

NAPLAN 2021

Technical Report

June 2022



National Assessment Program – Literacy and Numeracy (NAPLAN) 2021: technical report

Copyright

© Australian Curriculum, Assessment and Reporting Authority (ACARA) 2022, 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;
2. the Australian Government logo and the Education Services Australia Limited logo;
3. other trade mark protected material;
4. photographs; and
5. 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) 2022, 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, Centennial Plaza
280 Elizabeth Street
Sydney NSW 2000
T. 1300 895 563
F. 1800 982 118
www.acara.edu.au

The appropriate citation for this report is:

Australian Curriculum, Assessment and Reporting Authority
2022, *National Assessment Program – Literacy and Numeracy*
2021: *Technical Report*, ACARA, Sydney

Acknowledgements

ACARA worked with several contractors to successfully complete NAPLAN 2021. The contractors were: the Australian Council for Educational Research (ACER) for the central analysis of data, sampling and item development; Educational Measurement Solutions (EMS) for item trial analysis; University of Western Australia (UWA) for pairwise equating of writing data; and University of New South Wales Global Assessments (UNSWG) and National Foundation for Educational Assessments (NFER) for the development of items.

Contributors to the technical report are Yan Bibby, Xiaoxun Sun, Eunjung Lee, Nathan Zoanetti from ACER; Stephen Humphry from UWA; and Eveline Gebhardt, Anna Cohen, Rassoul Sadeghi, Leigh Paterson, Mark Glasby, Jasmine Cohen, Peter Noonan, Karina Welch, Ross Bindon and James Lay from ACARA.

ACARA would also like to acknowledge the technical input from the Measurement Advisory Group (MAG). MAG members are Ray Adams (chair), Barry McGaw, Siek Toon Khoo, David Andrich, Catherine McClellan, Gage Kingsbury and Mark Wilson

Contents

List of tables	7
List of figures	10
Chapter 1: Introduction	14
Chapter 2: Item development and item trial	16
Item development.....	16
Numeracy item development	16
Reading item development	17
Conventions of language item development	18
Writing task development	19
Item trial.....	19
Item trial test design.....	19
Test administration.....	22
Participants	23
Marking	23
Psychometric analysis of item trial data.....	24
Item selection for the 2021 NAPLAN tests	27
Chapter 3: NAPLAN test design	28
Multi-stage, tailored test design	28
Construction of NAPLAN Online tests	30
Test length	30
Difficulty of testlets	31
Item types for online tests.....	33
Curriculum coverage.....	34
Paper test design	42
Writing test design.....	44
Writing marking training and quality assurance.....	46
Example items in reporting bands	49
Setting branching rules	63
Branching rules for numeracy, reading and grammar and punctuation tests	64
Branching rules for spelling	67
Pathway utilisation	69
Chapter 4: Data collection and preparation	71
Data collection and delivery	71
Data cleaning validation process	72
Data preparation	72
Distribution of not reached items	74
Not reached items in online tests.....	74
Not reached items in paper tests	77
Final student participation rates	77

Chapter 5: Scaling methodology and outcomes	79
Scaling model.....	79
Software used for analyses.....	79
Item calibration	80
Review of test and item characteristics	81
Test reliability	81
Test targeting and item spread	82
Item fit	86
Differential Item Functioning (DIF) Analyses	89
Estimation of student ability and generation of PVs	98
Chapter 6: Equating procedures.....	101
Equating of numeracy, reading, spelling, and grammar & punctuation results	101
Horizontal equating shifts of the online tests	104
Horizontal equating shifts for paper tests	113
Vertical equating shifts of the online tests	122
Vertical link item review of paper tests	128
Horizontal–vertical regression (HVR) equating shifts (paper tests)	135
Final shifts.....	138
Scaling factors	138
Equating of writing results	140
Pairwise Study Results	141
Standardisation of scales from logits to reporting scales	145
Summary of equating parameter estimates for NAPLAN 2021	145
Estimating equating errors	147
Chapter 7: NAPLAN proficiency bands.....	151
Illustrations	156
Chapter 8: Reporting of national results	157
Calculation of statistics using plausible values	157
Computation of standard errors	157
Sampling error	158
Measurement error	158
Testing for differences	159
Effect sizes.....	160
Reporting of geographically classified statistics	161
References.....	162
Appendix A - Percentages and ability distribution by pathway.....	163
Appendix B - Item analysis details	164
Appendix C - Item summary tables	165
Appendix D - Item characteristic curves.....	166
Appendix E - Item-person maps.....	167

Appendix F - Gender DIF analysis	168
Appendix G - LBOTE DIF analysis	169
Appendix H - ATSI Status DIF analysis	170
Appendix I - DIF summary tables	171
Appendix J - Jurisdictional DIF	172
Appendix K - Horizontal link item comparison	173
Appendix L - Vertical link item comparisons	174
Appendix M - Exception report	175

