final national statistics to inform policymakers and researchers. Additional reporting is also provided on the website My School[13], with results for individual schools, and is accessible to the general public. This chapter describes analysis for the national results.

Calculation of statistics using plausible values

All statistics included in the national report were based on plausible values. Plausible values are a type of 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. The formal notation for this is:

$$\theta = \frac{1}{5}\sum_{i=1}^{5}\theta_i \qquad (8)$$

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).

Note that plausible values should never be averaged at the student level.

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 scores or performance differences are accurate. More precisely, appropriate standard errors are used for statistically testing the likelihood that observed performance differences arose by chance, before concluding that a statistically meaningful difference exists.

Because the NAPLAN scale has been reset and 2023 is the first year of a restarted time series, only 2 types of error were estimated and combined as the standard error; there is no equating error estimated in 2023 as there is no trend information on the new scale to be reported. 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.

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 2023. In addition, even in census assessments, there is a certain amount of non-response that must be considered. 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 1999) was used to construct an approximation to the functional form of the estimated population characteristic that is a linear function of the original observations and hence is amenable to construction of a variance estimator.

The process of linearisation 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

[11] www.nap.edu.au/docs/default-source/default-document-library/how-to-interpret-the-sssr.pdf?sfvrsn=10
[12] www.nap.edu.au/results-and-reports/student-reports
[13] www.myschool.edu.au/

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 1999).

Taylor series variance estimation can be done using commercially available statistical software. For NAPLAN 2023, the Complex Samples module implemented in the SPSS software package and the SURVEYMEANS procedure in the SAS software package were used in parallel processing for checking. Examples of these procedures are included in Figure 37. The sampling error is equal to the square root of the sampling variance.

SPSS	**SAS**

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.
	proc surveymeans data=temp;
	cluster school_ID ;
	domain grade <subgroups>;
	var PV1-PV5;
	ods output domain=PVout;
	run;

Figure 37: 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:

$$B = \frac{1}{4}\sum_{i=1}^{5}(\theta_i - \theta)^2 \qquad (9)$$

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:

$$V = U + \left(1 + \frac{1}{5}\right)B \qquad (10)$$

$$SE = \sqrt{V} \qquad (11)$$

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 2023 performance in the reports.

Testing for differences

Because 2023 is the first year of the new NAPLAN scale, the only differences that can be computed are between subgroups participating in NAPLAN 2023; for example, between male and female students, or between jurisdictions. Differences of this type can be tested for significance using the standard errors estimated from the sampling variance and the measurement variance.

To illustrate how statistical testing of the subgroup performance differences was carried out in the NAPLAN context, a hypothetical example – focusing on differences in mean scores – is provided below.

The example considers 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, having zero covariance), a standard error for the difference between them can be computed using the following formula:

$$SE_{DIFF} = \sqrt{{SE_A}^2 + {SE_B}^2} \quad (12)$$

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 probability level of 0.05 was used for all statistical tests, with corresponding critical values of ±1.96.

The 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.

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 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.

The effect size for differences in means is given by *Hedges' g*, whose formula is:

$$g = \frac{m_2 - m_1}{s_p} \quad (13)$$

where m_1 is the sample mean of the first group, m_2 is the sample mean of the second group and s_p is the pooled standard deviation; that is, the square root of the pooled within-groups variance, weighted by number of cases in each group.

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \quad (14)$$

where n_1 and n_2 are the number of cases in group 1 and 2, respectively, and s_1^2 and s_2^2 are their variances.

The effect size given by *Hedges' g* is known to yield a biased estimate for the population value and is corrected using the following formula:

$$g_{unbiased} = g_{biased}\left[1 - \frac{3}{4(n_1+n_2-2)}\right] \quad (15)$$

Significance testing and effect size were combined to report the "nature of the difference" for comparisons of NAPLAN performance between subgroups as follows:

- "substantially above/below" refers to a difference that is statistically significant and large in size, where large means an effect size of greater than 0.5 / less than -0.5

- “above/below” refers to a difference that is statistically significant and small in size, where small means an effect size between 0.2 and 0.5 / between -0.2 and -0.5
- “close to” refers to a difference that is either not statistically significant or negligible in size, where negligible means an effect size of less than 0.2 but greater than -0.2.

