NAPLAN 2024

Technical Report

March 2025

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

© Australian Curriculum, Assessment and Reporting Authority (ACARA) 2024, unless otherwise indicated. Subject to the exceptions listed below, copyright in this document is licensed under a Creative Commons Attribution 4.0 International (CC BY) licence (https://creativecommons.org/licenses/by/4.0/). This means that you can use these materials for any purpose, including commercial use, provided that you attribute ACARA as the source of the copyright material.

Exceptions

The Creative Commons licence does not apply to:

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

Attribution

ACARA requests attribution as: "© Australian Curriculum, Assessment and Reporting Authority (ACARA) 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 provides 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 adaptive 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 2024, 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 that 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 that were used in previous NAPLAN cycles.

In 2024, NAPLAN results were reported on these reset scales, and against the new proficiency levels.

Four outcome reports were produced for NAPLAN 2024.

- 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 and carers about their child's performance on the NAPLAN tests.
- The NAPLAN 2024 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 2024 and all previous cycles.
- My School 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 My School were not available in 2023 or 2024, 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 2024.

The aim of this technical report is to describe in detail the methodology used for NAPLAN 2024.

- Chapter 2 describes how items were developed, trialled, analysed and scaled in 2024 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.
- 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 link 2024 results to the NAPLAN measurement scales established in 2023.
- Chapter 7 describes the use of proficiency levels for reporting.
- Chapter 8 describes the methodology used for reporting of NAPLAN 2024 performance.

Chapter 2: Item development and item trial

This chapter describes the processes through which NAPLAN items, prompts and texts were developed and trialled in 2024, 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 are developed and trial tests constructed. The second part describes the item trial administration. 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 over 3 weeks across May and June 2024. The development and trialling of these items is 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 area specialists and proofreaders.

ACARA worked with a team of First Nations Australian educators to review the reading materials for inclusivity. For all informative and persuasive 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 2024, prompts for writing tests were developed and trialled according to the following process:

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

Expert panels undertook 4 separate reviews of the prompts as they were refined and developed to be trial-ready. An initial review was made 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 the first review, the panels made overall judgements of which writing tasks might be prioritised for administration in NAPLAN, providing feedback where necessary. In later reviews, 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 (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 are administered to a representative, stratified sample of schools and students. The trial tests include common items from the previous year’s NAPLAN tests so that the trial results can be equated to the historical NAPLAN scale by a common-item 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 enables review of the psychometric properties of the items at both year levels. Depending on these properties, the items can be used for the main study in only one year level or can 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 (CoL) test is separated into its 2 component sections: spelling, and grammar and punctuation. All spelling items are constructed response, so are classified instead into audio dictation (AD) or proofreading (PR) formats.

Table 1. Composition of the 2024 numeracy item trial domain

	MC	Other	Total
Year 3	117	105	222
Year 5	149	109	258
Year 7	179	133	312
Year 9	177	135	312
Total	622	482	1,104

Table 2. Composition of the 2024 reading item trial

	MC	Other	Total
Year 3	213	39	252
Year 5	287	49	336
Year 7	346	54	400
Year 9	257	43	300
Total	1,103	185	1,288

Table 3. Composition of the 2024 spelling item trial

	AD	PR	Total
Year 3	120	216	336
Year 5	120	216	336
Year 7	120	216	336
Year 9	120	216	336
Total	480	864	1,344

Table 4. Composition of the 2024 grammar and punctuation item trial

	MC	Other	Total
Year 3	147	189	336
Year 5	154	182	336
Year 7	154	182	336
Year 9	153	183	336
Total	608	736	1,344

Item trial design: writing

The 10 writing tasks were each trialled at Years 3, 5, 7 and 9. The tasks were administered in a rotational design based on classes, not individual students within each class as was the case for other domains. Some students completed 2 writing tasks. Students in Years 5, 7 and 9, and the majority of students in Year 3, completed their writing task(s) online. Some Year 3 students completed one task online and one task on paper.

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 240 schools each were chosen with probability proportional to school size. In the Year 9 secondary sample, only 226 of these schools were identified for selection as determined by the domain pair sequence allocation described below. 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, Brethren schools)
- schools without NAPLAN performance data
- schools that participated in the NAPLAN 2023 item trial.

Schools sampled for the 2024 NAP–Civics and Citizenship (NAP–CC) main study were also excluded from the NAPLAN 2024 item trial sample, whereas sample replacements from NAP–CC were not 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, which is 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 (Numeracy-Reading, Numeracy-CoL, Numeracy-Writing, Reading-CoL, Reading-Writing, CoL-Writing and Writing-Writing) 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 48 larger schools. For the secondary sample, 48 larger schools were allocated a second domain pair in Year 7, whereas only 46 larger schools were allocated a second domain pair in Year 9.

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 each year level.

		Domain pairs							
		CW	NC	NR	NW	RC	RW	WW	Total
Year 3	1st domain pair	54	53	14	13	53	13	40	240
	2nd domain pair	10	11	2	3	11	3	8	48
	Class total	64	64	16	16	64	16	48	288
Year 5	1st domain pair	54	53	14	13	53	40	13	240
	2nd domain pair	10	11	2	3	11	8	3	48
	Class total	64	64	16	16	64	48	16	288
Year 7	1st domain pair	54	53	14	13	53	40	13	240
	2nd domain pair	10	11	2	3	11	8	3	48
	Class total	64	64	16	16	64	48	16	288
Year 9	1st domain pair	54	53	14	13	53	13	26	226
	2nd domain pair	10	11	2	3	11	3	6	46
	Class total	64	64	16	16	64	16	32	272

Survey

A single-item survey was included at the start of all trial tests, collecting information about student gender.

The responses to this item were used in the analysis of student performance to determine whether there was evidence of differential item functioning (DIF) by gender.

Trial participation

A total of 469 schools across all states and territories participated. Note that while 240 primary schools and 240 secondary schools were sampled – from which only a selection were administered assessments in year 9, the total number of schools reflects the fact that some schools provided both primary and secondary classes.

The number of students who sat the tests in each non-writing domain (where this is defined as having responded to at least 5 items) 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	2,053	2,806	2,776	1,940	9,575
Conventions of language	4,187	4,218	4,101	3,919	16,425
Numeracy	2,235	2,184	2,072	2,027	8,518

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	357	350	361	334	1,402
Task 2	356	354	368	339	1,417
Task 3	355	350	362	333	1,400
Task 4	356	349	364	339	1,408
Task 5	355	331	347	316	1,349
Task 6	349	330	347	311	1,337
Task 7	345	335	353	311	1,344
Task 8	345	331	348	313	1,337
Task 9	333	426	360	330	1,449
Task 10	334	390	313	319	1,356
Task 1 paper	413	0	0	0	413
Task 5 paper	410	0	0	0	410
Total	4,308	3,546	3,523	3,245	14,622

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 event that was held across 3 weeks in May and June 2024. The trial was supported by trained invigilators in all schools.

Marking of writing responses

A team of experienced NAPLAN markers was engaged by the item trial administration contractor to mark the writing responses. Writing responses were extracted from the platform and provided to the contractor. Paper responses were returned to the contractor by invigilators and the responses scanned for uploading into the marking platform. ACARA's writing test manager attended the marking centre for the first week of the marking operation and facilitated the training of the markers. The ACARA writing team remained in communication with ACER staff during the marking and was able to oversee the marking process remotely during the second week of marking. Once the marking of each prompt was completed, a debriefing session was held with the markers, who also completed a short survey. 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

The trial data was extracted from the assessment platform and then sent, along with scores for the writing responses, to an external contractor for data processing, analysis, and scaling.

Analysis of 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 (missing responses that are followed by valid responses, plus the first missing response that is followed only by other missing responses) were coded "9", trailing missing responses (all other missing responses are of this type) were coded "M", and items not administered to a student were coded "R".

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 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 – measured by weighted mean-square (MNSQ) 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 is investigated using these measures. The procedures employed are very similar, whether they are 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 requires 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 requirement. 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 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 are able to 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.

Analysis of 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 lower and upper year levels.

The NTWG and MQT (Marking Quality Team) were consulted regarding the allocation of writing tasks to year levels and the final sequence across the 2-day writing window. This informed the design of the NAPLAN writing tests for 2025, 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 2024 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.

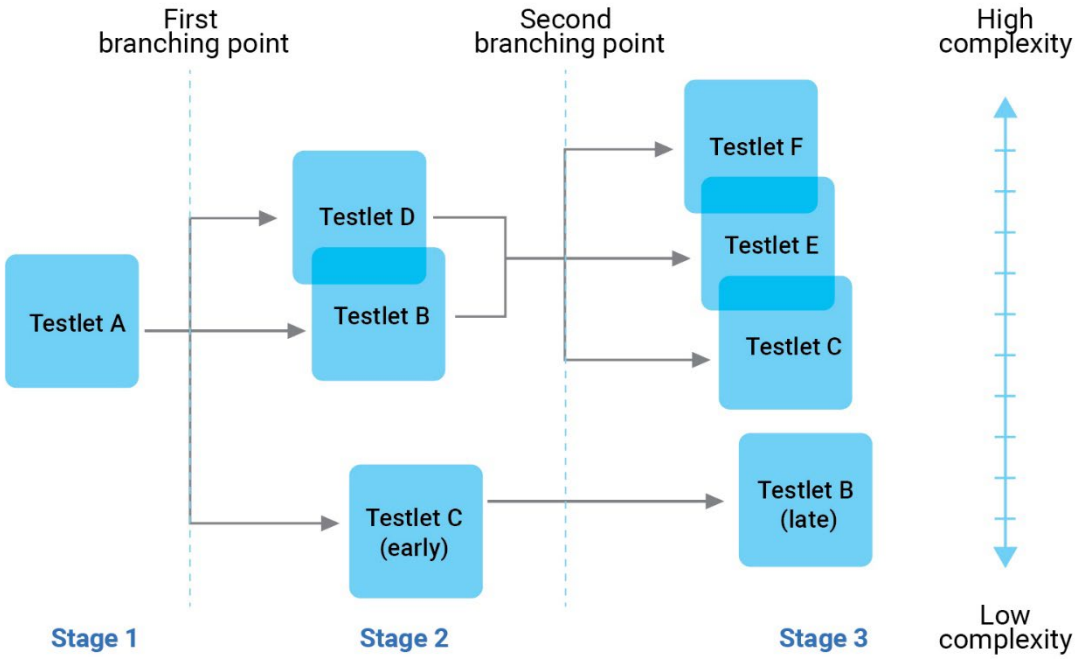


Figure 1. The multistage tailored test design for numeracy, reading and grammar and punctuation

Students at each year level start with testlet A. Each student's answers to testlet A determine the testlet they are 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 questions the student answers correctly and the average difficulty of the items 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 are advised that they cannot return to the non-calculator section once they have moved to the calculator-allowed section.

The conventions of language test includes a spelling section and a grammar and punctuation section, each with 2 branching points. Students are advised that they cannot return to the spelling section once they have 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.

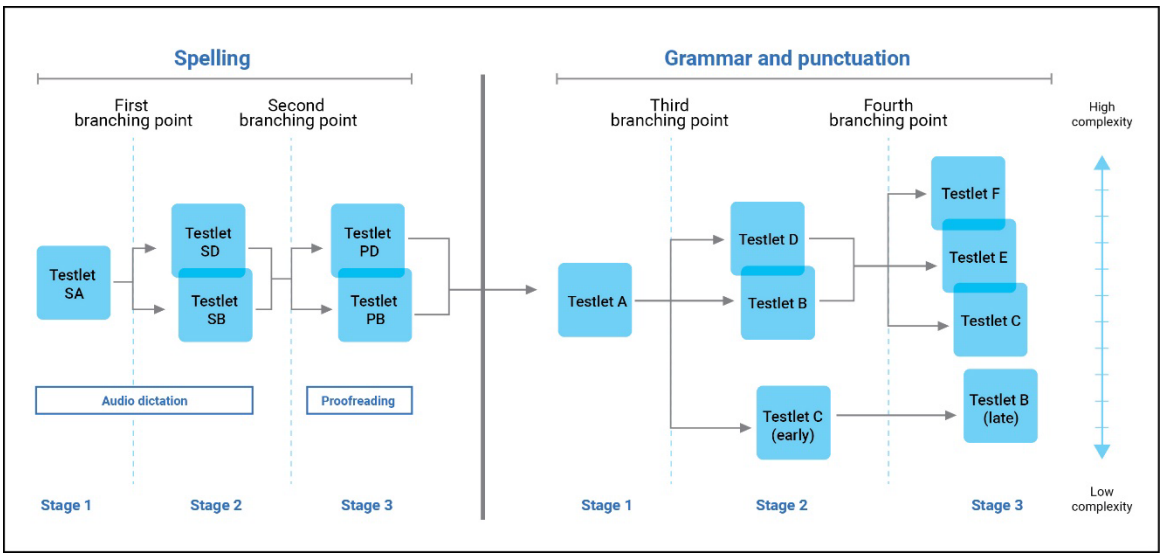


Figure 2: Online test design for conventions of language

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 2024 NAPLAN tests based on their performance in past item trials or in the 2023 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 between 2024 and 2023. 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 spelling and grammar and punctuation 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

Numeracy		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

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

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

Conventions of language	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] CA – calculator-allowed
[2] NC – non-calculator

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

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

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

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

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

Spelling	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

Grammar & punctuation	Lower bound	Upper bound	Average
A	-3.0	1.0	-0.5
B	-2.0	0.5	-0.8
C	-3.5	-0.5	-2.0
D	-0.5	2.0	0.8
E	-1.5	1.5	0.0
F	0.5	3.5	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: counts of item types by year level

Numeracy	MC/MCs items	CR items	Technology-enhanced items	Total in suite
Year 3	117	57	42	216
Year 5	147	62	43	252
Year 7	167	72	49	288
Year 9	157	82	49	288

Table 16. NAPLAN online reading: counts of item types by year level

Reading	MC/MCs items	CR items	Technology-enhanced items	Total in suite
Year 3	191	-	43	234
Year 5	196	-	38	234
Year 7	248	-	40	288
Year 9	241	-	47	288

Table 17. NAPLAN online conventions of language: counts of item types by year level

Conventions of language		MC/MCs items	CR items	Technology-enhanced items	Total in suite
Spelling	Year 3	0	129	0	129
	Year 5	0	129	0	129
	Year 7	0	129	0	129
	Year 9	0	129	0	129
Grammar and punctuation	Year 3	65	0	97	162
	Year 5	55	0	107	162
	Year 7	54	0	108	162
	Year 9	56	0	106	162

Numeracy test content

Items are written to cover the Australian Curriculum 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%	56%	56%	56%	56%
Measurement and geometry	25–35%	31%	31%	31%	31%	31%
Statistics and probability	10–20%	15%	14%	14%	14%	14%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Fluency	15–25%	20%	28%	23%	17%	16%
Understanding	25–35%	31%	39%	35%	31%	25%
Problem-solving	25–35%	29%	17%	23%	32%	36%
Reasoning	15–25%	20%	17%	19%	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%	52%	49%	52%	52%	52%
Measurement and geometry	25–35%	31%	34%	31%	31%	31%
Statistics and probability	10–20%	17%	17%	17%	17%	17%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Fluency	15–25%	20%	22%	17%	17%	21%
Understanding	25–35%	29%	37%	30%	28%	21%
Problem-solving	25–35%	30%	24%	32%	33%	36%
Reasoning	15–25%	21%	17%	21%	22%	23%

Table 20. NAPLAN numeracy Year 7 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Number and algebra	50–60%	54%	54%	54%	54%	54%
Measurement and geometry	25–35%	31%	31%	31%	31%	31%
Statistics and probability	10–20%	15%	15%	15%	15%	15%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Fluency	15–25%	20%	24%	23%	24%	22%
Understanding	25–35%	30%	37%	29%	24%	19%
Problem-solving	25–35%	29%	21%	28%	33%	36%
Reasoning	15–25%	21%	19%	19%	19%	23%

Table 21. NAPLAN numeracy Year 9 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Number and algebra	50–60%	54%	54%	54%	54%	54%
Measurement and geometry	25–35%	31%	31%	31%	31%	31%
Statistics and probability	10–20%	15%	15%	15%	15%	15%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Fluency	15–25%	19%	23%	25%	22%	17%
Understanding	25–35%	31%	39%	31%	28%	28%
Problem-solving	25–35%	28%	17%	22%	30%	36%
Reasoning	15–25%	22%	22%	23%	20%	19%

Reading test content

The reading tests primarily assess the *Literacy* strand of the Australian Curriculum, 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%	21%	21%	15%	18%	21%
Literature	5–15%	8%	6%	9%	9%	9%
Literacy	60–80%	72%	74%	75%	74%	70%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Locating and identifying	30–60%	50%	60%	51%	51%	45%
Integrating and interpreting	35–60%	48%	40%	48%	47%	52%
Analysing and evaluating	0–15%	2%	0%	1%	2%	3%
Text content	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Number of texts	-	-	7	6	6	6
Average word count	-	-	94	146	179	209

Table 23. NAPLAN reading Year 5 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Language	15–25%	19%	19%	19%	19%	21%
Literature	5–15%	8%	7%	9%	9%	8%
Literacy	60–80%	73%	74%	72%	72%	71%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Locating and identifying	30–60%	40%	47%	43%	35%	32%
Integrating and interpreting	35–60%	53%	46%	50%	57%	58%
Analysing and evaluating	0–15%	7%	7%	7%	8%	9%
Text content	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Number of texts	-	-	6	6	6	6
Average word count	-	-	191	218	255	276