List of tables

Table 1. Number of items developed for numeracy	17
Table 2. Composition of the trial numeracy item pool.....	20
Table 3. Composition of the trial reading item pool.....	21
Table 4. Composition of the trial grammar and punctuation item pool	21
Table 5. Composition of the trial spelling item pool	21
Table 6. Composition of the trial spelling item pool	21
Table 7. Writing by task and total responses	22
Table 8. Target and achieved number of students for the online item trial sample, by domain and year level.....	23
Table 9. NAPLAN Online Numeracy test: number of items and time available	31
Table 10. NAPLAN online reading test. number of items and time available.....	31
Table 11: NAPLAN Online Conventions of Language test: number of items and time available	31
Table 12: NAPLAN Online numeracy: predefined difficulty parameters for each testlet..	32
Table 13: NAPLAN online reading: predefined difficulty parameters for each testlet	32
Table 14: NAPLAN online grammar and punctuation: predefined difficulty parameters for each testlet.....	32
Table 15. NAPLAN Online spelling: predefined difficulty parameters for each testlet	33
Table 16. NAPLAN online numeracy: item types in the item pool by year level	33
Table 17. NAPLAN online reading: item types in the item pool by year level	33
Table 18. NAPLAN online conventions of language: item types in the item pool by year level	34
Table 19. NAPLAN Numeracy Year 3 curriculum coverage by mode and pathway	35
Table 20. NAPLAN Numeracy Year 5 curriculum coverage by mode and pathway	35
Table 21. NAPLAN Numeracy Year 7 curriculum coverage by mode and pathway	36
Table 22. NAPLAN Numeracy Year 9 curriculum coverage by mode and pathway	36
Table 23. NAPLAN Reading Year 3 curriculum coverage by mode and pathway	37
Table 24. NAPLAN Reading Year 5 curriculum coverage by mode and pathway	37
Table 25. NAPLAN Reading Year 7 curriculum coverage by mode and pathway	38
Table 26. NAPLAN Reading Year 9 curriculum coverage by mode and pathway	38
Table 27. NAPLAN Conventions of Language Year 3 curriculum coverage by mode and pathway.....	39
Table 28. NAPLAN Conventions of Language Year 5 curriculum coverage by mode and pathway.....	40
Table 29. NAPLAN Conventions of Language Year 7 curriculum coverage by mode and pathway.....	41
Table 30. NAPLAN Conventions of Language Year 9 curriculum coverage by mode and pathway.....	42
Table 31. NAPLAN numeracy paper test number of items and time available	43
Table 32. NAPLAN reading paper test number of items and time available	43
Table 33. NAPLAN language conventions paper test number of items and time available	43
Table 34. NAPLAN Writing prompt designation schedule according to test day	44
Table 35. Recommended allocation of time for the writing test.....	44

Table 36. NAPLAN Narrative marking criteria and skill focus descriptions	45
Table 37. NAPLAN Narrative marking criteria and score categories.....	46
Table 38. Writing scripts marked for each jurisdiction	46
Table 39. NAPLAN 2021 marking centre operational periods and duration by jurisdiction	47
Table 40. Approximate number of NAPLAN writing markers per day by jurisdiction.	47
Table 41. The number of Training, Practice and Control scripts developed for each prompt.....	48
Table 42. National marking protocols.....	48
Table 43. Numeracy example items in reporting bands.....	49
Table 44. Reading example items in reporting bands.....	52
Table 45. Grammar and punctuation example items in reporting bands.....	57
Table 46. Spelling items in bands.....	60
Table 47. Example writing prompt.....	63
Table 48. Stage 1 cut scores (Testlet A to C B D)	66
Table 49. Stage 2 cut scores (testlet AB to C E F)	66
Table 50. Stage 2 cut scores (testlet AD–C E F)	67
Table 51. Stage 1, Testlet SA–SB SD cut scores.....	68
Table 52. Stage 2, Testlets SA–SB to PB PD cut scores	69
Table 53. Stage 2, Testlet SA–SD to PB PD cut scores.....	69
Table 54: Rules for data coding.....	73
Table 55: Pathway assignment rules to incomplete online tests	74
Table 56: Student participation rates.....	78
Table 57. Reliability (WLE) for NAPLAN 2021 paper tests.....	82
Table 58. Summay of item statistics in NAPLAN 2021 online tests	87
Table 59. Summay of item statistics in NAPLAN 2021 paper tests.....	87
Table 60. Number of items showing gender DIF by domain by year level.....	90
Table 61. Number of Items Showing LBOTE DIF by Domain by Year Level.....	91
Table 62. Number of items showing Indigenous DIF by domain by year level	93
Table 63. Number of items showing state/territory DIF by domain by year level	95
Table 64. Number of students by device.....	97
Table 65. Number of items showing device DIF by domain by year level	98
Table 66. Equating design for both assessment modes.....	102
Table 67. Horizontal link review summary for online tests.....	112
Table 68. Horizontal equating shifts between 2021 item locations and 2019 item locations by year level for online tests	113
Table 69. Horizontal link review summary for paper tests	121
Table 70. Horizontal equating shifts between 2021 item locations and item locations on the historical NAPLAN scale for paper tests	121
Table 71. Vertical link review summary.....	134
Table 72. Vertical shift constants between adjacent year levels	135
Table 73. Vertical shift constants from each year level to Year 5.....	135
Table 74. Example of comparing horizontal shifts with vertical shifts (numeracy, paper test).....	136
Table 75. Regression intercepts and slopes of paper tests.....	137

Table 76. Final shifts applied for equating NAPLAN 2021	138
Table 77. Local means and scaling factors	139
Table 78. Domain mean and standard deviation for transforming logits to NAPLAN scale scores.....	145
Table 79. Summary of parameters for transforming the 2021 logit scores to the NAPLAN reporting scales.....	146
Table 80. Standard errors of equating	149
Table 81. Lower bounds of proficiency bands in scale scores and in logits.....	151
Table 82. Described scale for numeracy	152
Table 83. Described scale for reading.....	153
Table 84. Described scale for writing	154
Table 85. Described scale for conventions of language	155