References

Adams RJ, Wu ML, Cloney D and Wilson MR (2020) *ACER ConQuest: generalised item response modelling software* [computer software], version 5, Australian Council for Educational Research, Camberwell, Victoria.

Adams RJ 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) (2017) The Australian National Assessment Program Literacy and Numeracy (NAPLAN) assessment framework: NAPLAN Online 2017, ACARA, Sydney.

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 D (2007) "Automated Simultaneous Assembly of Multistage Testlets for a High-Stakes Licensing Examination", Educational and Psychological Measurement, 67(1): 5-20.

Camilli G and Shepard LA (1994) *Methods for identifying biased test items* (Vol. 4), Sage, Thousand Oaks.

Eggen TJ and Verhelst ND (2011) "Item calibration in incomplete testing designs", *Psicológica, 32*(1):107–132.

Hendrickson A (2007) "An NCME Instructional Module on Multistage Testing", *Educational Measurement: Issues and Practice, 26(*2).

Levy PS and Lemeshow S (1999) *Sampling of populations: methods and applications*, John Wiley & Sons, New York.

Luecht RM, Brumfield T and Breithaupt K (2006) "A testlet assembly design for adaptive multistage tests", *Applied Measurement in Education, 19*(3):189–202.

Masters GN (1982) "A Rasch model for partial credit scoring", *Psychometrika 47*: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, 293–360.

Rasch G (1960) *Probabilistic models for some intelligence and attainment tests*, Danmark Paedagpgiske Institut, Copenhagen.

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

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):427–50.

Wolter KM (1985) *Introduction to Variance Estimation*, Springer-Verlag, New York.

Appendices

Appendix A: Percentages and ability distribution by pathway

https://nap.edu.au/docs/default-source/naplan/appendix-a---percentages-and-ability-distribution-by-pathway-2023.pdf

Appendix B: Item analysis details

https://nap.edu.au/docs/default-source/naplan/appendix-b---item-analysis-details-2023.pdf

Appendix C: Item summary tables

https://nap.edu.au/docs/default-source/naplan/appendix-c---item-summary-tables-2023.pdf

Appendix D: Item characteristic curves

https://nap.edu.au/docs/default-source/naplan/appendix-d---item-characteristic-curves-2023.pdf

Appendix E: Expected score curves (writing)

https://nap.edu.au/docs/default-source/naplan/appendix-e---expected-score-curves-(writing)-2023.pdf

Appendix F: Item-person maps

https://nap.edu.au/docs/default-source/naplan/appendix-f---item-person-maps-2023.pdf

Appendix G: Gender DIF analysis

https://nap.edu.au/docs/default-source/naplan/appendix-g---gender-dif-analysis-2023.pdf

Appendix H: Language background DIF analysis

https://nap.edu.au/docs/default-source/naplan/appendix-h---language-background-dif-analysis-2023.pdf

Appendix I: Indigenous status DIF analysis

https://nap.edu.au/docs/default-source/naplan/appendix-i---indigenous-status-dif-analysis-2023.pdf

Appendix J: DIF summary tables

https://nap.edu.au/docs/default-source/naplan/appendix-j---dif-summary-tables-2023.pdf

Appendix K: Jurisdictional DIF

https://nap.edu.au/docs/default-source/naplan/appendix-k---jurisdictional-dif-2023.pdf

Appendix L: Device DIF

https://nap.edu.au/docs/default-source/naplan/appendix-l---device-dif-2023.pdf

Appendix M: Vertical link item comparisons

https://nap.edu.au/docs/default-source/naplan/appendix-m---vertical-link-item-comparisons-2023.pdf

Appendix N: Data cleaning and validation exception rules

https://nap.edu.au/docs/default-source/naplan/appendix-n---data-cleaning-and-validation-exception-rules-2023.pdf