Table 24. NAPLAN reading Year 7 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Language	15–25%	20%	19%	18%	20%	22%
Literature	10–20%	11%	7%	8%	13%	15%
Literacy	55–75%	69%	74%	74%	67%	63%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Locating and identifying	15–45%	34%	44%	43%	35%	28%
Integrating and interpreting	40–65%	55%	50%	53%	56%	59%
Analysing and evaluating	5–30%	10%	6%	4%	9%	13%
Text content	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Number of texts	-	-	9	9	9	9
Average word count	-	-	224	275	299	318

Table 25. NAPLAN reading Year 9 test content by pathway

Strand	Target range	Overall	ABC	ABE	ADE	ADF
Language	15–25%	22%	19%	22%	26%	24%
Literature	10–20%	10%	7%	11%	12%	11%
Literacy	55–75%	68%	74%	67%	63%	65%
Proficiency	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Locating and identifying	15–45%	31%	42%	35%	26%	22%
Integrating and interpreting	40–65%	55%	52%	56%	58%	56%
Analysing and evaluating	5–30%	15%	6%	9%	15%	22%
Text content	**Target range**	**Overall**	**ABC**	**ABE**	**ADE**	**ADF**
Number of texts	-	-	9	9	9	9
Average word count	-	-	214	275	311	328

Conventions of language test content

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

- Audio dictation: the student plays a recording of the word, along with a sentence where the word is used in context, then the student is asked to correctly spell the word.
- Proofreading (mistake identified): a sentence contains a misspelled word that is highlighted for the student. The student is asked to correctly spell the word.
- Proofreading (mistake not identified): a sentence contains a misspelled word that is not highlighted for the student. The student is 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 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

Year 3	Target range	Overall	SASBPB	SASBPD	SASDPB	SASDPD
Audio dictation	55–65%	58%	64%	64%	64%	64%
Mistake identified	15–25%	24%	21%	20%	21%	20%
Mistake not identified	15–25%	18%	15%	16%	15%	16%

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

Year 3	Target range	Overall	ABC	ABE	ADE	ADF
Grammar	65–75%	65%	67%	64%	64%	67%
Punctuation	25–35%	35%	33%	36%	36%	33%

Table 28: NAPLAN spelling Year 5 test content by pathway

Year 5	Target range	Overall	SASBPB	SASBPD	SASDPB	SASDPD
Audio dictation	55–65%	58%	64%	64%	64%	64%
Mistake identified	15–25%	22%	21%	17%	21%	17%
Mistake not identified	15–25%	19%	15%	19%	15%	19%

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

Year 5	Target range	Overall	ABC	ABE	ADE	ADF
Grammar	65–75%	65%	65%	64%	65%	67%
Punctuation	25–35%	35%	35%	36%	35%	33%

Table 30: NAPLAN spelling Year 7 test content by pathway

Year 7	Target range	Overall	SASBPB	SASBPD	SASDPB	SASDPD
Audio dictation	55–65%	58%	64%	64%	64%	64%
Mistake identified	15–25%	22%	19%	20%	19%	20%
Mistake not identified	15–25%	19%	17%	16%	17%	16%

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

Year 7	Target range	Overall	ABC	ABE	ADE	ADF
Grammar	65–75%	65%	65%	65%	65%	64%
Punctuation	25–35%	35%	35%	35%	35%	36%

Table 32: NAPLAN spelling Year 9 test content by pathway

Year 9	Target range	Overall	SASBPB	SASBPD	SASDPB	SASDPD
Audio dictation	55–65%	58%	64%	64%	64%	64%
Mistake identified	15–25%	20%	19%	16%	19%	16%
Mistake not identified	15–25%	22%	17%	20%	17%	20%

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

Year 9	Target range	Overall	ABC	ABE	ADE	ADF
Grammar	65–75%	65%	65%	65%	65%	65%
Punctuation	25–35%	35%	35%	35%	35%	35%

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 difficulty 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 (CA) items preceded the non-calculator (NC) 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

	Subdomain	Number of items	Time available
Year 3	Spelling	25	45 minutes
	Grammar and punctuation	25	
Year 5	Spelling	25	45 minutes
	Grammar and punctuation	25	
Year 7	Spelling	25	45 minutes
	Grammar and punctuation	25	
Year 9	Spelling	25	45 minutes
	Grammar and punctuation	25	

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%	52%	56%	56%
Measurement and geometry	25–35%	31%	33%	29%	29%
Statistics and probability	10–20%	14%	14%	15%	15%
Proficiency	**Target range**	**Year 3**	**Year 5**	**Year 7**	**Year 9**
Fluency	15–25%	19%	21%	19%	19%
Understanding	25–35%	31%	29%	29%	31%
Problem-solving	25–35%	31%	29%	33%	29%
Reasoning	15–25%	19%	21%	19%	21%

Table 38: Test content – reading paper tests

Strand	Target range	Year 3	Year 5	Year 7	Year 9
Language	10–20%	13%	21%	17%	17%
Literature	10–20%	10%	10%	11%	15%
Literacy	50–70%	77%	69%	72%	69%
Proficiency	**Target range**	**Year 3**	**Year 5**	**Year 7**	**Year 9**
Locating and identifying	20–40%	56%	44%	30%	27%
Integrating and interpreting	40–60%	38%	49%	57%	54%
Analysing and evaluating	20–40%	5%	8%	13%	19%
Text content	**Target range**	**Year 3**	**Year 5**	**Year 7**	**Year 9**
Stimulus texts		6	6	8	8
Average word count		183	240	289	299

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%	72%	72%	68%	72%
Punctuation	25–35%	28%	28%	32%	28%

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 2024, 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 2024, 4 writing prompts were used for the paper and online modes of the writing tests across Years 3, 5, 7 and 9. Of these 4 prompts, one was assigned to the Year 3 test. Two prompts – the Year 3 prompt plus another – were assigned to the Year 5 tests. The remaining 2 prompts were assigned to the Years 7 and 9 tests. A further 3 prompts were kept in reserve in case of widespread technical issues or a security breach. No reserves were required in 2024. 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 33). Each prompt has closely scripted scaffolding, or instructions. All prompts had been trialled and the prompts selected for the 2024 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 see or hear the prompt, whereas students doing the test on paper see the paper prompt and have it read to them immediately prior to the start of the test timer. Therefore, an additional 2 minutes is allocated to the online tests to allow students to read and/or listen to the audio recording of the prompt. It is recommended that students divide their time 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 one million student scripts were marked nationally across the 5 marking operations. In 2024, 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 training using the same materials and were subject to comparable quality assurance measures.

Nationally, all markers were trained with the same content to ensure continuity with previous years and consistency across jurisdictions.

ACARA provided 2 comprehensive online Writing Marker Training courses to test administration authorities for use in training new and experienced markers and leaders. The courses were 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.

Training was conducted in the lead-up to the marking period. Training consisted of intensive training in the writing criteria, using the marking guide exemplars and training scripts with detailed commentaries explaining the criterion scores. Markers also completed practice scripts and qualification scripts to demonstrate their capabilities before commencing marking of student writing.

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.

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 is issued when the first marking centre commences marking, and the last control is issued on the final day of the last marking centre. However, as each jurisdiction has a slightly different marking window, not all controls are 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 required by the national protocols. Check-marking occurs for each marker and is 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 must cover at least 10% of all scripts marked across the marking operation (although in some centres 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 to 4 points discrepant	5 to 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 1 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 criterion	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 that 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 multistage 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 2024 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.

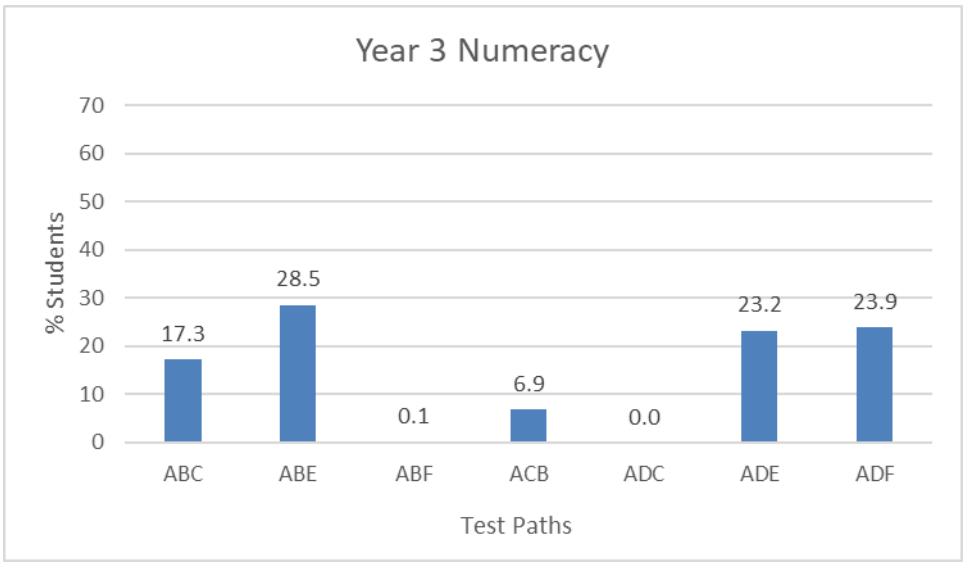


Figure 3. Percentage of students assigned to each pathway in Year 3 numeracy

Ability distributions by pathway are illustrated in Figure 4. Patterns of ability distributions across pathways were roughly as expected. That is, students ending with testlet F had the highest ability distribution. 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.

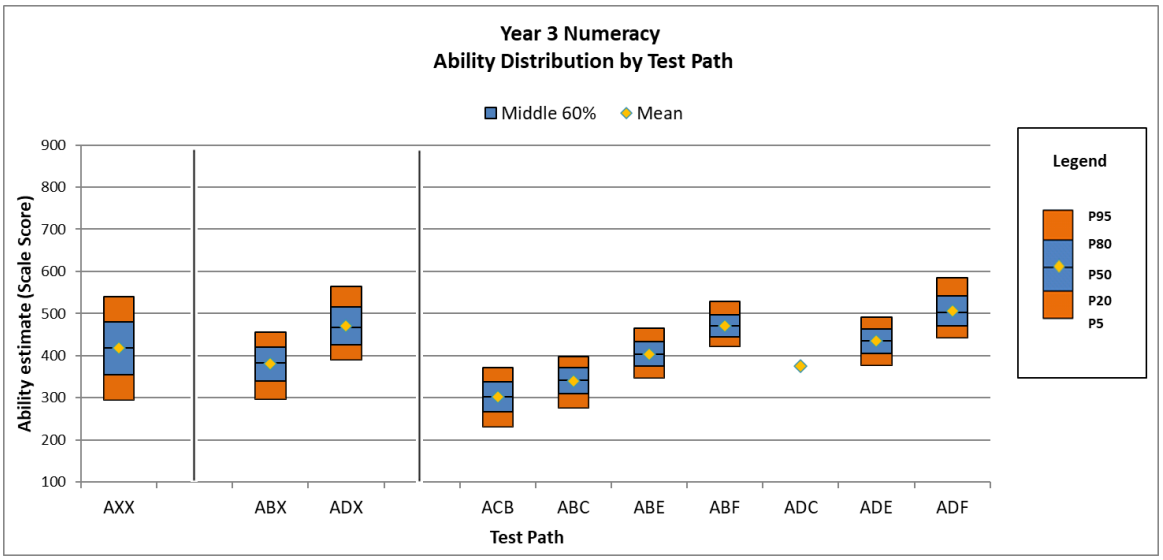


Figure 4. Ability distribution by pathway for Year 3 numeracy

Chapter 4: Data collection and preparation

This chapter describes data collection and delivery, data validation and data preparation for NAPLAN 2024. The chapter focuses on how data for online and paper tests is collected by test administration authorities (TAAs) from each jurisdiction and delivered to ACARA. It also describes 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 NAPLAN 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 provide guidance to schools and school systems in the collection of information on student background characteristics, using the nationally agreed standard measures of the characteristics. The manual is 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 Central Analysis of Data contractor in 5 main batches:
- online test data, sequentially by test domain, including both scored and raw response data, which is used for item calibration
- NAP Analysis Extracts (NAE) for preliminary analysis and to generate initial student and school summary reports (SSSRs)
- calibration extracts to calibrate writing criteria
- 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
- 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 2024 National Results.

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

Online tests

Education Services Australia (ESA) managed the online national assessment platform (the platform) through which the NAPLAN 2024 online tests were delivered. The Australian Council for Educational Research (ACER), as the analysis contractor in 2024, received the online test data extracted from the platform. Data files were 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. A list of the exception rules is included in Appendix B.

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.

In 2024, one TAA was unable to provide the SMF on time along with the additional participation codes (Z and B) for the writing domain, which required that writing be calibrated with partially incomplete data. The aim of the writing calibration and equating process is to establish whether the base year (2023) calibration can be used in subsequent years. Given the lack of complete data, this process was carried out in 2 ways: one excluding this TAA's data and the other including this TAA's data but with modifications to maintain the same ratio of Z and B students as in 2023. Results from these 2 approaches produced essentially the same item parameters, both being consistent with those from 2023. The calibration and equating results excluding this TAA's data are presented in the 2024 Technical Report.

Data preparation

Test data was recoded by the contractor prior to data analysis. The recoding rules depend on participation status, and are shown below.

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 NAPLAN national protocols for test administration, section 5.4.
- Data treatment
  - Item response data were 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 NAPLAN national protocols for test administration, section 5.5.
- Data treatment
  - Item response data were 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 NAPLAN national protocols for test administration, section 5.3.
- Data treatment
  - Item response data are 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 NAPLAN national protocols for test administration, section 5.2.
- Data treatment
  - Item response data are 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 45. 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 45: 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 paths, the most difficult test path (A1-D1-F1) and the test path 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.