List of figures

Figure 1. A sample ICC for a poorly performing item.....	25
Figure 2. A sample MC distractor curve for a poorly performing item	25
Figure 3. A sample ICC for a well-performing item	25
Figure 4. A sample MC distractor curve for a well- performing item.....	26
Figure 5. A sample ICC displaying gender DIF in favor of girls	27
Figure 6. A sample ICC displaying gender DIF in favor of boys	27
Figure 7. The multistage tailored test design for numeracy and reading	29
Figure 8: Online test design for conventions of language.....	30
Figure 9. Test information functions: curves for testlets C, B and D	65
Figure 10. Stage 1. Testlet A–C B D cut scores	65
Figure 11. Stage 2. Testlet AB–C E F cut scores	66
Figure 12. Stage 2. Testlet AD–C E F cut scores	67
Figure 13. Stage 1. Testlet SA–SB SD cut scores.....	68
Figure 14. Stage 2. Testlet SA–SB to PB PD cut scores.....	68
Figure 15. Stage 2. Testlets SA–SD to PB PD cut scores.....	69
Figure 16. Percentage of students assigned to each pathway in Year 3 numeracy	70
Figure 17. Ability distribution by pathway for Year 3 numeracy.....	70
Figure 18. Trailing missing percentages in numeracy online test.....	75
Figure 19. Trailing missing percentages in reading online test	75
Figure 20. Trailing missing percentages in spelling online test	76
Figure 21. Trailing missing percentages in grammar & punctuation online test.....	76
Figure 22. Trailing missing percentages in numeracy, reading, spelling and grammar & punctuation paper tests.....	77
Figure 23. Wright map for Year 3 numeracy online test (an example).....	83
Figure 24. Wright map for online writing test (a polytomous example)	84
Figure 25. Thurstonian thresholds for online writing test	85
Figure 26. Item characteristic curves for an item with $infit = 1.00$	88
Figure 27. Item characteristic curves for an item with $infit = 1.29$	89
Figure 28. Example of item characteristic curves displaying gender DIF†.....	91
Figure 29. Example of item characteristic curves displaying LBOTE DIF†	92
Figure 30. Example of item characteristic curves displaying Indigenous DIF†	93
Figure 31. Example of item characteristic curves displaying jurisdictional DIF	94
Figure 32. Conditioning variables for the multidimensional item response model with latent regression model.....	100
Figure 33. Scatterplot of numeracy, horizontal equating items between 2021 and 2019 for Year 3 online students	104
Figure 34. Scatterplot of numeracy, horizontal equating items between 2021 and 2019 for Year 5 online students	105
Figure 35. Scatterplot of numeracy, horizontal equating items between 2021 and 2019 for Year 7 online students	105
Figure 36. Scatterplot of numeracy, horizontal equating items between 2021 and 2019 for Year 9 online students	106
Figure 37. Scatterplot of reading, horizontal equating items between 2021 and 2019 for Year 3 online students	106

Figure 38. Scatterplot of reading, horizontal equating items between 2021 and 2019 for Year 5 online students	107
Figure 39. Scatterplot of reading, horizontal equating items between 2021 and 2019 for Year 7 online students	107
Figure 40. Scatterplot of reading, horizontal equating items between 2021 and 2019 for Year 9 online students	108
Figure 41. Scatterplot of spelling, horizontal equating items between 2021 and 2019 for Year 3 online students	108
Figure 42. Scatterplot of spelling, horizontal equating items between 2021 and 2019 for Year 5 online students	109
Figure 43. Scatterplot of spelling, horizontal equating items between 2021 and 2019 for Year 7 online students	109
Figure 44. Scatterplot of spelling, horizontal equating items between 2021 and 2019 for Year 9 online students	110
Figure 45 Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2019 for Year 3 online students	110
Figure 46 Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2019 for Year 5 online students	111
Figure 47. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2019 for Year 7 online students	111
Figure 48. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2019 for Year 9 online students	112
Figure 49. Scatterplot of numeracy, horizontal equating items between 2021 and 2009 for Year 3 paper students	113
Figure 50. Scatterplot of numeracy, horizontal equating items between 2021 and 2009 for Year 5 paper students	114
Figure 51. Scatterplot of numeracy, horizontal equating items between 2021 and 2009 for Year 7 paper students	114
Figure 52. Scatterplot of numeracy, horizontal equating items between 2021 and 2009 for Year 9 paper students	115
Figure 53. Scatterplot of reading, horizontal equating items between 2021 and 2009 for Year 3 paper students.....	115
Figure 54. Scatterplot of reading, horizontal equating items between 2021 and 2009 for Year 5 paper students)	116
Figure 55. Scatterplot of reading, horizontal equating items between 2021 and 2009 for Year 7 paper students.....	116
Figure 56. Scatterplot of reading, horizontal equating items between 2021 and 2009 for Year 9 paper students.....	117
Figure 57. Scatterplot of spelling, horizontal equating items between 2021 and 2009 for Year 3 paper students.....	117
Figure 58. Scatterplot of spelling, horizontal equating items between 2021 and 2009 for Year 5 paper students.....	118
Figure 59. Scatterplot of spelling, horizontal equating items between 2021 and 2009 for Year 7 paper students.....	118
Figure 60. Scatterplot of spelling, horizontal equating items between 2021 and 2009 for Year 9 paper students.....	119
Figure 61. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2009 for Year 3 paper students.....	119
Figure 62. Scatterplot of grammar and punctuation, horizontal equating items between	

2021 and 2009 for Year 5 paper students.....	120
Figure 63. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2009 for Year 7 paper students.....	120
Figure 64. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2009 for Year 9 paper students.....	121
Figure 65. Scatterplot for vertical link item review for numeracy between Year 3 and Year 5 online tests.....	122
Figure 66. Scatterplot for vertical link item review for numeracy between Year 5 and Year 7 online tests.....	123
Figure 67. Scatterplot for vertical link item review for numeracy between Year 7 and Year 9 online tests.....	123
Figure 68. Scatterplot for vertical link item review for reading between Year 3 and Year 5 online tests.....	124
Figure 69. Scatterplot for vertical link item review for reading between Year 5 and Year 7 online tests.....	124
Figure 70. Scatterplot for vertical link item review for reading between Year 7 and Year 9 online tests.....	125
Figure 71. Scatterplot for vertical link item review for spelling between Year 3 and Year 5 online tests.....	125
Figure 72. Scatterplot for vertical link item review for spelling between Year 5 and Year 7 online tests.....	126
Figure 73. Scatterplot for vertical link item review for spelling between Year 7 and Year 9 online tests.....	126
Figure 74. Scatterplot for vertical link item review for grammar and punctuation between Year 3 and Year 5 online tests.....	127
Figure 75. Scatterplot for vertical link item review for grammar and punctuation between Year 5 and Year 7 online tests.....	127
Figure 76. Scatterplot for vertical link item review for grammar and punctuation between Year 7 and Year 9 online tests.....	128
Figure 77. Scatterplot for vertical link item review for numeracy between Year 3 and Year 5 paper tests.....	128
Figure 78. Scatterplot for vertical link item review for numeracy between Year 5 and Year 7 paper tests.....	129
Figure 79. Scatterplot for vertical link item review for numeracy between Year 7 and Year 9 paper tests.....	129
Figure 80. Scatterplot for vertical link item review for reading between Year 3 and Year 5 paper tests.....	130
Figure 81. Scatterplot for vertical link item review for reading between Year 5 and Year 7 paper tests.....	130
Figure 82. Scatterplot for vertical link item review for reading between Year 7 and Year 9 paper tests.....	131
Figure 83. Scatterplot for vertical link item review for spelling between Year 3 and Year 5 paper tests.....	131
Figure 84. Scatterplot for vertical link item review for spelling between Year 5 and Year 7 paper tests.....	132
Figure 85. Scatterplot for vertical link item review for spelling between Year 7 and Year 9 paper tests.....	132
Figure 86. Scatterplot for vertical link item review for grammar and punctuation between Year 3 and Year 5 paper tests.....	133

Figure 87. Scatterplot for vertical link item review for grammar and punctuation between Year 5 and Year 7 paper tests.....	133
Figure 88. Scatterplot for vertical link item review for grammar and punctuation between Year 7 and Year 9 paper tests.....	134
Figure 89. Comparisons of horizontal and vertical shifts of the paper tests.....	137
Figure 90. Scatterplot for writing criteria between 2021 and 2019 online and paper tests	140
Figure 91. Scatterplot of the NAPLAN rubric and pairwise scale locations for 2016 and 2019 paper performances.....	142
Figure 92. Scatterplot of the NAPLAN rubric and pairwise scale locations for 2021 online performances and 2019 online performances.....	143
Figure 93. Scatterplot of the NAPLAN rubric and pairwise scale locations, for all 2016 and 2021 performances used in the pairwise equating.	144
Figure 94. A schematic of the equating errors accumulated across NAPLAN administrations	148
Figure 95. Schematic picture of proficiency bands by year levels.....	156
Figure 96. Examples in SPSS and SAS for estimating sampling variance	158

Chapter 1: Introduction

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

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

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

The NAPLAN 2021 tests were administered nationally in May. As in previous cycles of NAPLAN, students at each year level were assessed in five domains: reading, writing, language conventions (spelling, grammar and punctuation), and numeracy.

The Australian Council for Educational Research (ACER) was appointed by the Australian Curriculum, Assessment and Reporting Authority (ACARA) to undertake the central analysis of test data from the NAPLAN 2021 administration.

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

NAPLAN results are reported using five 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. There are also 10 proficiency bands that span Years 3, 5, 7 and 9. Each year level is reported against six of these bands.

Over one million students in Years 3, 5, 7 and 9 in all states and territories of Australia participated in NAPLAN 2021. From 2008 to 2017, NAPLAN delivered only paper-based tests. From 2018¹ NAPLAN delivered both paper-based tests and online multistage adaptive tailored tests. The online tailored tests in reading, spelling, grammar and punctuation, and numeracy were delivered to students in participating schools. In 2021, approximately 50 per cent of students took the NAPLAN test online (30% in 2019 and 15% in 2018). NAPLAN in 2020 was cancelled due to the COVID-19 pandemic.

¹ Education ministers made the decision to cancel NAPLAN in 2020 due to the COVID-19 pandemic. As a result, NAPLAN tests constructed for the 2020 assessment cycle were used in the 2021 NAPLAN.

Five outcome reports were produced for NAPLAN 2021. The first report was the Student and School Summary reports (SSSR). This interactive report was for online schools only, it provided an opportunity for schools to take a first glance at the achievement of their students. The second report was a report with preliminary national outcomes, also called the Summary Report. The first cut of the census data was used for this report. The third report type was the Individual Student Report (ISR), providing information to parents about their children's performance on the NAPLAN tests. The fourth report was the official NAPLAN 2021 National Report that was based on the second cut of the census data. The National Report for 2021 and all previous NAPLAN assessments are available on the ACARA website. The final cut of the census data was used for the school-level online *My School* reports, which are beyond the scope of this technical report.

The aim of this technical report is to describe in detail the methodology used for NAPLAN 2021. Chapter 2 of this report describes the NAPLAN 2021 item trial. Chapter 3 describes the test design. Chapter 4 describes the data preparation process. Chapter 5 describes scaling methodology and outcomes. Chapter 6 describes the test equating processes to place the NAPLAN 2021 tests on the NAPLAN historic scales. Chapter 7 describes the proficiency bands on the NAPLAN scales. Chapter 8 describes the methodology used for reporting of NAPLAN 2021 performance.