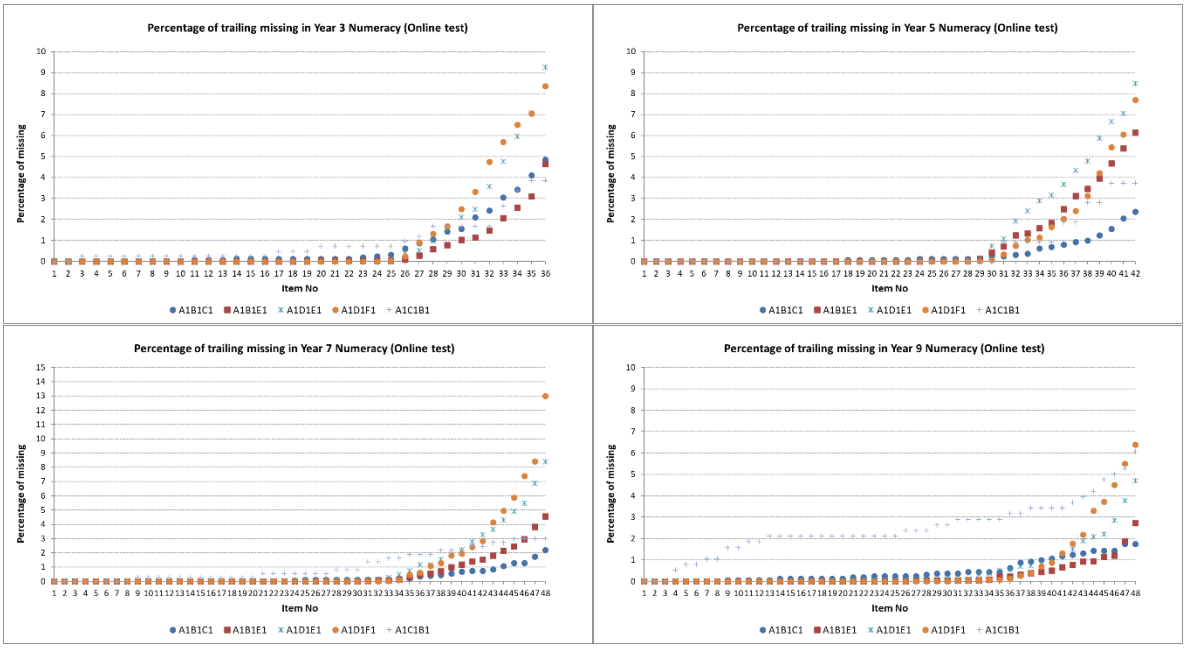


Figure 5: Trailing missing percentage in numeracy

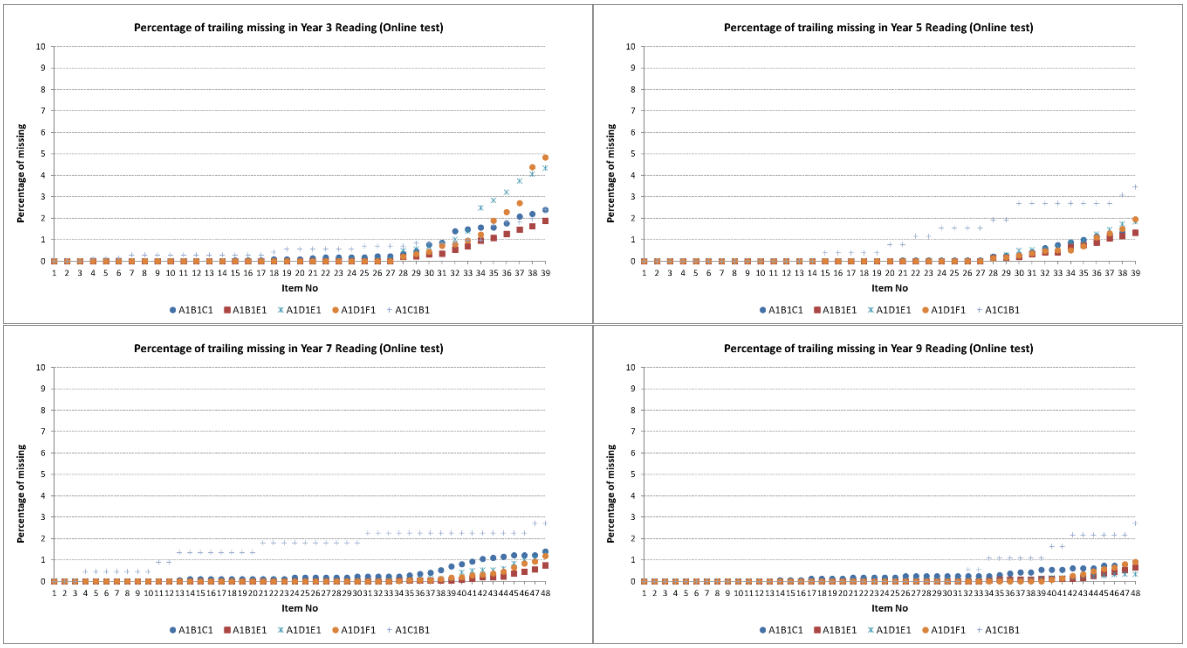


Figure 6: Trailing missing percentage in reading

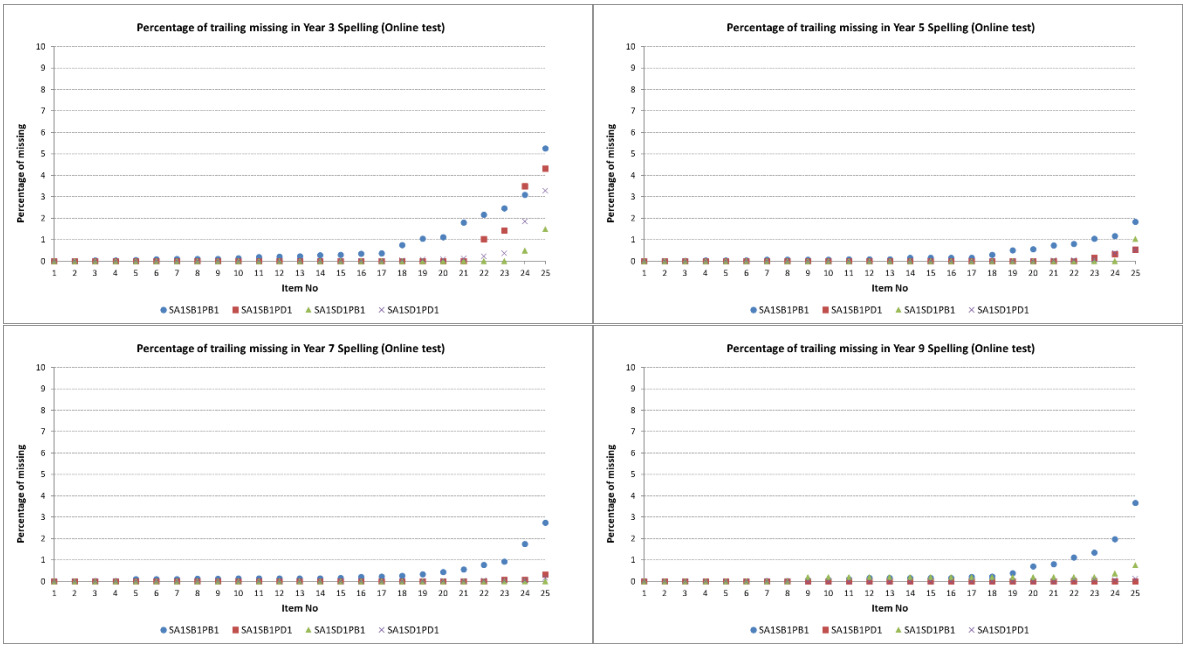


Figure 7: Trailing missing percentage in spelling

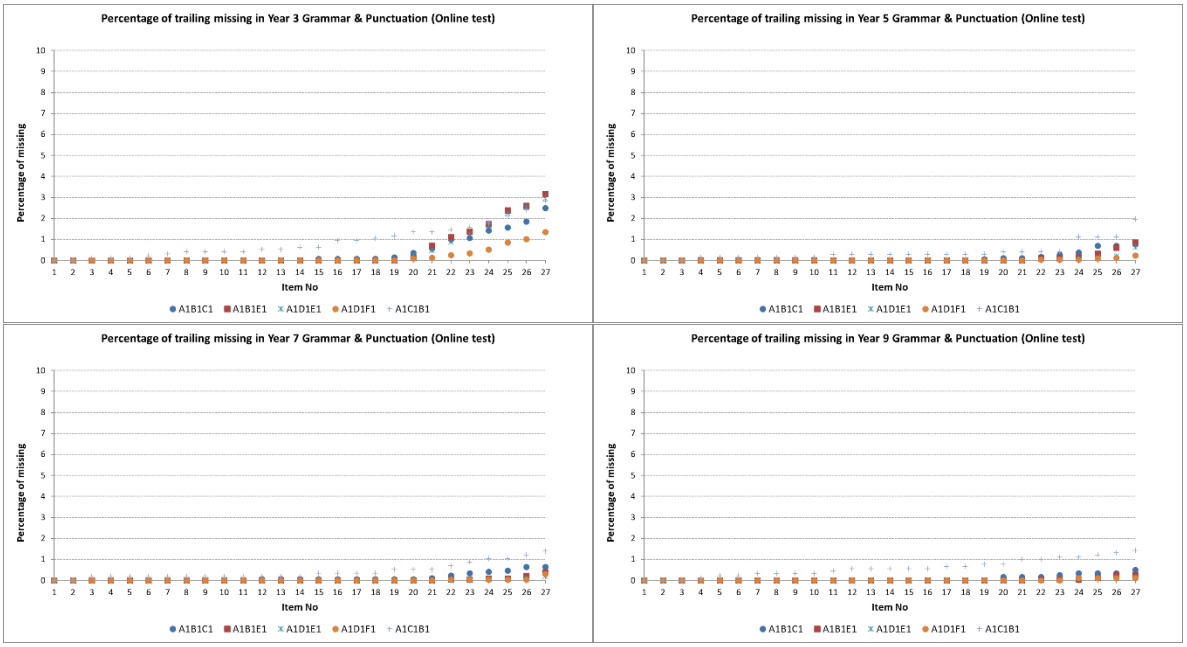


Figure 8: Trailing missing percentage in grammar and punctuation

Final student participation rates

The participation category diagram for NAPLAN 2024, 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.

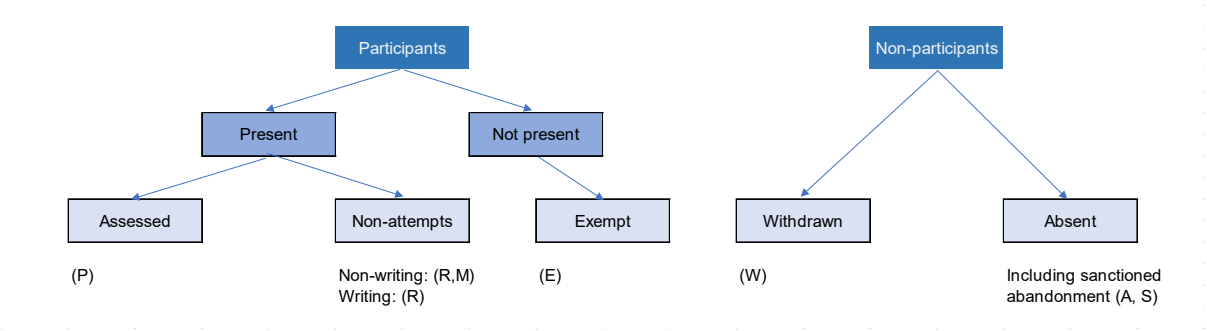


Figure 9: NAPLAN 2024: Participation Categories

Final student participation rates for NAPLAN 2024 are recorded in Table 46 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 46.

Table 46: Student participation rates

TAA	Year level	Numeracy (%)	Reading (%)	Writing (%)	Spelling (%)	Grammar and punctuation (%)	At least one test (%)
NSW	3	96.4	96.9	95.9	96.5	96.5	97.5
Vic.	3	95.2	95.6	94.7	95.1	95.1	96.8
Qld	3	92.0	93.2	92.6	92.4	92.4	94.4
WA	3	95.3	96.0	95.9	95.5	95.5	97.0
SA	3	94.8	95.3	94.2	94.8	94.8	96.1
Tas.	3	96.2	96.5	95.6	95.9	95.9	97.7
ACT	3	94.5	94.6	93.2	93.9	93.9	95.7
NT	3	81.5	83.3	82.2	81.3	81.3	87.5
Aus.	**3**	**94.7**	**95.4**	**94.6**	**94.8**	**94.8**	**96.4**
NSW	5	96.9	97.4	96.9	97.0	97.0	98.0
Vic.	5	95.7	96.2	95.9	95.6	95.6	97.2
Qld	5	92.1	93.4	93.3	92.5	92.5	94.5
WA	5	95.8	96.6	96.5	96.1	96.1	97.3
SA	5	95.0	95.7	95.3	95.1	95.1	96.5
Tas.	5	96.2	97.0	96.4	96.4	96.4	97.8
ACT	5	94.2	95.2	94.7	94.5	94.5	96.1
NT	5	84.4	85.7	85.8	84.6	84.6	89.3
Aus.	**5**	**95.1**	**95.9**	**95.6**	**95.2**	**95.2**	**96.7**
NSW	7	95.7	96.5	96.6	96.0	96.0	97.9
Vic.	7	95.2	95.8	96.1	95.2	95.2	97.5
Qld	7	88.1	89.6	90.3	88.6	88.6	92.2
WA	7	94.8	96.0	96.0	95.2	95.2	97.9
SA	7	93.6	94.8	94.7	93.8	93.8	96.4
Tas.	7	94.2	95.2	95.1	94.1	94.1	97.2
ACT	7	94.2	94.9	94.8	94.1	94.1	96.7
NT	7	77.5	80.0	81.2	78.4	78.4	85.8
Aus.	**7**	**93.4**	**94.5**	**94.7**	**93.7**	**93.7**	**96.3**
NSW	9	92.1	93.2	93.3	92.4	92.4	95.3
Vic.	9	91.1	92.2	92.4	91.2	91.2	94.7
Qld	9	80.0	81.8	82.8	80.7	80.7	85.5
WA	9	92.3	93.3	93.4	92.1	92.1	95.5
SA	9	89.0	90.4	90.4	89.4	89.4	92.9
Tas.	9	88.6	89.8	90.5	88.7	88.7	93.5
ACT	9	88.4	89.6	90.0	88.7	88.7	92.4
NT	9	69.0	70.9	72.8	70.7	70.7	78.5
Aus.	**9**	**88.6**	**89.9**	**90.3**	**88.9**	**88.9**	**92.6**

Chapter 5: Scaling methodology and outcomes

This chapter describes the processes and methodologies used in the NAPLAN 2024 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 2024 tests were undertaken with the Rasch model (Rasch 1960), 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:

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

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:

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

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. τij is a step parameter of score *j* on item *i*. 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 Version 5* (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. 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 path with no trailing missing responses. In total, the number of students included in the estimation of each domain was between 184,734 and 230,497 by year level.

For the 2024 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 2024, there was only a small number of students who participated in NAPLAN paper tests. 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 2024 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 were 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 rubric scores were not observed for some year levels. Writing is calibrated only when all jurisdictions have completed marking; effectively, the whole population is available for calibration[6].

In the estimation of parameters, only students with complete test paths were included in the non-writing calibration data. Students with an incomplete test path 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, NSW.

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 2024 tests are given in Appendix C. 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.

The Rasch item parameter estimates and fit statistics are summarised in Appendix D 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

[6] Data from one jurisdiction was excluded from the reported 2024 writing calibration results and it was subsequently verified that this had no material impact on the comparability of the 2024 parameters with those from 2023.

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 E. 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 F. These show a comparison of the observed and the modelled expected score curve for each criterion.

Test reliability

Table 57 shows the IRT-based weighted reliabilities, calculated using weighted likelihood estimates (WLEs) or plausible values (EAP/PV) for each test.

The WLE reliability coefficients were between 0.90 and 0.94 for the numeracy tests, between 0.90 and 0.92 for the reading tests, between 0.90 and 0.93 for the spelling tests, and between 0.83 and 0.86 for the grammar and punctuation tests. The EAP/PV reliabilities were very similar to the WLE reliabilities: between 0.90 and 0.94 for the numeracy tests, between 0.90 and 0.92 for the reading tests, between 0.90 and 0.94 for the spelling tests, and between 0.82 and 0.86 for the grammar and punctuation tests. The reliabilities for the writing test were 0.96 for both WLE reliability and EAP/PV reliability.

Table 47. Reliability (EAP/PV, WLE) for NAPLAN 2024 tests

	Numeracy		Reading		Spelling		Grammar and punctuation		Writing*	
	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV
Year 3	0.90	0.90	0.92	0.92	0.93	0.94	0.86	0.86	0.96	0.96
Year 5	0.92	0.92	0.90	0.90	0.93	0.93	0.83	0.82		
Year 7	0.94	0.94	0.92	0.92	0.92	0.92	0.85	0.84		
Year 9	0.93	0.94	0.91	0.91	0.90	0.90	0.86	0.86		

*Concurrent Years 3, 5, 7 and 9 with data from 7 of the 8 jurisdictions

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 or thresholds (on the right side of the map). Item, step and person parameters are plotted on a common scale on a map. Appendix G 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 282 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* or *lower* 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)

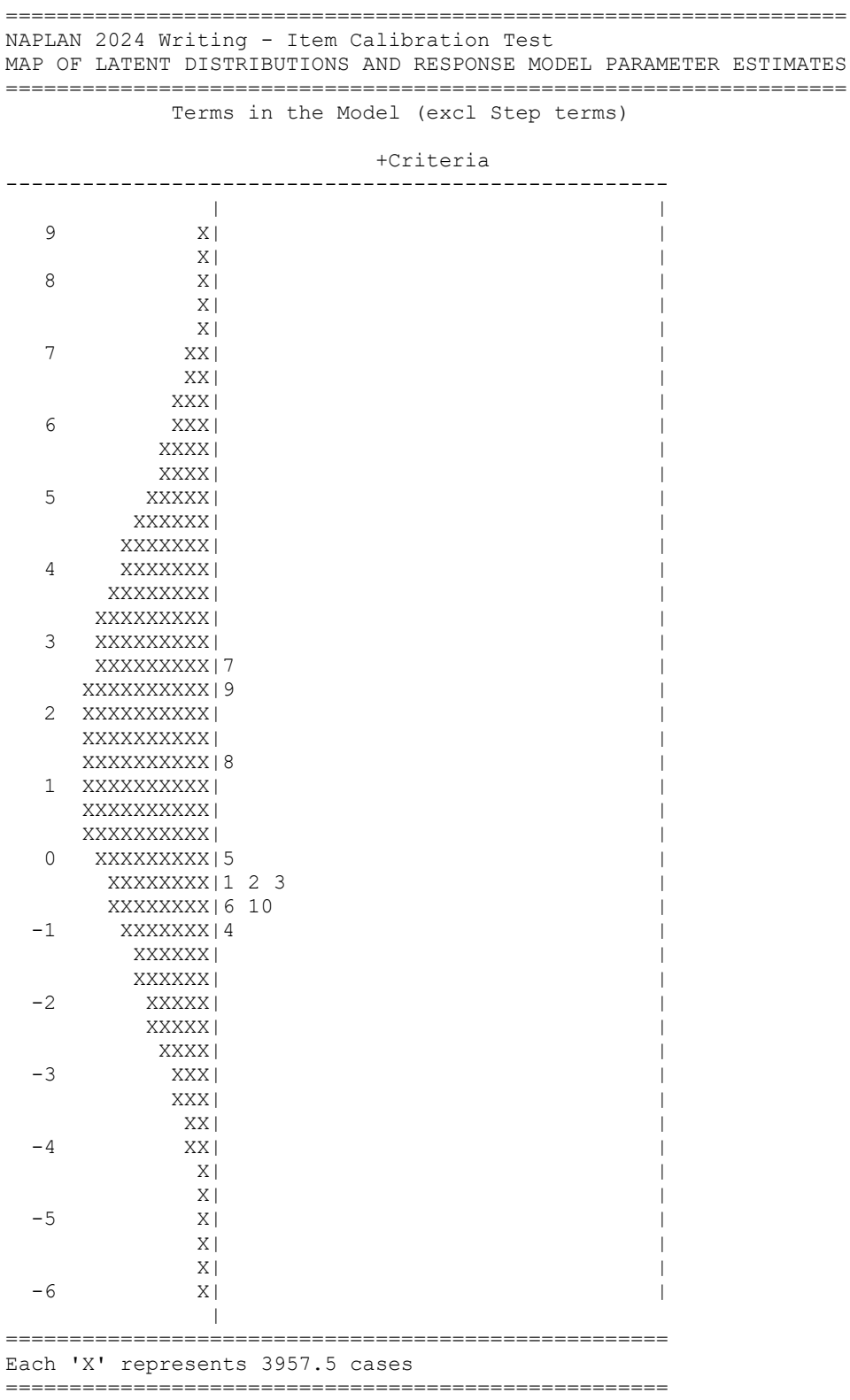


Figure 11. Wright map for writing test (a polytomous example)