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

Chapter 2: Item development and item trial

The aim of this chapter is to describe the item development and trial activities for the NAPLAN 2021 test. There are three main components in the NAPLAN item trial: 1) item development, 2) item trialing and 3) psychometric analysis. The first part of this chapter describes the item and test development process, the second part the item trial administration and the third part focuses on the psychometric analysis.

Item development

External contractors were hired for developing new items in each of the assessment domains. Item development required contractors to conform to the following documents:

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

Contractors delivered items in batches across the project period, from September 2018 until June 2019. Items in each batch were reviewed by ACARA and the National Testing Working Group (NTWG). Feedback was synthesised by ACARA and the items were returned to the contractors for revisions. All modified items were reviewed by ACARA before final delivery in May-June 2019.

Contractors submitted compliance tables showing how the items met the specifications outlined in the contracts. Source files of all graphics were also supplied.

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

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

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

Numeracy item development

Items for the NAPLAN 2021 Numeracy tests were procured from two separate contractors. The main contractor, the University of New South Wales Global through the business group University of New South Wales Global Assessments (UNSWG) provided ACARA with items from the Number and Algebra, Measurement and Geometry, and Statistics and Probability strands.

The second contractor, the National Foundation for Educational Research (NFER), provided year 7/9 link items from the Number and Algebra, Measurement and Geometry, and Statistics and Probability strands.

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

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

Table 1. Number of items developed for numeracy

	Measurement and Geometry	Statistics and Probability	Number and Algebra
Year 3	34	19	66
Year 5	28	16	56
Year 7	40	20	70
Year 7/9	16	8	24
Year 9	40	20	71
Total	158	83	287

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

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

Reading item development

ACARA contracted UNSWG to produce 66 reading units predominantly targeting the lower and upper end of the performance scale for Years 3, 5, 7 and 9. UNSWG's final delivery included 67 stimulus texts and 536 items.

NFER was contracted to provide 45 items to supplement pre-existing reading units, most of which had been trialled but not yet used in a main study. These additional items were required to ensure the pre-existing units could readily fit testlet boundaries.

ACER was contracted to produce 40 standalone items (10 at each of Years 3, 5, 7 and 9). Standalone items are items targeting specific skills that can be used on their own or with a very short stimulus text. These items were designed to target the lower and upper ends of the performance scale.

An additional package, procured from NFER in April 2019, consisted of 12 units (96 items) designed to target the F testlet at each NAPLAN year level. These included 'paired units'; two stimulus texts with items focused on each stimulus text as well as a small number of items requiring students to synthesise information from both stimulus texts.

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

The difficulty of items, to a large extent, was dependent on the complexity of the stimulus texts. A common concern for NAPLAN reading items was appropriate targeting for early childhood and entry-level texts for all years. Entry-level texts target students working at a skill level 1–3 years below their school year level using subject matter that is still engaging and age appropriate for these students. All Year 3 texts and entry-level Year 5 texts were

reviewed by experienced pre-primary and/or primary teachers. Entry-level Year 7 and Year 9 texts were also reviewed by teachers who have extensive experience with students of lower reading ability.

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

- online readability, particularly in font selection, and text layouts aimed at reducing the need for scrolling
- accessibility for visually impaired students, taking into account ACARA's guidelines for colour, contrast and font selection
- resource file size was kept at a maximum of 150 kb per text

The requirement to provide texts in the HTML format created some challenges, as this was the first time contractors were required to provide all stimulus texts in HTML.

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

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

ACARA facilitated five reading reviews of the reading stimuli and items over a six-month period. Feedback was sought from the NTWG and ACARA's student diversity specialist. ACARA synthesised the feedback, and items were returned to contractors classified as 'accepted', 'needing modification as specified' or 'needing replacement'. Items continued to be refined until the final delivery was made in May 2019.

Conventions of language item development

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

Spelling items were developed by the ACARA Writing / Conventions of Language team. Target words were sourced from past NAPLAN writing trial scripts. The team identified the words students commonly misspell as well as likely error patterns. The words were used in simple age-appropriate context sentences that provided enough support for the misspelt words to be readily understood. Items were allocated to audio dictation, mistake-identified or mistake-not-identified (proofreading) sections of the spelling test and assigned targeted testlets according to year level, predicted difficulty, skill focus and item type.

ACARA developed 270 audio dictation items, 72 mistake-identified items and 112 mistake-not-identified spelling items, and facilitated three reviews of the spelling items over a six-month period. Feedback was sought from the NTWG and ACARA's student diversity specialist. All modifications to items were made by ACARA. Audio was recorded for all audio dictation items prior to trialling.

Grammar and punctuation items were developed by NFER and UNSWG. These contractors delivered four batches of items, totaling approximately 351 grammar and 94 punctuation items; six testlets for each of Years 3, 5, 7 and 9. ACARA facilitated five reviews of the grammar and punctuation items over a six-month period. Additional feedback on accessibility alternative items was sought from NTWG and ACARA's student diversity specialist. All modifications to items were made by ACARA.

Writing task development

Prompts for the NAPLAN 2021 Writing were developed and trialled according to the following process:

1. Education experts from all jurisdictions developed a large pool of writing tasks to engage students in Years 3 and 5, and Years 7 and 9. Each jurisdiction convened panels of experts with significant experience in writing assessment, and educators representing key special needs groups.
2. Expert panels undertook four stages of review of all writing tasks in the pool to ensure that they were accessible for students from a range of backgrounds. Panels considered what students might write about and whether the task would be fair for students. In early stages of the review, the panels prioritised the national pool of writing topics, providing feedback where necessary. In later stages of the review, they distilled the suitable tasks and suggested changes to wording and images.
3. A shortlist of eight topics was chosen and refined. Approximately five-thousand students each responded to two of these tasks under test conditions. The student writing from the trials was marked and markers gave feedback on how students engaged with each task. The marking data were analysed to discern which tasks best ensured fairness and measurement reliability. Psychometric analysis of the tasks confirmed that scores were reliable and valid for each year group. At least three tasks were selected for each of Years 3 and 5, and Years 7 and 9.
4. The National Testing Working Group gave advice regarding the final sequence and allocation of writing tasks.

Item trial

In the item trial process, items were trialled to obtain critical item performance data used to guide construction of the final NAPLAN tests and build each domain's item bank. Trialling allowed additional quantitative and qualitative feedback on the tests to be gathered, including time on task, engagement with test content and identification of online display issues. Individual items and suites of test items (based on common stimulus texts) were administered to samples of students within Australia. Psychometric analysis of the data, conducted after the trial, was used to evaluate the performance of each individual item. The Educational Measurement Solution (EMS) was engaged to analyse items that were included in tests for each of the test domains.

Item trial test design

As there was no NAPLAN testing in 2020 due to the impacts of COVID, new items for the NAPLAN 2021 tests were drawn from the item trial conducted in 2019.

To support the placement of items on the NAPLAN scale, the test is administered to a representative, stratified sample of schools and students (within the constraints of technology transition). The trial test includes items from the previous main study so that the trial results can be equated to the historical NAPLAN scale.

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 testlets were presented at differing positions within the tests. To illustrate, Year 3 reading had the following rotational design:

- Twenty testlets plus one testlet of stand-alone items²
- Three nodes: node 1 had one testlet with approximately 10 stand-alone items. Nodes 2 and 3 had ten testlets each.
- Students started by answering a single stand-alone item from node 1, then *either* one testlet from node 2 followed by one testlet from node 3, *or* one testlet from node 3 followed by one testlet from node 2. As such, every item was trialled in two different positions, with half of the students seeing the item in the first half of the test and half seeing the item in the final half of the test.
- The equating units were placed in the trial design to approximate their position in the main study. Testlet A units were placed towards the start of a testlet, and testlet E units, towards the end of a testlet.

A number of items were included in adjacent NAPLAN year levels (for example, Year 3 and Year 5.) This enables reviewing psychometric properties of the items for several 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.

The total item pool for numeracy was 792; for reading, 1,814; for grammar and punctuation, 1086; and for spelling, 599. Table 2 to Table 7 below show the composition of the trial pools by domain and by item format: multiple choice (MC) and constructed response (CR) which includes technology enhanced items.

Table 2. Composition of the trial numeracy item pool

	MC	CR	Total
Year 3	78	66	144
Year 5	89	79	168
Year 7	130	110	240
Year 9	144	96	240
Total	393	399	792

² An item set consisting of a very short stimulus text and usually just one item. These are used to target specific reading skills and/or locations on the scale.

Table 3. Composition of the trial reading item pool

	MC	CR	Total
Year 3	331	99	430
Year 5	317	123	440
Year 7	326	151	477
Year 9	324	153	477
Total	1,298	526	1,814

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

	MC	CR	Total
Year 3	128	144	272
Year 5	138	134	272
Year 7	161	110	271
Year 9	171	100	271
Total	598	488	1086

Table 5. Composition of the trial spelling item pool

	MC	CR	Total
Year 3	0	150	150
Year 5	0	149	149
Year 7	0	150	150
Year 9	0	150	150
Total	0	599	599

Table 6. Composition of the trial spelling item pool

	AD	MNI/MI	Total
Year 3	98	149	150
Year 5	95	152	149
Year 7	95	156	150
Year 9	97	154	150
Total	385	611	599

Eight writing tasks were trialled at each of Years 3, 5, 7 and 9. These included prompts for both the persuasive and narrative genres. The tasks were administered in a rotational design based on classes, not individual students. For example, Class A was allocated tasks 1 and 8, Class B was allocated tasks 2 and 7, Class C was allocated tasks 3 and 6,

etc. Students in Years 5, 7 and 9, and the majority of students in Year 3 completed two tasks online. Approximately 250 Year 3 students completed one task online and one task on paper.

Table 7. Writing by task and total responses

Prompt	Year 3	Year 5	Year 7	Year 9	Total
Task 1	318	375	311	277	1,281
Task 2	325	339	314	309	1,287
Task 3	330	360	345	312	1,347
Task 4	322	353	336	300	1,311
Task 5	325	316	318	265	1,224
Task 6	357	262	278	257	1,154
Task 7	345	307	220	237	1,109
Task 8	322	353	336	300	1,311
Task 1 paper	137				137
Task 7 paper	110				110
Total	2,569	2,312	2,122	1,957	8,960

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

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

Test administration

The Educational Services Australia (ESA) test delivery platform was used to administer the trial tests in a sample of schools in Australia for all domains of the NAPLAN program. Schools from all states and territories participated in the trial from 29 July to 16 August 2019.

A trained invigilator was sent to each trial school to deliver and collect the trial assessment materials (to ensure the security of the materials), and to also observe and support the classroom teacher throughout the assessment and student survey. At the completion of each assessment and student survey session, the invigilator and the classroom teacher each completed a session report to provide feedback about aspects of the trial administration. This feedback, in conjunction with feedback from a range of other sources, informed the selection and refinement of items for the final pool of assessment items and the design of the 2021 NAPLAN tests.