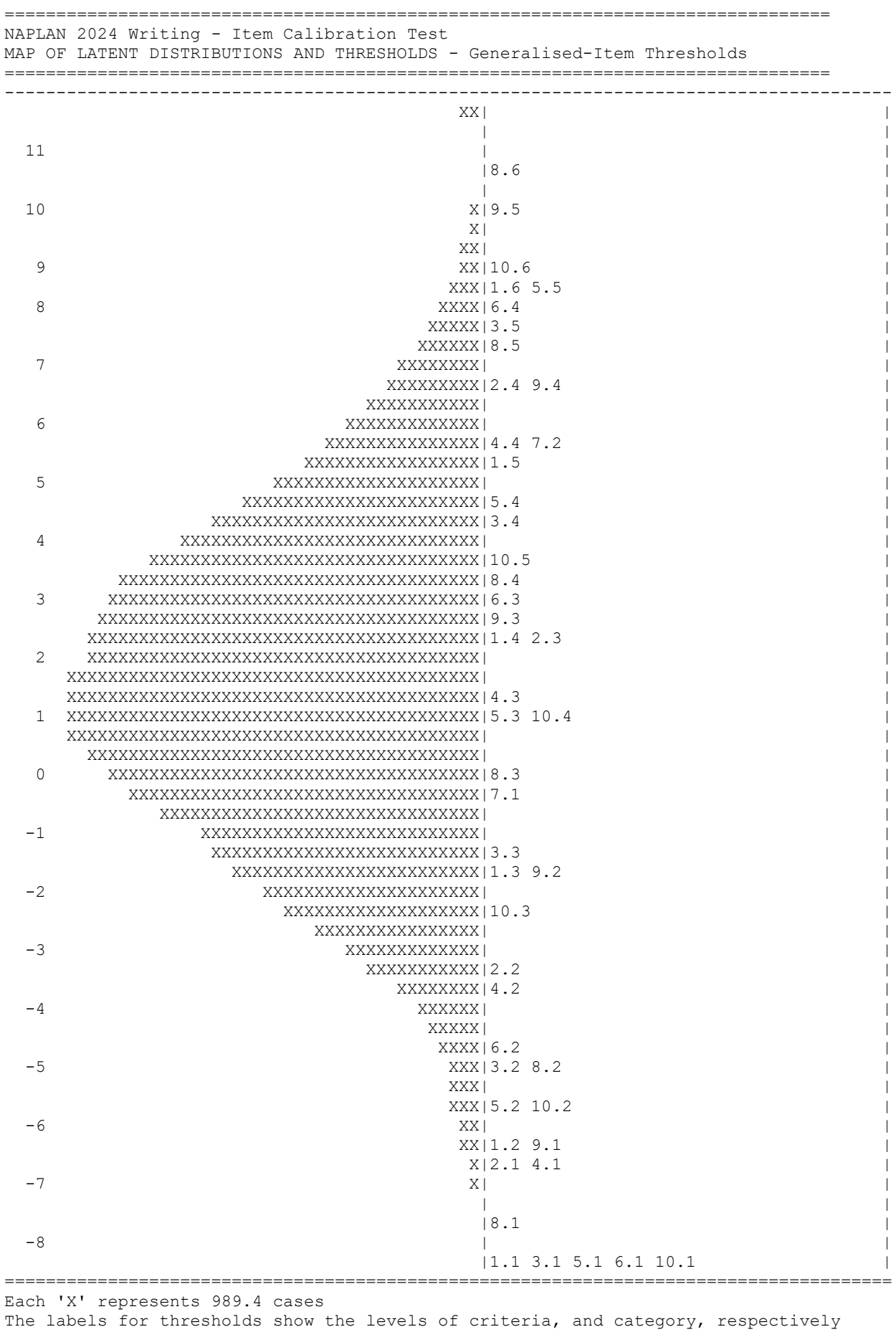


Figure 12. Thurstonian thresholds for writing test

Item fit

The evaluation of goodness of fit to the Rasch model for individual items was based on the weighted mean square (infit mean square) statistics. Infit compares the observed residual variance with the expected residual variance if the data fit the model. Infit mean square is an IRT-based index for the degree to which an item discriminates between low- and high-achieving students. Values larger than one indicate low discrimination (or flatter ICC slope than expected) and values smaller than one 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 C.

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 E.

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 58 presents summaries of item statistics in the NAPLAN 2024 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 58, 42 out of 3,252 items from 16 non-writing online tests had infit greater than 1.20. There were 81 items with facility higher than 0.90 and 39 items with facility less than 0.10. Figure 13 shows the ICC of one numeracy Year 3 item (item x00167153) with an infit statistic equal to 1.00. In contrast, Figure 14 shows the ICC of one Year 3 reading item (item x000170099) with an infit statistic (1.35) 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 D 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 1.45 and 1.64 respectively. None of the other criteria exhibited misfit to the Rasch partial credit model. Inspection of the ICCs did not reveal large differences between the empirical and the expected curves for any of the 10 criteria; with small discrepancies visible for the criteria with the highest infit (punctuation). The ICCs of the 10 writing criteria for writing are included in Appendix E.

Table 48. Summary of item statistics in NAPLAN 2024 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	4	0
	5	252	3	4	0
	7	288	4	5	2
	9	288	8	2	1
Reading	3	234	1	1	1
	5	234	1	5	0
	7	277	4	6	0
	9	284	0	7	2
Spelling	3	129	6	11	6
	5	129	2	5	9
	7	129	2	10	8
	9	129	3	6	9
Grammar and punctuation	3	162	3	2	0
	5	162	0	4	0
	7	162	2	5	0
	9	162	1	4	1
Writing	3, 5, 7 & 9	10*	2	N/A	N/A

* Item in Writing is criterion.

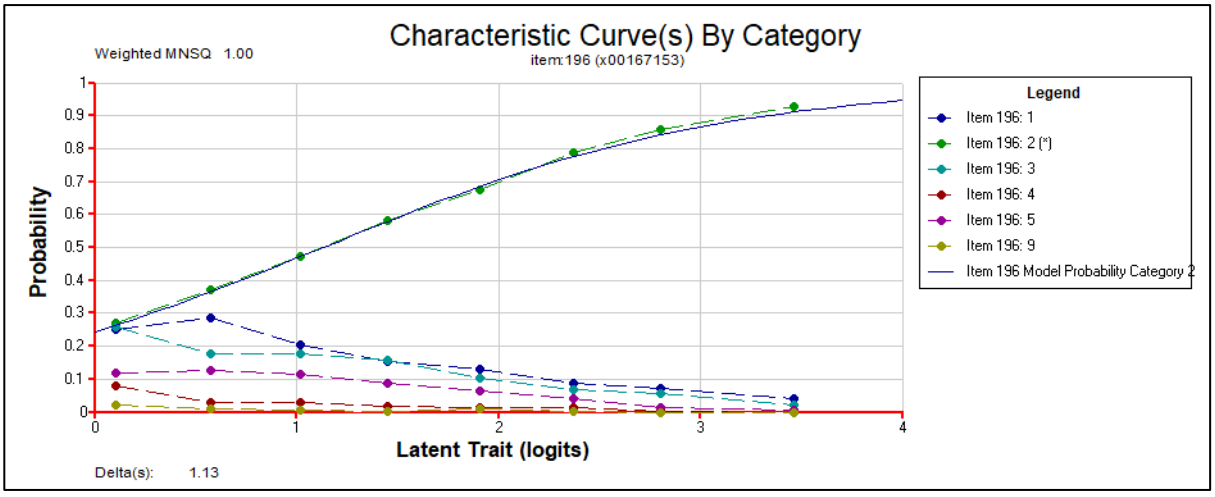


Figure 13. Item characteristic curves for an item with infit = 1.00

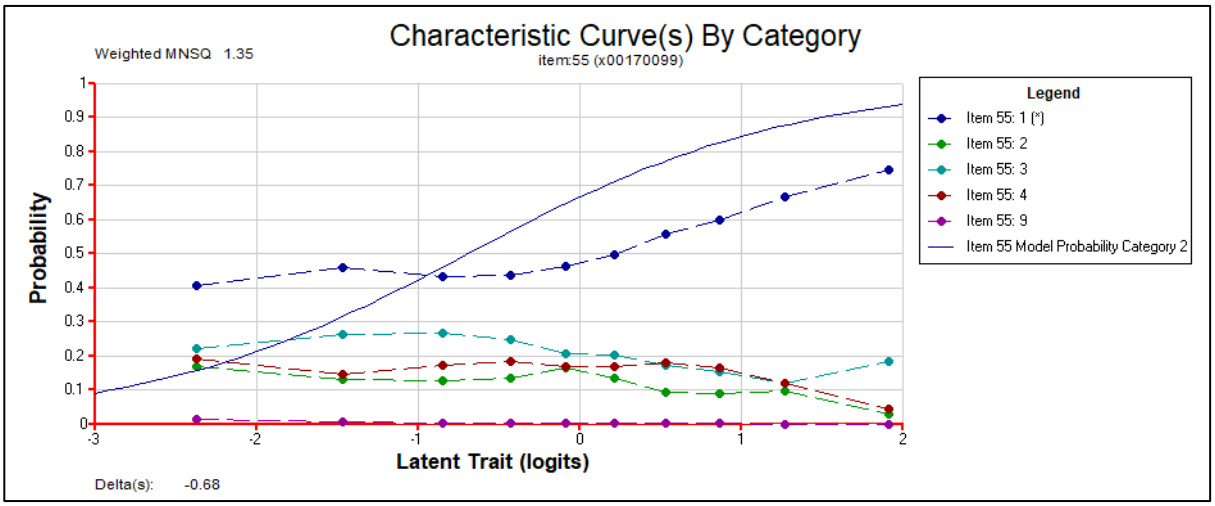


Figure 14. Item characteristic curves for an item with infit = 1.35

Differential item functioning (DIF) analyses

The functioning of the items was also evaluated through various differential item functioning (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[7], 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, $P(\theta)$, of a correct response differs for two groups. (Camilli and Shepard 1994)

In the analysis for NAPLAN, subgroups were arbitrarily categorised as either reference or focal groups. While male students, LBOTE students and Indigenous students were assigned to the reference group, female students, 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, male students versus female students). 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 H, I and J (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.

[7] 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, relative to minimum sample size thresholds in DIF literature, the analysis of gender DIF was limited to comparisons between male and female students.

The distance of a point from the diagonal reflects the magnitude of DIF. Due to the large sample sizes, confidence bands were very narrow.

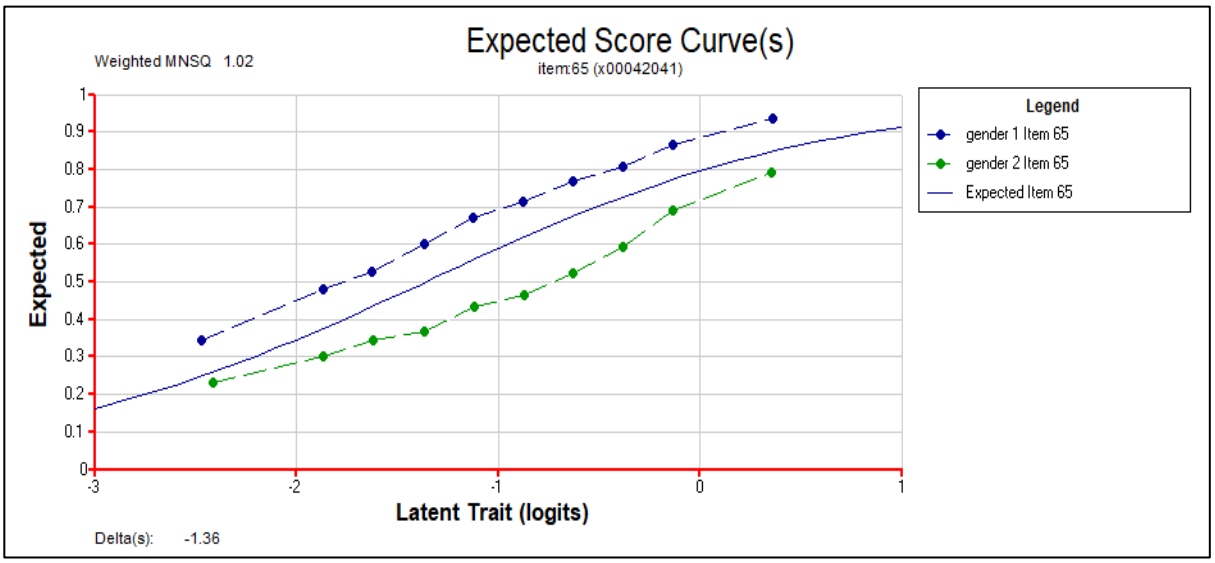
Gender DIF

Appendix H 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 60 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[8]. Figure 15 depicts an example of an item that displayed gender DIF. Appendix H includes DIF plots that show for each of the items the observed curves by gender group compared with the expected ICC.

Table 49. Number of items showing gender DIF by domain by year level

	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
Year 3	5/216	0/234	1/129	0/162	0/10
Year 5	5/252	0/234	2/129	0/162	
Year 7	7/288	1/277	5/129	0/162	
Year 9	2/288	2/284	10/129	0/162	



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

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

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

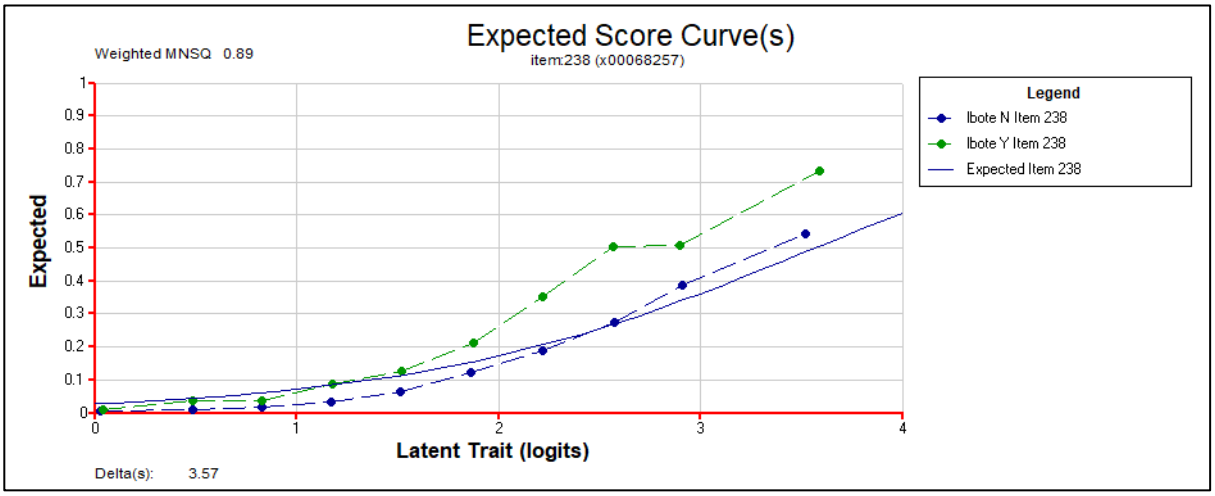
Language background DIF

Appendix I 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 61 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 29 depicts an example of an item that displayed LBOTE DIF.

Table 50. Number of items showing LBOTE DIF by domain by year level

	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
Year 3	1/216	0/234	1/129	1/162	0/10
Year 5	3/252	0/234	0/129	0/162	
Year 7	1/288	0/277	1/129	2/162	
Year 9	3/288	0/284	1/129	2/162	



† "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 J 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 62 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 an example of an item that displayed Indigenous DIF.

Appendix J provides the item DIF plots for items listed in Table 62. 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 51. Number of items showing Indigenous DIF by domain by year level

	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
Year 3	2/216	0/234	1/129	1/162	0/10
Year 5	2/252	0/234	0/129	2/162	
Year 7	2/288	1/277	0/129	0/162	
Year 9	6/288	0/284	1/129	2/162	

† "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 (see below), are presented in Appendix K.

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 63. In the headings of Table 63, "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 63 can be read in conjunction with Appendix L, which contains item DIF plots for items showing jurisdictional DIF for

items listed in Table 63. The plots show, for each of these items, the observed curves by state/territory compared with the expected ICC. Figure 18 depicts one Year 9 numeracy test item (item x000154356) showing jurisdictional DIF. This item was relatively hard for NT students.

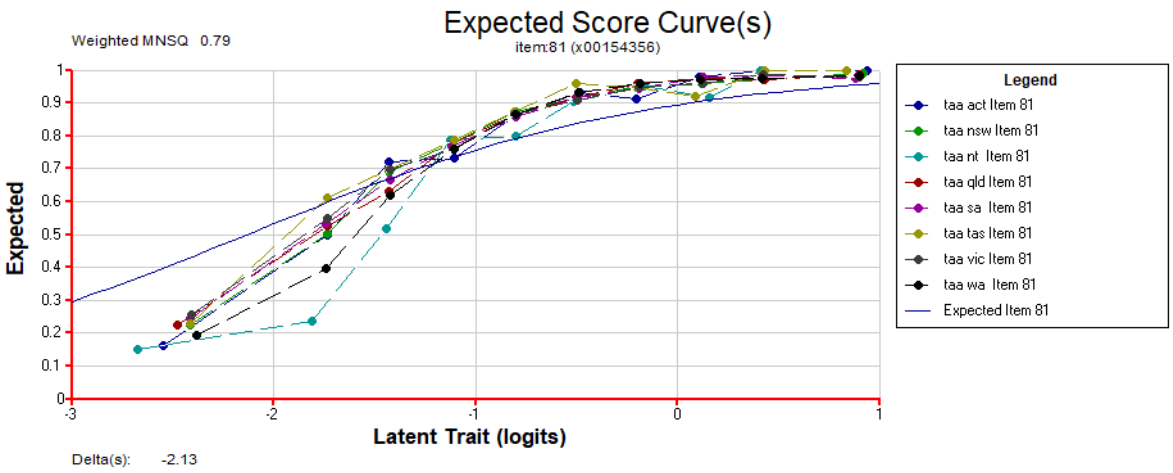


Figure 18. Example of item characteristic curves displaying jurisdictional DIF

Table 52. 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	1	1	-	6	7	-	-	-	-	1	2	-	-	1	-
	5	6	1	-	2	9	7	2	-	1	-	3	8	-	-	-	-
	7	3	-	3	1	9	12	-	-	-	-	1	5	2	-	1	-
	9	3	2	1	1	23	18	-	-	-	-	1	3	1	-	9	2
Reading	3	1	-	-	-	3	8	-	-	-	-	2	-	-	-	-	-
	5	2	-	-	-	5	7	-	-	-	-	1	-	-	-	-	-
	7	2	1	2	-	4	12	-	-	-	-	1	-	1	-	-	-
	9	1	-	-	-	9	13	-	-	-	-	-	1	-	-	6	1
Spelling	3	-	2	-	-	2	2	-	-	1	1	4	2	-	1	1	1
	5	4	1	-	-	5	7	-	-	2	1	3	5	1	1	1	1
	7	2	1	-	-	4	8	-	-	-	-	2	4	1	-	2	-
	9	2	3	-	-	4	6	-	-	-	-	5	1	-	-	2	-
Grammar and punctuation	3	-	1	4	-	4	2	1	-	-	-	4	5	-	-	-	1
	5	9	2	-	-	9	12	2	-	1	-	1	3	2	1	-	1
	7	-	1	-	-	4	10	3	-	-	-	3	2	-	-	-	-
	9	2	4	1	-	5	2	3	-	-	-	-	1	-	-	4	1
Writing	3, 5, 7 & 9	1	-	1	-	-	2	N/A	N/A	-	-	-	2	-	1	-	-

Note 1. "E" indicates that the item is relatively easy for the jurisdiction, and "H" indicates that the item is relatively hard for the jurisdiction.

Note 2. Results for 7 of the 8 jurisdictions were included in the reported writing calculation.

Device DIF

For online tests, a device DIF analysis was also carried out for each domain, as there were different devices used by different students. Writing device DIF was investigated for Years 5, 7 and 9, but not for Year 3 as all Year 3 students completed the writing test on paper. 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 64 shows the number of students using each device type at each grade 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 65. Table 65 shows that only Mac devices had any items demonstrating DIF, mainly in numeracy and reading. Graphs showing device DIF by item are shown at Appendix M.

Table 53. Number of students by device

Domain	Year level	Chromebook	iOS	Mac	Windows
Numeracy	3	39,077	77,654	1,747	83,244
	5	39,707	54,325	5,427	98,529
	7	19,781	16,746	32,764	136,050
	9	17,682	14,266	36,156	127,319
Reading	3	41,932	82,576	2,122	94,374
	5	44,191	61,590	6,761	117,938
	7	20,920	17,831	35,953	154,840
	9	18,784	15,669	39,277	142,981
Spelling	3	36,838	72,016	1,669	77,069
	5	38,591	53,324	5,603	96,801
	7	19,295	16,525	32,686	134,801
	9	16,714	13,762	35,083	121,076
Grammar and punctuation	3	36,120	71,776	1,645	75,170
	5	38,741	53,586	5,619	97,235
	7	19,479	16,656	32,864	136,208
	9	16,918	13,978	35,457	122,893
Writing	5, 7 & 9	96,306	71,982	88,073	380,369

Table 54. 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	-	-	-	-	5	1	-	-
	5	-	-	-	-	4	1	-	-
	7	-	-	-	-	3	1	-	-
	9	-	-	-	-	2	-	-	-
Reading	3	-	-	-	-	-	-	-	-
	5	-	-	-	-	2	-	-	-
	7	-	-	-	-	5	-	-	-
	9	-	-	-	-	1	-	-	-
Spelling	3	-	-	-	-	1	1	-	-
	5	-	-	-	-	1	-	-	-
	7	-	-	-	-	-	-	-	-
	9	-	-	-	-	1	-	-	-
Grammar and punctuation	3	-	-	-	-	1	-	-	-
	5	-	-	-	-	3	-	-	-
	7	-	-	-	-	-	-	-	-
	9	-	-	-	-	1	-	-	-
Writing	5, 7 & 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 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 and 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 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.

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

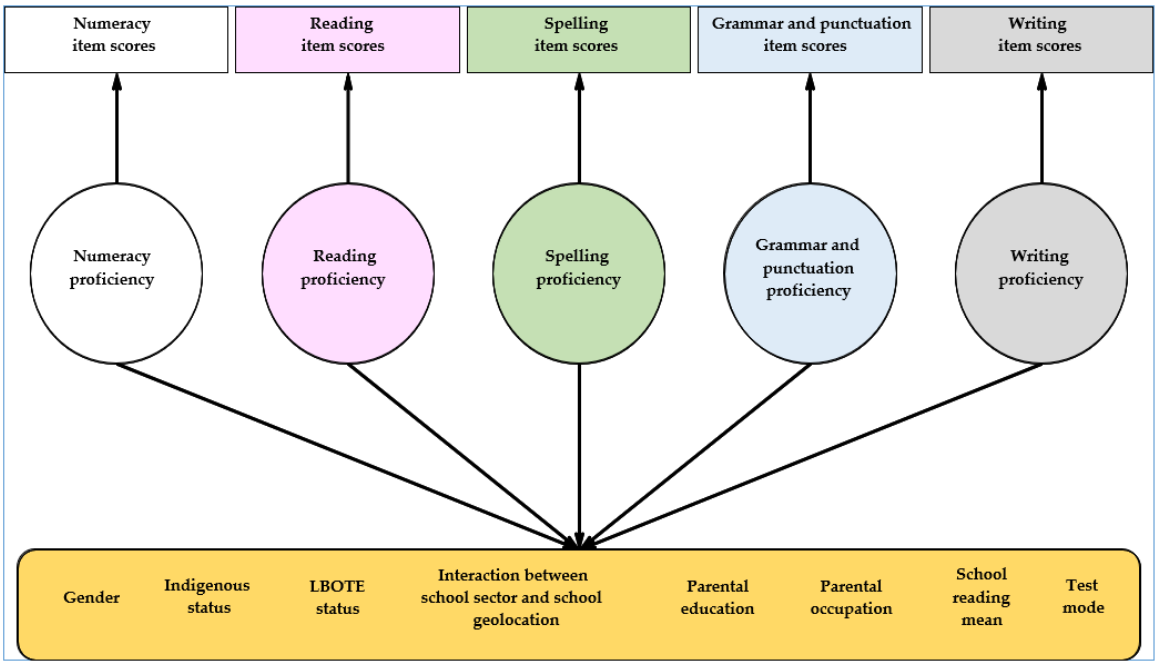


Figure 19. Conditioning variables for the multidimensional item response model with latent regression

Chapter 6: Equating procedures

In 2023, the NAPLAN scales were reset. This chapter describes the process of equating the 2024 tests to the reset NAPLAN scales. The first section describes the equating procedures for each of the 2024 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

The NAPLAN scales in each assessment domain were reset in 2023 because of the full transition to the adaptive online assessment and the change in testing window from May to March. In each domain, the new reporting scales were established by placing all year levels onto the same scale, using vertical link items. In 2024, in order to monitor student achievement over time, the scales were horizontally equated to the 2023 scales.

Each of the 2024 reading, spelling, grammar and punctuation, and numeracy tests across Years 3, 5, 7 and 9 contained a large number of items that were also administered in the 2023 tests. This allowed the application of common-item equating to place the NAPLAN 2024 results onto the reset NAPLAN 2023 scale. The 2024 tests were first placed onto the NAPLAN 2023 delta-centred scales and were then transformed onto to the 2023 reset NAPLAN scale by applying the shifts and transformations that were used in 2023 to set the new NAPLAN scale.

In addition, vertical link items were embedded in tests at adjacent year levels: Years 3 and 5, 5 and 7, or 7 and 9. In 2024, vertical equating was used as a quality assurance procedure to verify the horizontal shifts (rather than as the primary method for establishment of the scales, as was the case in 2023). This verification was conducted through horizontal–vertical regression (HVR). The HVR analysis established that the scales as calibrated in 2024 through horizontal equating were consistent with the scales that would have been developed through vertical equating, as they were in 2023 when they were initially established. Further details on the HVR methodology can be found in previous years' technical reports.

Additionally, all items included in the 2024 paper tests were also embedded in the 2024 online tests. This allowed the online test item parameters to be used for all paper items.

For each of the 4 domains, before calculating the horizontal equating shifts, the quality of the common items was systematically reviewed for their functioning as equating links. Only items that showed satisfactory and similar psychometric properties in 2023 and 2024 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 horizontal link items was undertaken as follows:

- Initial cross-test-form scatter plots 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 cross-test-form DIF, difficulties of the set of common items were centred on zero for each test form. For each pair of linked tests, one set of item difficulties (for example, of 2024 Year 3 link items) was then plotted against the other set of item difficulties (of 2023 Year 3 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 2 steps:
    1. Outliers (items with an absolute difference larger than 0.9 of a logit between their relative difficulties) were broken, and the process was repeated if necessary.
    2. Any other items with an absolute difference of more than 0.4 logits between their relative difficulties were broken in the second step, and the process was repeated if necessary.

- 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, node (A, B, C, D, E or F), item facility, (average) position of the item in the pathway, infit MNSQ and gender DIF are compared between the 2 linked tests. While items are not individually excluded based on these criteria, the link sets are audited to ensure that they have similar specifications: to each other, and to the entire set of items within each test.

Each 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, 2024 Year 3 SD / 2023 Year 3 SD). The ideal ratios between equated tests should be 1.00. Ratios that fall between 0.9 and 1.1 are considered to be of very high quality. The actual ratios for 2024 were between 0.94 and 1.04.

After the review and evaluation of the equating items between the 2024 and 2023 tests, a final set of link items was identified for each domain and year level. The final sets of link items were used to calculate the preliminary horizontal shifts from 2024 to 2023.

The outcome of the review of horizontal link items is summarised in Figure 20 to Figure 35: all common items shown in the left-hand graphs, and the final link sets shown in the right-hand graphs. These plots show the comparisons of item difficulty estimates of each link set between 2024 and 2023 for each year level 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 scatter plot.

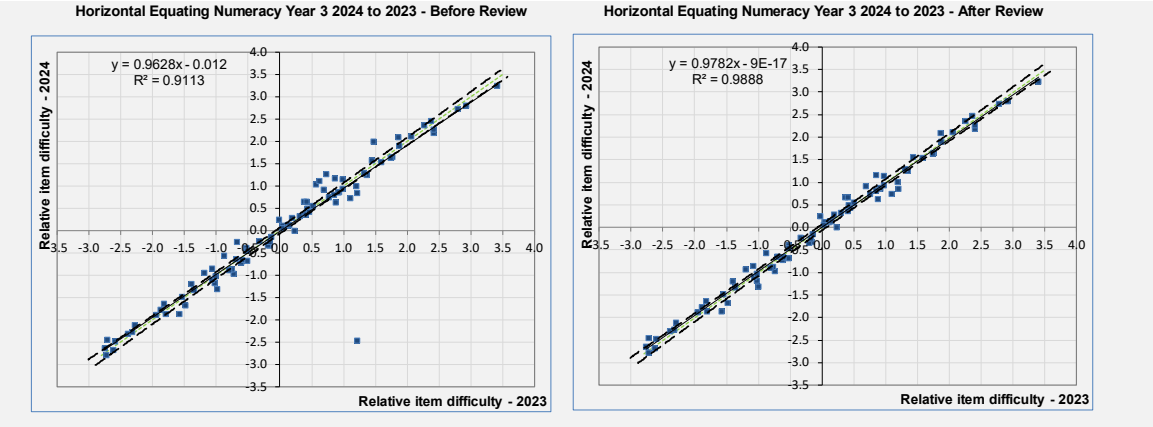


Figure 20. Scatter plot of numeracy, horizontal equating items between 2024 and 2023 for Year 3 students

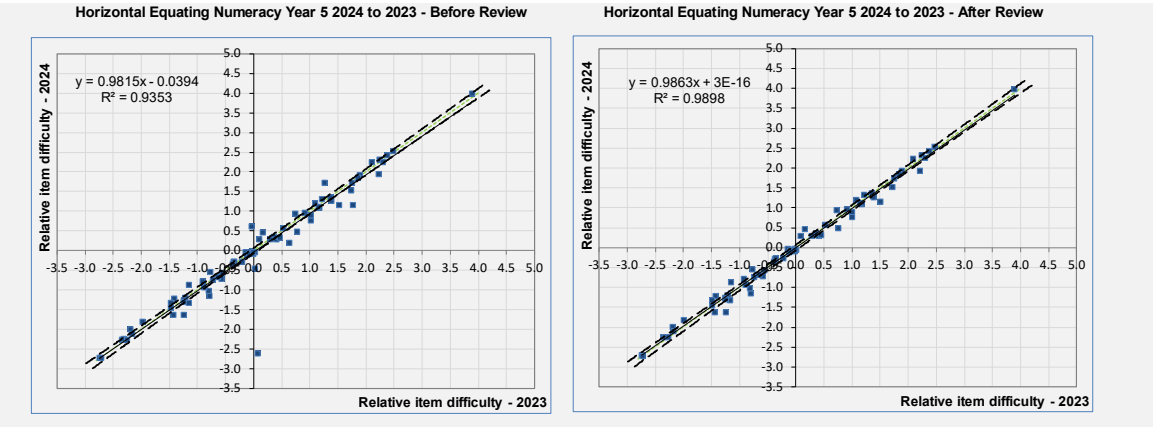


Figure 21. Scatter plot of numeracy, horizontal equating items between 2024 and 2023 for Year 5 students

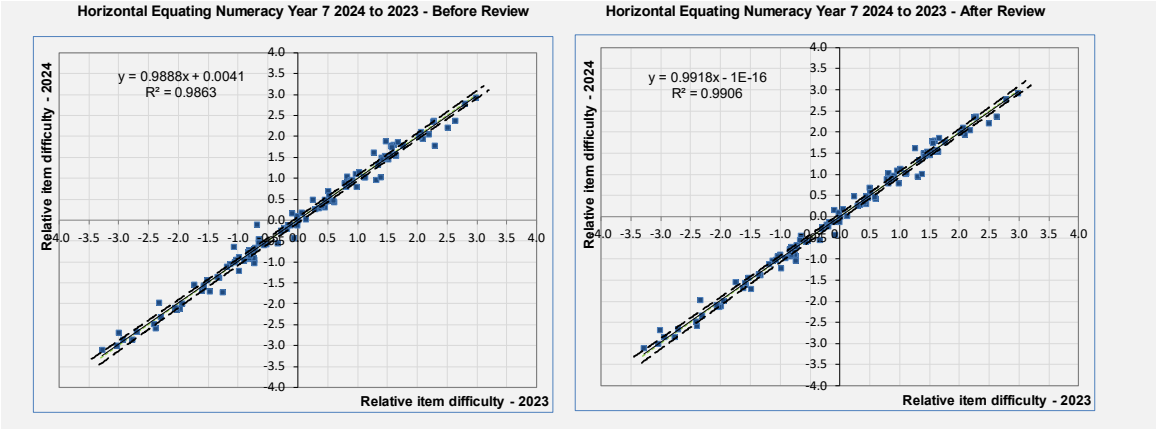


Figure 22. Scatter plot of numeracy, horizontal equating items between 2024 and 2023 for Year 7 students

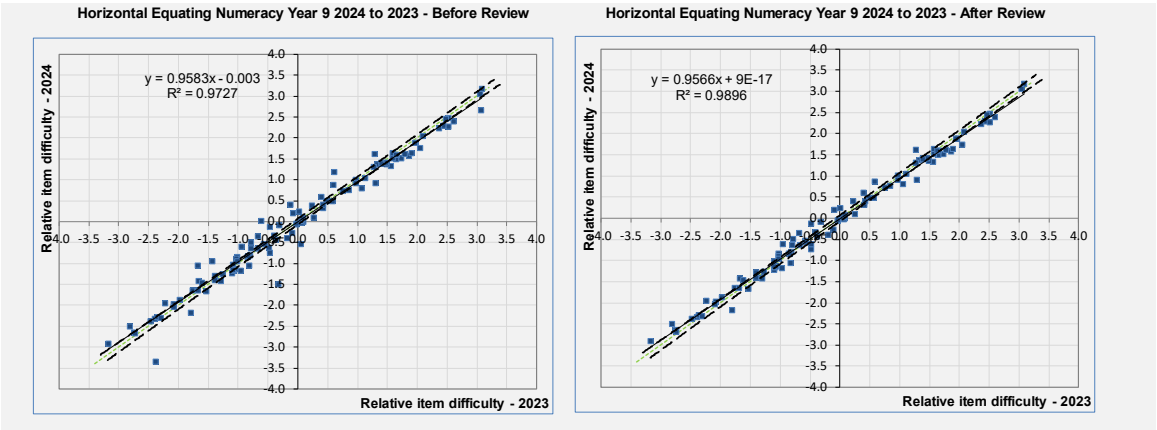


Figure 23. Scatter plot of numeracy, horizontal equating items between 2024 and 2023 for Year 9 students