Participants

A sample of 401 schools across all states and territories participated. The trial schools were selected to reflect the range of educational contexts around the nation and included schools from government, Catholic and independent sectors; low and high socioeconomic areas; metropolitan and regional locations; large and small schools; and students from a variety of language backgrounds. The following schools were not included in the sample:

- very remote schools
- schools with less than 15 students in targeted years
- schools that participated in the previous year's trial or equating study
- schools participating in NAP Sample field trial or main study
- distance education schools
- Montessori, Steiner, Waldorf schools
- non-mainstream schools
- schools without NAPLAN performance data.

Schools across all states and territories participated. The target student sample size for each domain and year level, and subsequent achieved sample, is presented in Table 8. Each student completed tests from two different domains. The minimum number of responses for each item was set at 250 to achieve stable item parameters.

Table 8. Target and achieved number of students for the online item trial sample, by domain and year level

Domain	Sample	Year 3	Year 5	Year 7	Year 9	Total
Reading	Target	3000	3000	3000	3000	12000
	Achieved	3731	3730	3200	3104	13765
CoL	Target	1800	1800	1800	1800	7200
	Achieved	2200	2235	2072	1970	8477
Numeracy	Target	1200	1200	1200	1200	4800
	Achieved	1499	1468	1753	1698	6418
Writing (online)	Target	1000	1000	1000	1000	4000
	Achieved	1295	1273	1198	1137	4903

Marking

Pearson were contracted to develop marking materials and manage marking operations for the NAPLAN 2021 trial of writing tasks. A team of experienced NAPLAN markers was engaged by Pearson for marking the writing scripts. ACARA's writing test manager supported Pearson's training of the markers, and remained on-site to oversee the marking process. Once the marking of each prompt was completed, a debriefing session was held

with the test developers and amendments were made to the training materials as necessary. Qualitative feedback on the marking of each prompt was gathered to be used alongside the quantitative data when selecting prompts for the main study.

Psychometric analysis of item trial data

The trial data were extracted from the assessment platform and then sent to an external contractor, Educational Measurement Solutions (EMS), for analysis. Writing data was marked by another contractor and the marked data were also sent to EMS for analysis.

The following steps have been taken to analyse data from the item trial conducted in 2019:

1. *Data validation and recoding*: In order to ensure the data were of high quality and could be used in the analysis, each data set was validated separately and anomalies were removed. Raw data were also recoded to suit the purposes of analysis: embedded missing responses were coded '9', and items not administered to a student were coded '8'.
2. *Year level analysis*: Data for each year level were analysed separately for each domain. Two rounds of analyses were undertaken:
 - a. The purpose of the first round of analyses was to identify mis-keyed items. Output files were sent to ACARA's NAPLAN team for identification of possible mis-keys and identification of items with poor psychometric properties (and thus should be omitted from all subsequent analysis).
 - b. The purpose of the second round of analyses (with acceptable items) was to calibrate items and place them on the historical NAPLAN scale..

The Rasch measurement model (Rasch, 1960), using ACER Conquest (Adams, Wu, Cloney & Wilson; 2020) and RUMM software, was used for item calibration. In the Rasch model, the probability of a correct response to an item is modeled as a logistic function of the difference between person ability and item difficulty. The Rasch measurement models permit the separation of the item difficulty and student ability parameters. In practical terms, this means that if data conform to the underlying model, then the measurement of students on the variable is independent of the difficulty of items used to obtain the measures. Similarly, the item difficulty can be determined through a process of item calibration independent of the distribution of achievement of students involved in the data collection. The mathematical form of the model is provided in Chapter 5.

Key criteria for judging the performance of items were measures of item fit statistics (weighted MNSQ) and item performance illustrated by item characteristic curves (ICCs). Sample Item Characteristic Curve (ICC) and MC distractor curves are displayed in Figure 1 to Figure 4. In these graphs, student abilities are on the horizontal axis, and probability of responding correctly is on the vertical axis. The solid lines are the expected curves from the model, the broken lines are the observed proportions from the data. For multiple choice items, the graphs include a curve for each response category. Items that do not fit the model, do not discriminate well between high- and low- performing students. The items have a high MNSQ value (larger than approximately 1.2) and the curve for the correct response has a slope flatter than the expected curve. Facilities, item-rest correlations and point-biserial correlations were noted, but only informed decisions to eliminate items if other indices were poor.