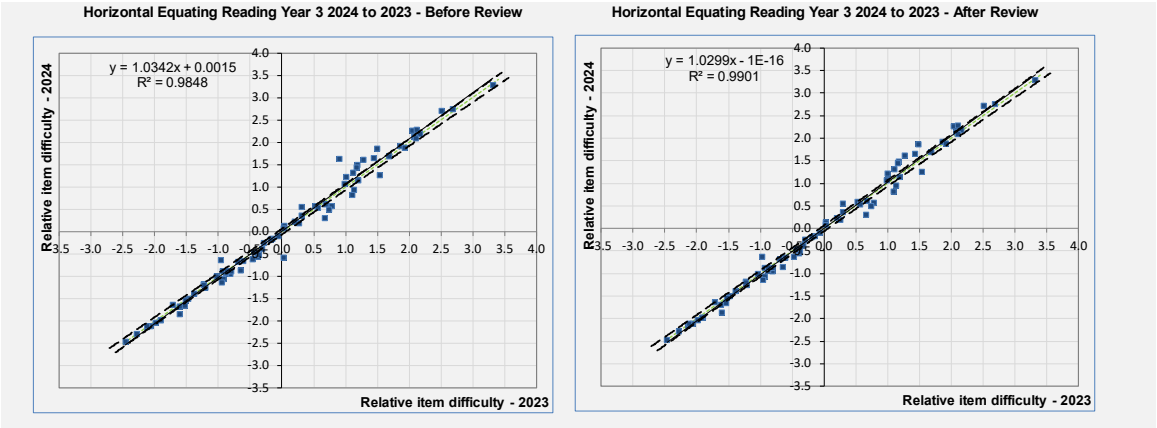


Figure 24. Scatter plot of reading, horizontal equating items between 2024 and 2023 for Year 3 students

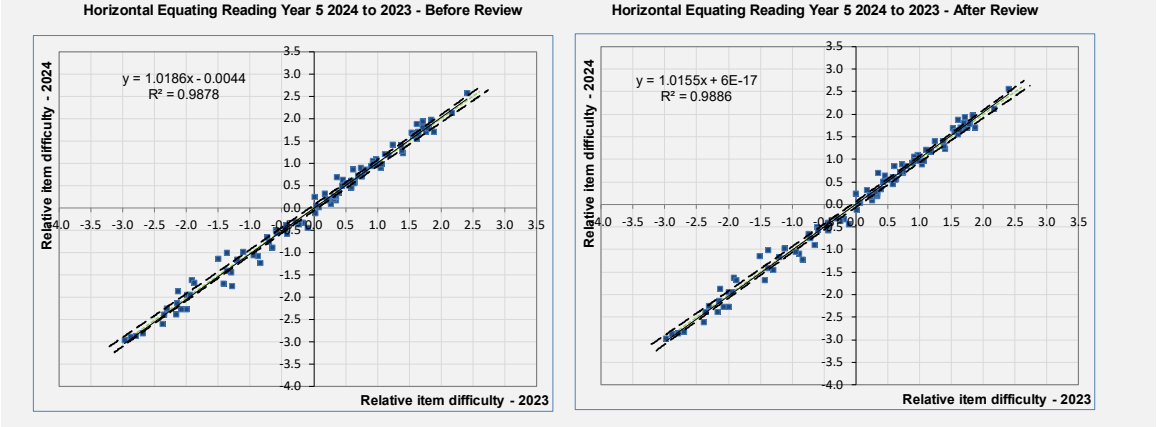


Figure 25. Scatter plot of reading, horizontal equating items between 2024 and 2023 for Year 5 students

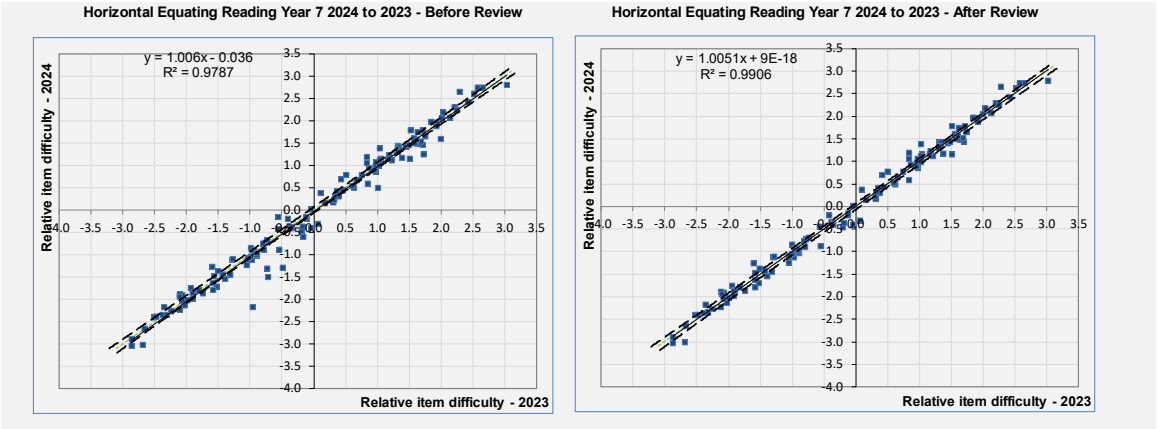


Figure 26. Scatter plot of reading, horizontal equating items between 2024 and 2023 for Year 7 students

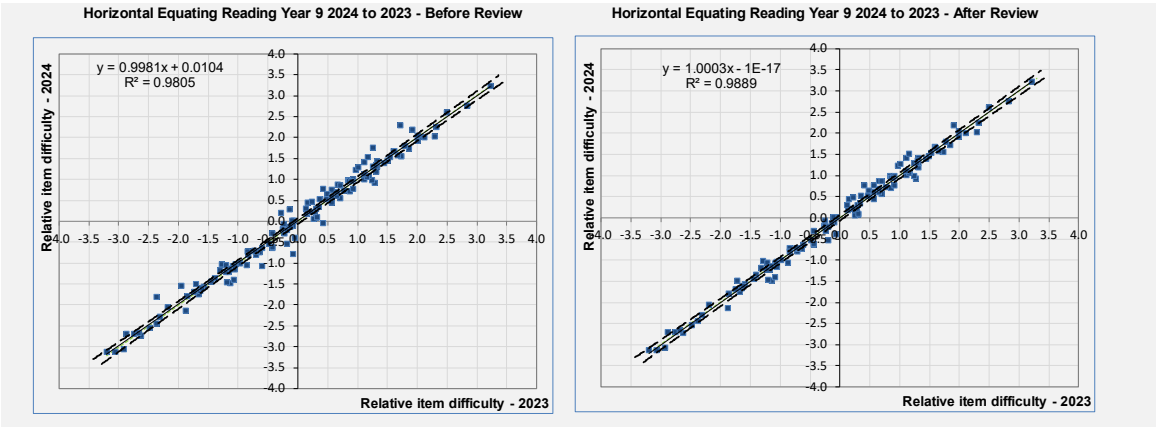


Figure 27. Scatter plot of reading, horizontal equating items between 2024 and 2023 for Year 9 students

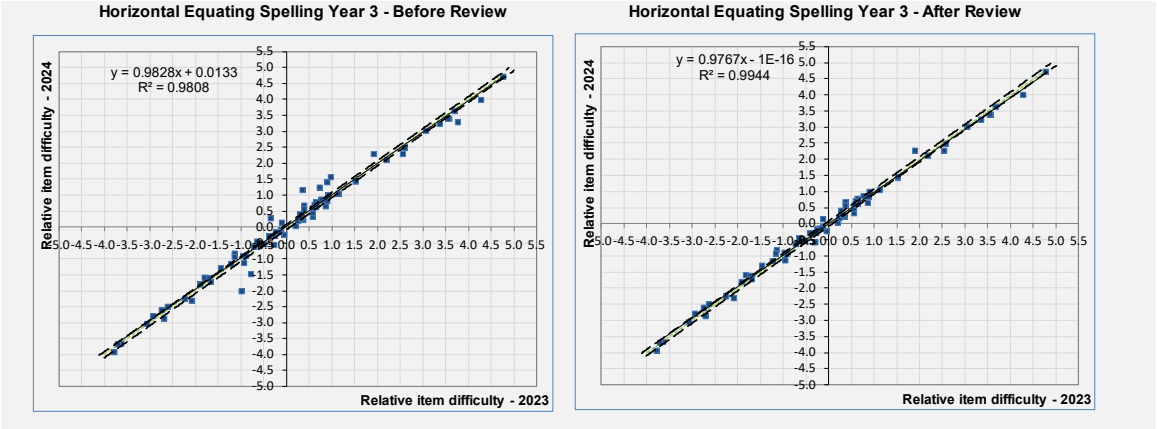


Figure 28. Scatter plot of spelling, horizontal equating items between 2024 and 2023 for Year 3 students

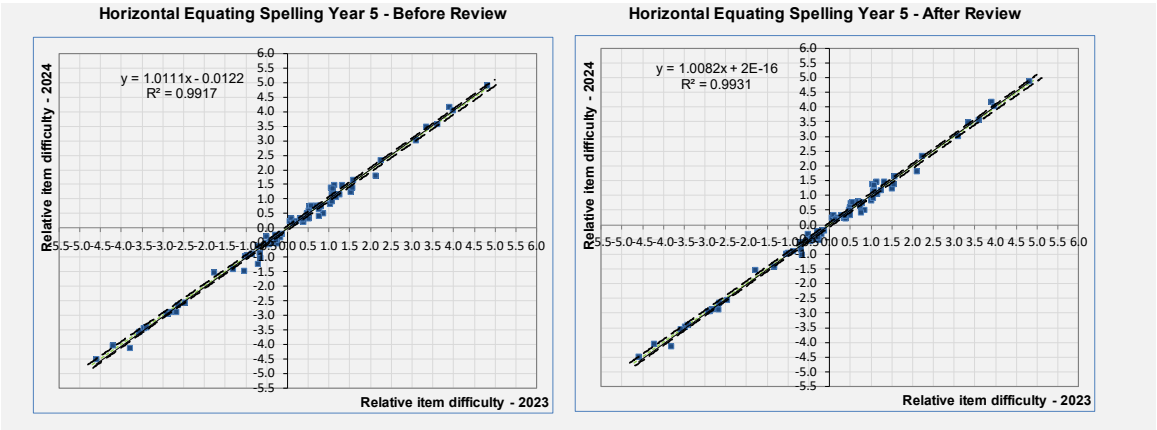


Figure 29. Scatter plot of spelling, horizontal equating items between 2024 and 2023 for Year 5 students

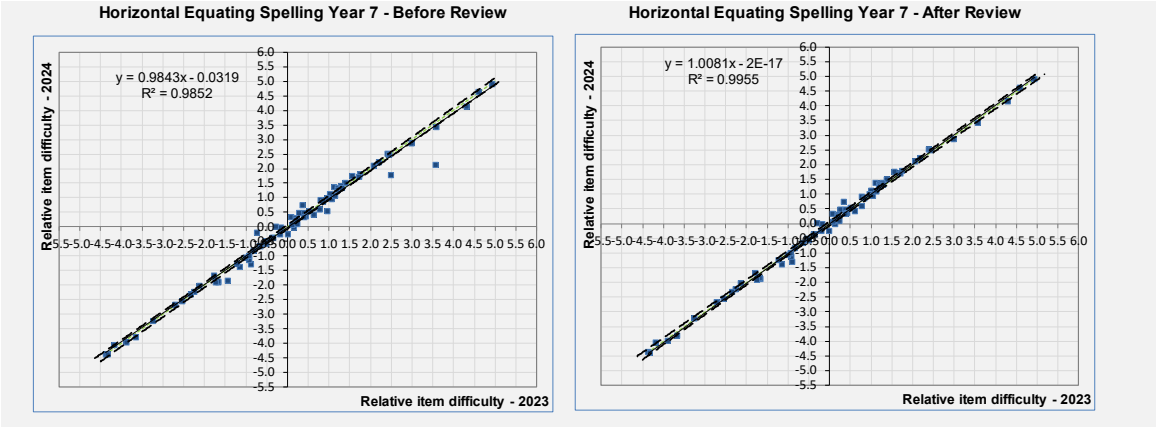


Figure 30. Scatter plot of spelling, horizontal equating items between 2024 and 2023 for Year 7 students

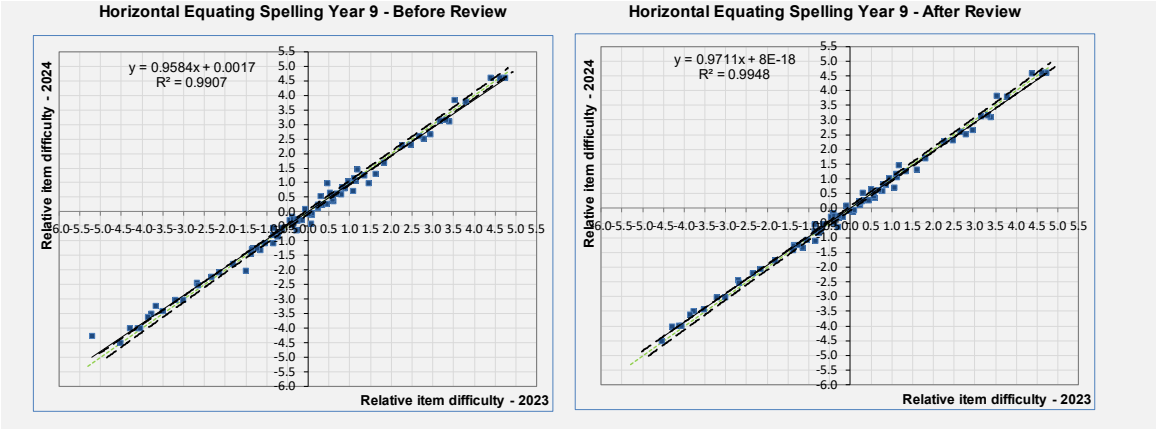


Figure 31. Scatter plot of spelling, horizontal equating items between 2024 and 2023 for Year 9 students

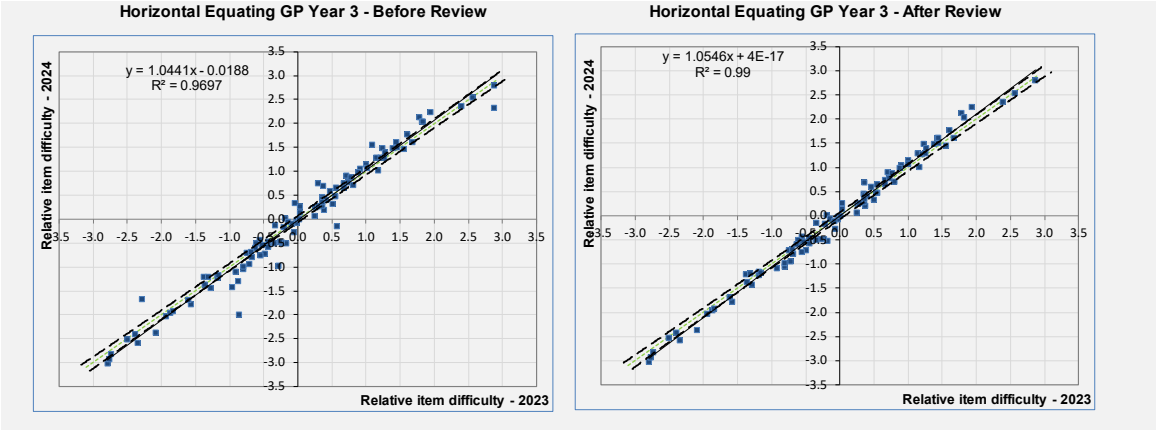


Figure 32. Scatter plot of grammar and punctuation horizontal equating items between 2024 and 2023 for Year 3 students

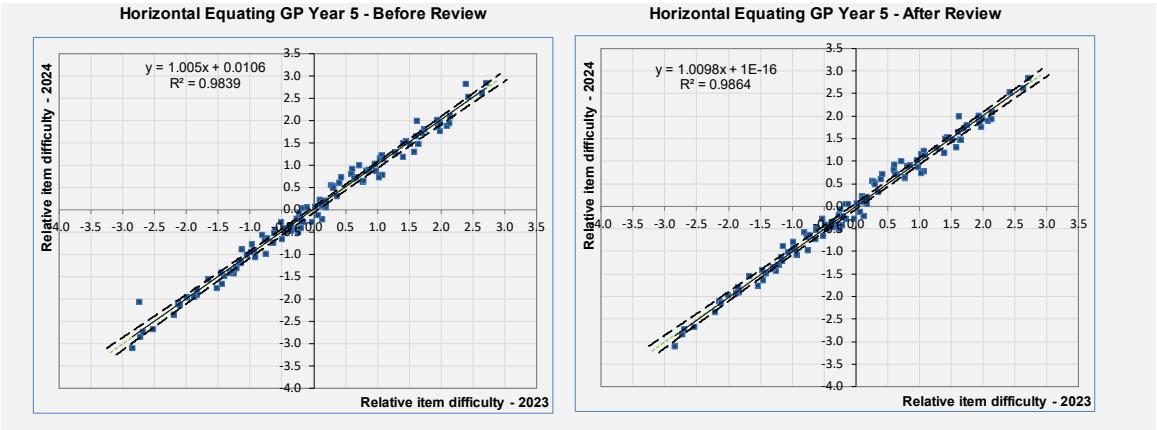


Figure 33. Scatter plot of grammar and punctuation, horizontal equating items between 2024 and 2023 for Year 5 students

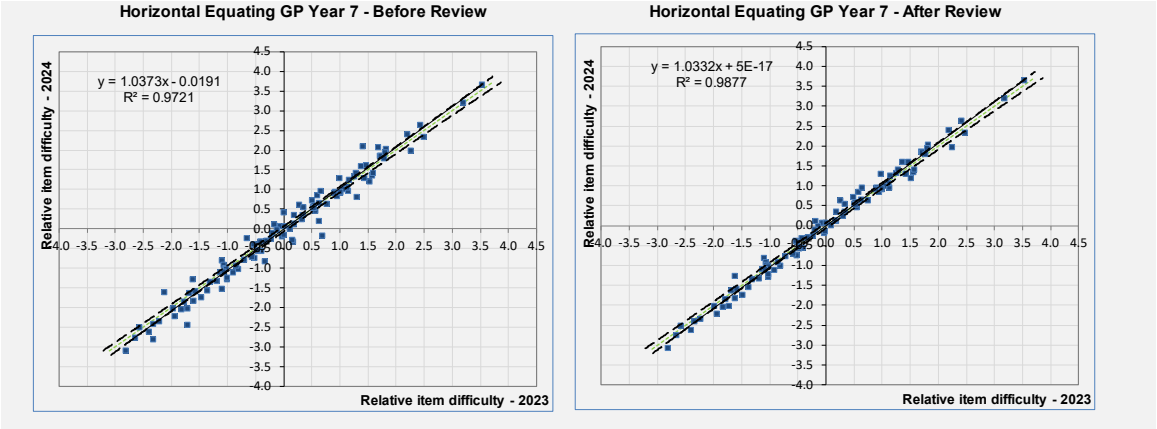


Figure 34. Scatter plot of grammar and punctuation, horizontal equating items between 2024 and 2023 for Year 7 students

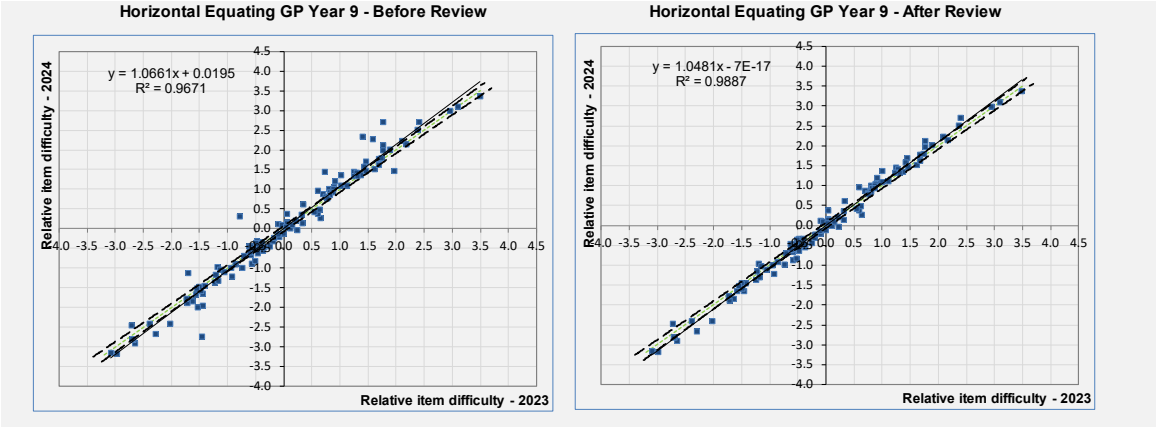


Figure 35. Scatter plot of grammar and punctuation, horizontal equating items between 2024 and 2023 for Year 9 students

The numbers of common items between 2023 and 2024 in the test design and the numbers retained as links for each test are shown in Table 55. The horizontal shift-constants for each domain at each year level are summarised in Table 56, and the final shifts to place 2024 tests onto the NAPLAN reporting scale are summarised in Table 57.

Appendix N presents the 2024 horizontal link item locations (Rasch difficulties), standard errors, and differences in the item locations by domain and year level. Appendix O contains the information for vertical links.

Table 55. Horizontal link review summary (Number of used links/Number of common items in test design)

	Numeracy	Reading	Spelling	Grammar and punctuation
Year 3	83/89	74/76	66/74	94/104
Year 5	69/75	96/97	72/74	105/107
Year 7	109/114	123/132	69/74	95/109
Year 9	103/112	124/133	75/81	104/114

Table 56. Horizontal equating shifts between 2024 and 2023 item locations and their associated equating errors by domain and year level

	Numeracy		Reading		Spelling		Grammar and punctuation	
	Shift	Error	Shift	Error	Shift	Error	Shift	Error
Year 3	-0.17672	0.01718	-0.26577	0.01756	-0.43161	0.01848	-0.20856	0.01543
Year 5	-0.23199	0.01767	-0.08820	0.01537	-0.13339	0.01932	0.10948	0.01531
Year 7	-0.01786	0.01379	-0.06491	0.01365	-0.13475	0.01662	-0.12676	0.01690
Year 9	-0.14542	0.01613	-0.04028	0.01337	-0.08081	0.01906	-0.06784	0.01662

Table 57. Final equating shifts applied for each test by year level by domain

	Numeracy	Reading	Spelling	Grammar and punctuation
Year 3	-1.51224	-1.28452	-2.55691	-1.00818
Year 5	-0.23199	-0.08820	-0.13339	0.10948
Year 7	0.63293	0.54149	1.19436	0.60980
Year 9	1.20764	1.00785	1.94106	0.79648

Equating of writing results

As described in Chapter 5, the writing data from all 4 year levels were concurrently calibrated to construct the vertical writing scale. Because this process placed the 4 year levels on the same scale, no separate vertical equating process was required.

To equate the writing test to the writing reporting scale that was established in 2023, the anchoring method was used. Before anchoring the item (criterion) difficulties to the 2023 parameters, the appropriateness of this method was assessed in 2 ways. First, the relative item difficulty steps were compared with those from 2023. Second, achievement drift caused by any systematic changes in marking over time was examined.

To review the stability of item difficulty steps, the freely calibrated 2024 writing data was compared to the item difficulties of the 2023 tests since the writing genre was narrative in both 2024 and 2023. The scatter plot between the 2 calendar years is shown in Figure 36. They indicate that the consistency of relative difficulties supported using the anchoring method in 2024.

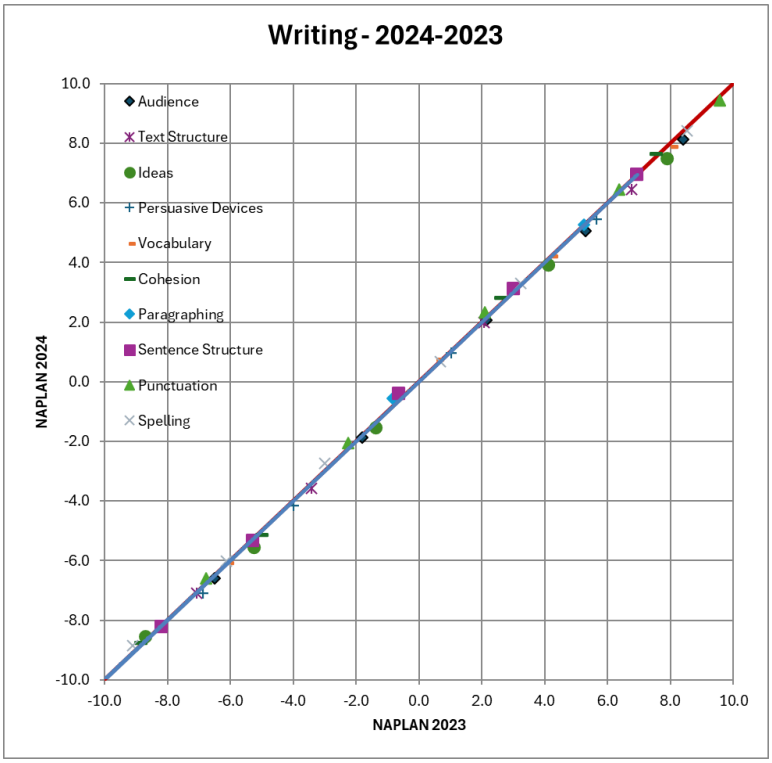


Figure 36. Scatter plot for writing criteria between 2024 and 2023 tests

In addition to comparing relative item difficulties, an equating verification study was conducted using pairwise comparisons of scripts in order to investigate if a shift in marking may have occurred.

Pairwise equating verification of writing

The purpose of the pairwise study is to ascertain whether rubric marks are consistent across calendar years using a pairwise scale as a common reference point. In particular, the objective is to examine whether there is evidence for changes in marker harshness or other changes that might affect the comparability of results.

Pairwise comparisons provide a direct means of ordering scripts. In this analysis, scripts from 2023 and 2024 were ordered and scaled together to form a common scale. The study then triangulated rubric locations (student ability estimates on the NAPLAN writing scale, obtained using a score equivalence table from the raw score) from 2023 and 2024 with pairwise locations on the combined 2023/2024 scale, based on a sample of scripts from several states. This allowed evaluation of whether, for a given scale location based on pairwise comparisons, a similar rubric location is predicted for 2023 and 2024 scripts.

Pairwise study design

The equating design involved pairwise comparisons of 299 writing responses from 2024 (230 online and 69 paper), and 300 writing responses from 2023 (230 online, 70 paper). Writing samples were obtained from all tasks administered to students, to minimise task effects. Scripts were selected using an approximately uniform score distribution in terms of total rubric scores.

All pairs of scripts were compared using 2 criteria: authorial choices and conventions. Markers judged which script is better on each of these 2 criteria.

For the 2024 pairwise equating project, 42 judges compared 31,220 pairs of scripts in total. Of these, there were 7,728 comparisons of 2023 against 2023 scripts, 15,816 comparisons of 2023 against 2024 scripts and 7,676 comparisons of 2024 against 2024 scripts. 774 comparisons were made between 2023 paper scripts and 2024 paper scripts.

Pairwise study results

The Bradley-Terry-Luce (BTL) model (Bradley and Terry 1952; Luce 1959) was used to analyse the data. To evaluate fit to the model, judge outfit indices were calculated after removing extreme observations (comparisons for which the standardised residuals were greater than 7). For the 2024 pairwise study, all judges had good outfit indices (less than 1.4). The person separation index was 0.986, indicating very high internal consistency of judgements overall. Most writing samples had acceptable outfit values, with only 16 of 599 exceeding values of 1.5.

Figure 37 shows the locations on the pairwise scale of the 2024 scripts from: (a) 2023 vs 2024 (*y*-axis) paired comparisons; and (b) 2024 vs 2024 paired comparisons (*x*-axis). Figure 36 shows a very strong linear correspondence between 2023 vs 2024 estimates and 2024 vs 2024 estimates, with points scattered very closely around the identity line, indicating very high comparability of the scales. The person separation index for 2023 vs 2024 estimates is 0.983 and for 2024 vs 2024 estimates is 0.973. The correlation between 2023 vs 2024 estimates and 2024 vs 2024 estimates is 0.929 and the disattenuated correlation is 0.979. The very high correlation shows that 2024 script locations are effectively the same whether based on paired comparisons between 2023 and 2024 scripts or paired comparisons between 2024 and 2024 scripts. Thus, the estimates are invariant whether they are obtained from comparisons within a single calendar year or comparisons of scripts from the different calendar years. As a result, the locations are robust over the 2 calendar years and support the consistency of marking across those years.

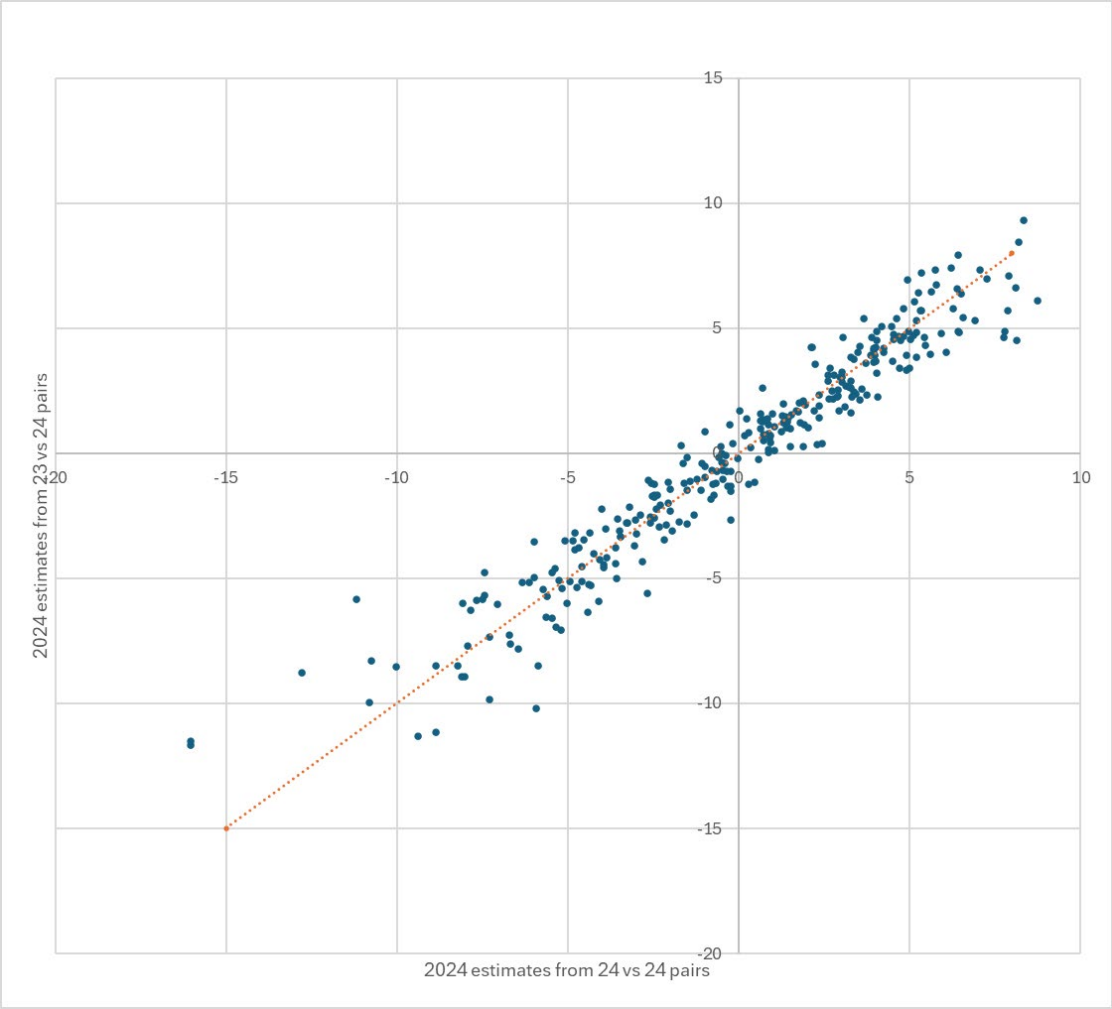


Figure 37. Pairwise locations for 2024 scripts from 2024 vs 2024 pairs and 2024 vs 2023 pairs.

Figure 38 shows the pairwise scale locations (*x*-axis) plotted against the NAPLAN rubric locations (*y*-axis) for the 2023 and 2024 scripts, across all year levels and all writing tasks. The pairwise scale locations show the ordering of the scripts based on direct comparisons whereas the NAPLAN scale locations are based on rubric marking.

The fitted curves in Figure 38 are somewhat curvilinear, and show a very close relationship between the 2023 scripts and the 2024 scripts. The overall correlation between the pairwise and rubric locations is 0.954 based on a polynomial regression model that allows for the curvilinear relationship. The close agreement of both fitted curves and data points for the 2023 data and 2024 data provide evidence that marking in 2023 was highly consistent with marking in 2024.

The correlation and nature of the relationship are similar for both calendar years to the relationship observed in previous calendar years of NAPLAN. The correlation between pairwise location and rubric location for 2024 is 0.961 and for 2023 is 0.948. The correlation for 2024 is notably very high for the selected sample in 2024, indicating tight marking of the scripts.

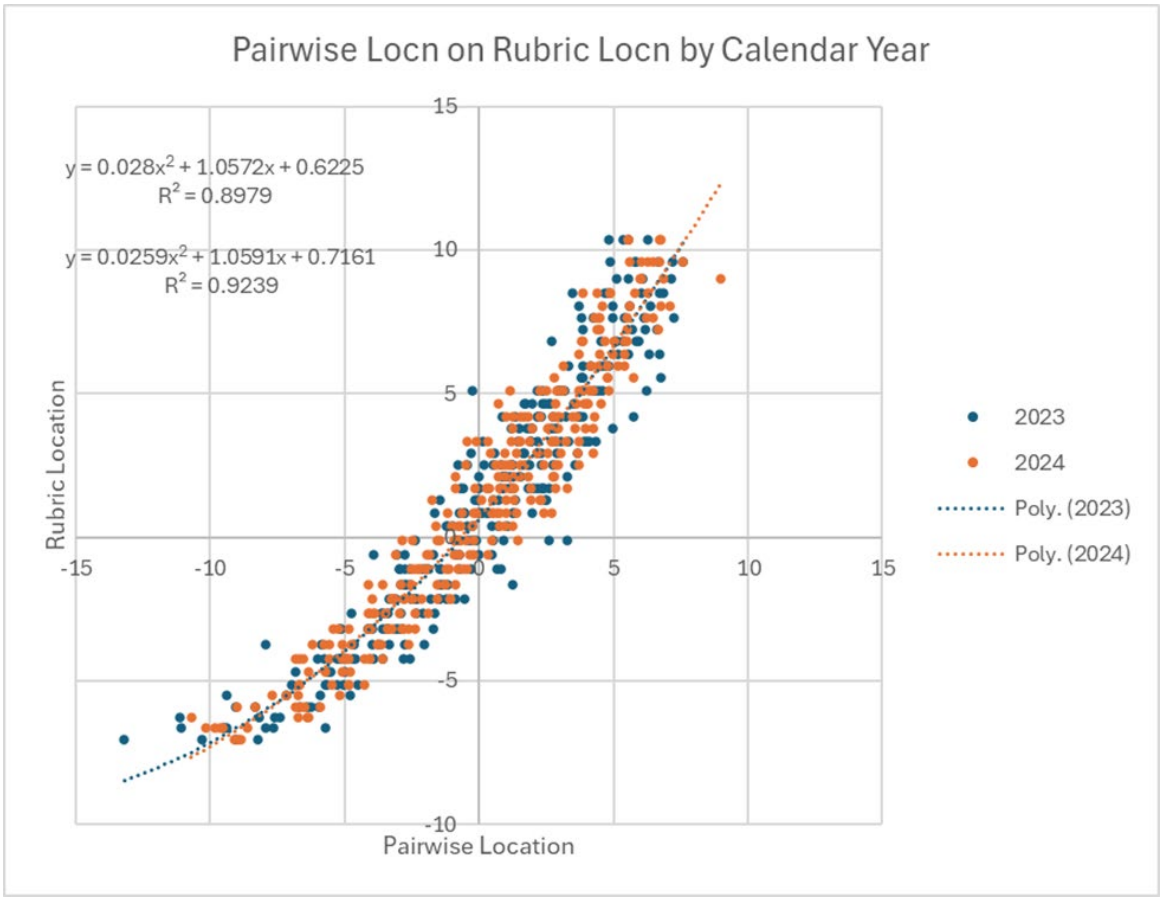


Figure 38. Rubric location estimates (y-axis) plotted against the pairwise location estimates from the 2023 project for the 2023 and 2024 scripts (x-axis).

Figure 39 shows the pairwise scale locations (x-axis) plotted against the NAPLAN rubric locations (*y*-axis) for the 2023 and 2024 paper scripts only. The data points in Figure 39 show the locations of the Year 3 scripts only, as only Year 3 students provided responses on paper. The data points for the 2023 scripts are shown in a different colour to the data points for the 2024 scripts, and separate regression curves are shown.

It can be seen in Figure 39 that the regression curves effectively have the same line of best fit, indicating very consistent marking for the Year 3/paper scripts.

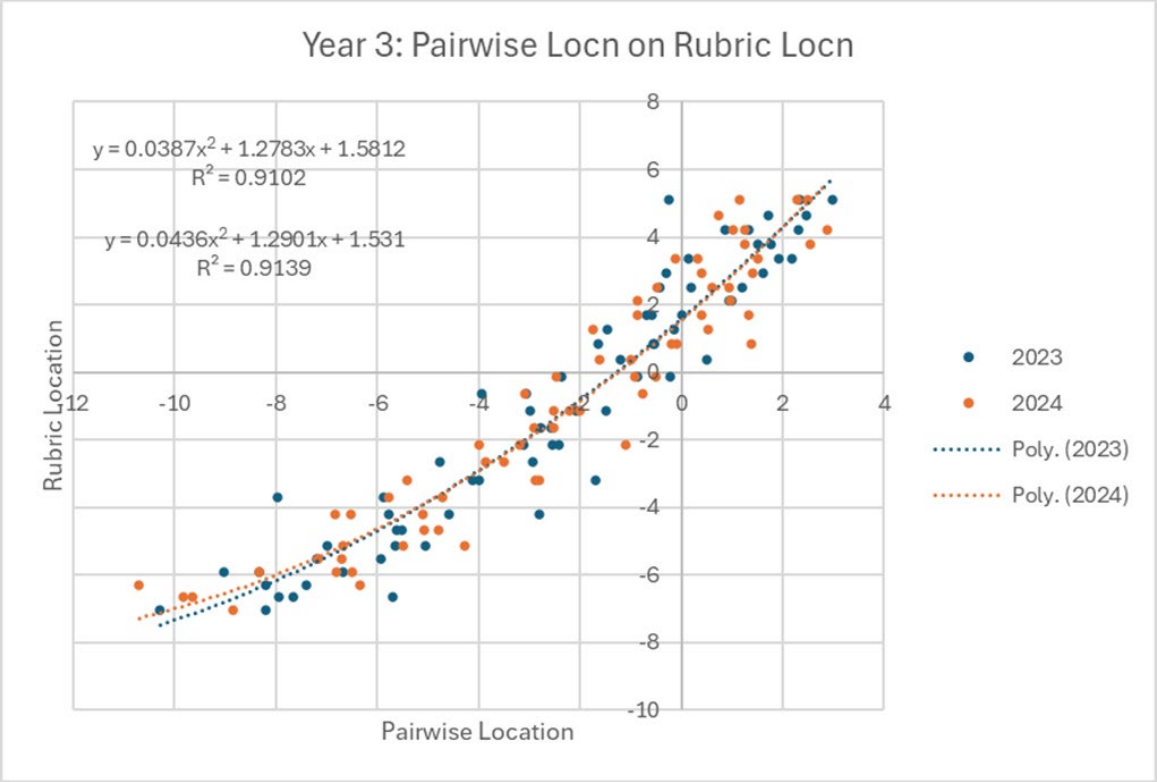


Figure 39. Rubric location estimates plotted against the pairwise location estimates from the 2024 project for the 2023 and 2024 year 3 paper scripts.

Overall, the 2024 pairwise study showed that for the selected sample, rubric scores in 2024 are highly consistent with rubric scores in 2023 for the selected samples of scripts. Figure 38 shows that for any given location along the paired comparison scale (x-axis), the predicted rubric scores for 2023 and 2024 are highly similar, and the distribution and range of actual rubric scores is generally similar. The results showed that 2023 and 2024 performances scaled together well to form a single scale, which indicates that there is a common pairwise scale. Together these observations imply that with the pairwise comparison scale as a reference, marking for the selected sample was consistent across 2023 and 2024.

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 58. These were used to standardise each domain scale to have an overall mean of 500 and standard deviation of 100 as follows:

$$Score_{NAPLANScale} = 100 \cdot \frac{Score_{logit} - DomainMean_{2023}}{DomainStdDeviation_{2023}} + 500 \qquad (5)$$

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:

$$SE_{NAPLANScale} = 100 \cdot \frac{SE_{logit}}{DomainStdDeviation_{2023}} \qquad (6)$$

Table 58. Domain mean scores and standard deviations for transforming logits to NAPLAN scale scores

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

Summary of equating parameter estimates for NAPLAN 2024

In 2024, the NAPLAN scales for each domain were equated to the 2023 delta-centred scale separately, then the vertical shifts used in 2023 were applied to place 2024 results onto the reset NAPLAN scale. For each domain, the same shifts were applied to the students' ability estimates and then transformed to the NAPLAN scale score as below:

$$\theta^{x}_{NAPLAN} = \theta^{xy}_{2024} + HorizontalShift^{xy}_{2024to2023} + VerticalShift^{xy}_{2023} \quad (7)$$

$$Score^{xy}_{NAPLANScale} = \frac{(\theta^{xy}_{NAPLAN} - Mean^{x}_{2023})}{StdDeviation^{x}_{2023}} * 100 + 500 \quad (8)$$

where:

- θ^{xy}_{2024} is the 2024 achievement score in logits on the Year y delta-centred scale for domain x
- θ^{x}_{NAPLAN} is the equated 2024 achievement score in logits on the 2023 reset NAPLAN scale for domain x
- $HorizontalShift^{xy}_{2024to2023}$ is the horizontal shift from the 2024 delta-centred scale to the 2023 delta-centred scale for Year y and domain x
- $VerticalShift^{xy}_{2023}$ is the vertical shift from Year y to Year 5 for domain x used in 2023.

All shifts are listed in Table 4. For writing, both $HorizontalShift^{xy}_{2024to2023}$ and $VerticalShift^{xy}_{2023}$ equal zero.

$Score^{xy}_{NAPLANScale}$ is the scale score for Year *y* and 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 58. The same transformation was applied to all 4 year levels.

Table 59. Summary of parameters for transforming the 2024 logit scores to the NAPLAN reporting scales

		Horizontal shift	Vertical shift	Mean	Standard deviation
Numeracy	Year 3	-0.17672	-1.33551	0.24273	1.67176
	Year 5	-0.23199	0.00000	0.24273	1.67176
	Year 7	-0.01786	0.65079	0.24273	1.67176
	Year 9	-0.14542	1.35306	0.24273	1.67176
Reading	Year 3	-0.26577	-1.01874	0.21845	1.41868
	Year 5	-0.08820	0.00000	0.21845	1.41868
	Year 7	-0.06491	0.60641	0.21845	1.41868
	Year 9	-0.04028	1.04814	0.21845	1.41868
Spelling	Year 3	-0.43161	-2.12529	0.24156	2.77813
	Year 5	-0.13339	0.00000	0.24156	2.77813
	Year 7	-0.13475	1.32911	0.24156	2.77813
	Year 9	-0.08081	2.02187	0.24156	2.77813
Grammar and punctuation	Year 3	-0.20856	-0.79962	0.26014	1.29412
	Year 5	0.10948	0.00000	0.26014	1.29412
	Year 7	-0.12676	0.73657	0.26014	1.29412
	Year 9	-0.06784	0.86432	0.26014	1.29412
Writing	Year 3	0.00000	0.00000	0.62741	3.16266
	Year 5	0.00000	0.00000	0.62741	3.16266
	Year 7	0.00000	0.00000	0.62741	3.16266
	Year 9	0.00000	0.00000	0.62741	3.16266

Estimating equating errors

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

The uncertainty that results from the selection of a subset of link items is referred to as *equating error*. This error should be taken into account when making comparisons between the results from different data collections across time (see Chapter 8). The exact magnitude of the equating error cannot be determined. We can, however, estimate the likely range of magnitudes for this error and take this error into account when interpreting results. As with sampling or measurement errors, the likely range of magnitude for the combined errors is represented as a standard error of each reported statistic.

In 2024, the equating errors were determined in order to compare student achievement for numeracy, reading, spelling, and grammar and punctuation between 2024 and 2023. The equating between 2024 and

2023 NAPLAN tests was through a set of horizontal link items. Hence, the equating error between 2024 and 2023 tests was the standard error associated with the final selection of link items (see the section Equating of numeracy, reading, spelling, and grammar and punctuation above).

Table 60 shows the standard errors of equating associated with each test domain and year level in logits and in scale scores. The scale scores were transformed from the logit values, by applying the factors from formula (2); that is, the 2023 standard deviation and 100.

Table 60. Standard errors of equating

		Logit	Scale score
Numeracy	Year 3	0.01718	1.0277
	Year 5	0.01767	1.0571
	Year 7	0.01379	0.8250
	Year 9	0.01613	0.9646
Reading	Year 3	0.01756	1.2381
	Year 5	0.01537	1.0832
	Year 7	0.01365	0.9625
	Year 9	0.01337	0.9425
Spelling	Year 3	0.01848	0.6653
	Year 5	0.01932	0.6955
	Year 7	0.01662	0.5983
	Year 9	0.01906	0.6859
Grammar and punctuation	Year 3	0.01543	1.1927
	Year 5	0.01531	1.1834
	Year 7	0.01690	1.3062
	Year 9	0.01662	1.2843
Writing*	Years 3, 5, 7 & 9	0.11881	3.7627

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

Estimation of equating error for writing

In 2024, writing results were equated to the reset 2023 scale by fixing the criterion and step calibrations to their 2023 values. This does not allow an equating error to be estimated. However, there is still variability associated with the transition between the 2023 and 2024 scales. The pairwise study was used to provide an estimate of this variability, which stands as a proxy for equating error when calculating the statistical significance of year-on-year changes in writing performance. The calculation is shown below.

Let n_{2023} and n_{2024} be the number of 2023 and 2024 scripts that were placed on the pairwise scale, respectively.

Let β_i be the location of script *i* in logits obtained through marking against the NAPLAN rubric.

Let γ_i be the location of script *i* in logits obtained through pairwise comparisons.

Then γ_i is transformed onto the NAPLAN rubric scale, standardising using the means and standard deviations of the βs and γs as follows:

$$\gamma_i^* = \frac{(\gamma_i - \bar{\gamma}) * SD(\beta)}{SD(\gamma)} + \bar{\beta} \qquad (9)$$

Then $SE_{2023\ to\ pairwise} = \sqrt{\frac{\sum_{i=1}^{n_{2023}} (\beta_i - \gamma_i^*)^2}{n_{2023}}}$, $SE_{2024\ to\ pairwise} = \sqrt{\frac{\sum_{i=1}^{n_{2024}} (\beta_i - \gamma_i^*)^2}{n_{2024}}}$,

and $SE_{2023\ to\ 2024} = \sqrt{SE^2_{2023\ to\ pairwise} + SE^2_{2024\ to\ pairwise}}$

This gives an approximation of the writing equating error.

The equating errors were taken into account, together with sampling and measurement errors, in estimating the standard errors used to determine statistical significance in the comparisons between mean scores across years in NAPLAN reports. The equating errors are not included when estimating standard errors of estimates used to determine statistical significance in the comparisons between mean scores of different subgroups within NAPLAN 2024. This is further explained in Chapter 8.

Estimates of standard errors of equating for percentages of students at or above a proficiency level in different calendar years required a different estimation process. Further details regarding the application of standard errors to testing the statistical significance of performance differences are given in Chapter 8.

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:

- **Needs additional support (NAS):** 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.

The cut-points on the NAPLAN scale for the proficiency levels in each domain and year level were set in the lead-up to the 2023 assessment. This involved a 3-step process.

- **standard-setting** by panels of experienced and expert teachers, along with curriculum and assessment specialists from states and territories
- **statistical analysis** of the cut-points to ensure that they reflected a smooth growth trajectory from Year 3 to Year 9
- **validation** of the skill descriptions of the associated with each proficiency level.

This process was described in detail in the NAPLAN 2023 Technical Report.

Reporting against proficiency levels

In 2024, NAPLAN was reported against proficiency levels in various ways, depending on the report.

- **Individual student report**

The student's proficiency level in each domain was reported.

- **Student and school summary report**

The proficiency level of each student in the school was reported.

- **National results**

The percentage of student results falling in each proficiency level was reported: nationally, and for each state, territory or demographic subgroup.

- **My School**

No proficiency level information was reported.

Proficiency level cut-points for NAPLAN

The complete set of proficiency level cut-points for NAPLAN is shown in Table 61.

Table 61: Proficiency level cut-points for NAPLAN

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

These cut-points, established in 2023, remain in place for 2024 and will continue to do so for future years as a benchmark of the proficiency levels. Changes in performance for a cohort of students 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)[10] is a preliminary report with student- and school-level results for school staff. The individual student report (ISR)[11] 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[12], 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:

$$\theta = \frac{1}{5}\sum_{i=1}^{5}\theta_i \qquad (10)$$

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.

Three types of errors were estimated and different types of combinations of the standard errors were used for different types of comparisons. The first type of error was the uncertainty caused by the selection of students participating in the study: the sampling error. The second type of error was uncertainty caused by the measurement tool (the tests): the measurement error. The third type was uncertainty caused by the equating design: the equating error. Estimation of the equating error was explained in Chapter 6. The other 2 types of errors are explained in this chapter.

Sampling error

The inclusion of sampling error might be considered surprising in that all students in the target year levels were included in the assessment. However, the aim of NAPLAN is to make inferences about trends in the educational systems over time and not about the specific student cohorts in 2024. 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.

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

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 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 2024, 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 40. 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 40. 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:

$$B = \frac{1}{4}\sum_{i=1}^{5}(\theta_i - \theta)^2 \qquad (11)$$

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:

$$V = U + \left(1 + \frac{1}{5}\right) B \qquad (12)$$

$$SE = \sqrt{V} \qquad (13)$$

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 2024 performance in the reports.

Testing for differences

Two types of differences could be computed and tested for significance. The first type of comparison was between subgroups within the NAPLAN 2024 data; 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:

$$SE_{DIFF} = \sqrt{{SE_A}^2 + {SE_B}^2} \qquad (14)$$

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.

The second example involves statistical testing of performance differences between calendar years. This requires inclusion of the equating error in the calculation of SE_{DIFF}. Drawing on the previous example, if we now consider the difference between group A's mean score in 2024 and 2023, we need to add the equating error between these 2 years, $SE_{2024to2023}$, to the calculation in the following way:

$$SE_{DIFF} = \sqrt{{SE_{A23}}^2 + {SE_{A24}}^2 + {SE_{2024to2023}}^2} \qquad (15)$$

The same procedure as shown in the previous example can then be applied to evaluate the statistical significance of the difference. Actual equating errors for comparisons of mean scale scores involving 2024 NAPLAN with 2023 for each domain and year level are included in Chapter 6.

Only when differences between subgroups are compared between calendar years – for example, the gap between Indigenous and non-Indigenous students over time – does the equating error not need to be taken into account. This is because both group statistics are equally affected by uncertainty due to equating, which is therefore cancelled out. This type of comparison, however, is not included in the NAPLAN 2024 National Report.

Effect sizes

All significance testing in NAPLAN is accompanied by an effect size measure, which indicates the magnitude of any difference as opposed to indicating the likelihood that the difference could have arisen through chance alone. The incorporation of 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:

$$g = \frac{m_2 - m_1}{s_p} \quad (16)$$

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

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \quad (17)$$

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:

$$g_{unbiased} = g_{biased}\left[1 - \frac{3}{4(n_1+n_2-2)}\right] \quad (18)$$

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, Camberwell, Victoria: Australian Council for Educational Research.

Adams JR and Lazendic G (2013) *Observations on the Feasibility of a Multistage Test Design for NAPLAN*, unpublished technical report.

ACARA (Australian Assessment, Curriculum and Reporting Authority) (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.

Bradley R and Terry M (1952) "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons", *Biometrika*, 39 (3/4):324–345.

Breithaupt K and Hare D R (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), Thousand Oaks: Sage.

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*, New York: John Wiley & Sons.

Lord FM and Novick MR (1968) *Statistical Theories of Mental Test Scores,* Addison-Wesley: Menlo Park.

Luce R D (1959) *Individual Choice Behavior: A Theoretical Analysis*, Wiley.

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*, Copenhagen: Danmark Paedagpgiske Institut.

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

Rubin DB (1991) "EM and beyond", *Psychometrika*, 39:111–21.

Warm TA (1989) "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika,* 54(3): 427–50.

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