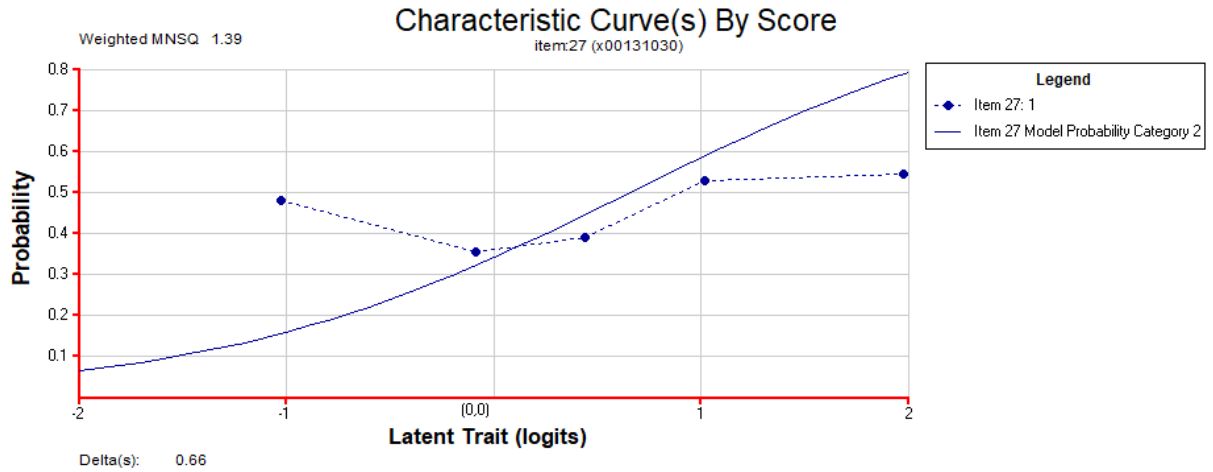


Figure 1. A sample ICC for a poorly performing item

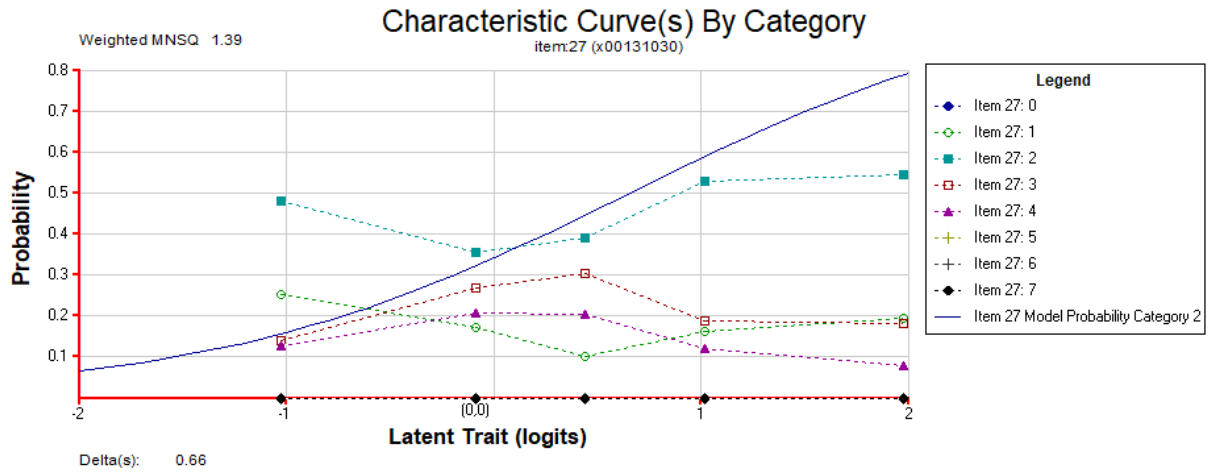


Figure 2. A sample MC distractor curve for a poorly performing item

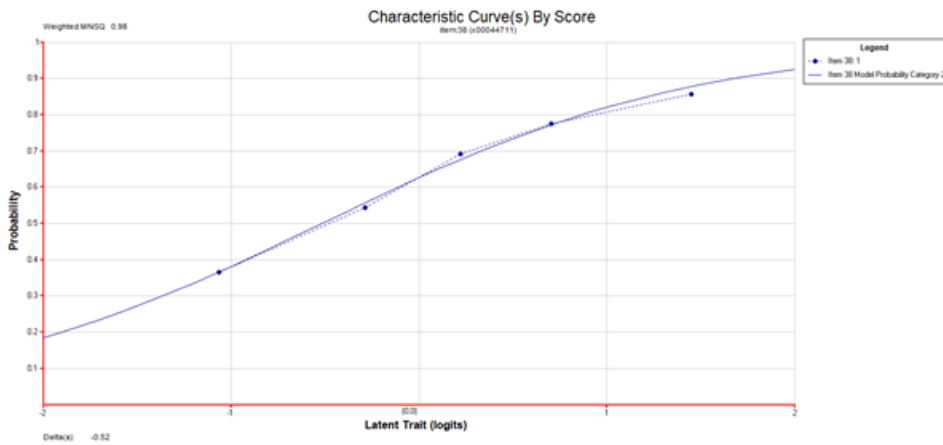


Figure 3. A sample ICC for a well-performing item

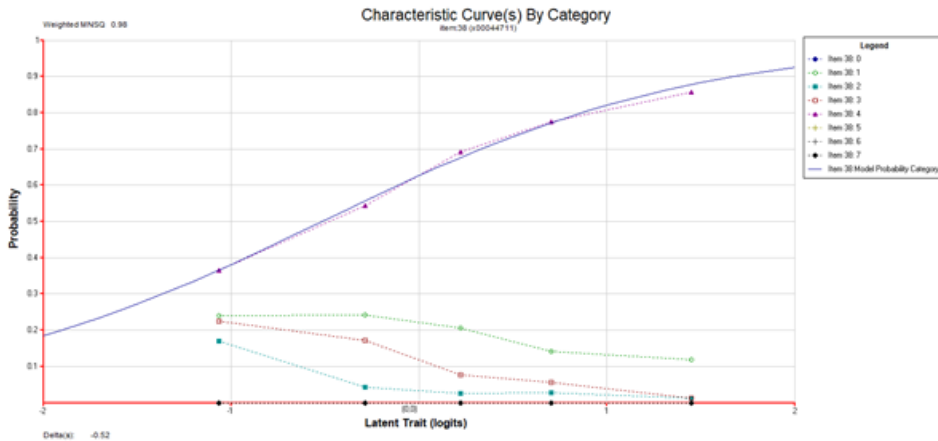


Figure 4. A sample MC distractor curve for a well- performing item

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

When the interaction term was significantly different from zero at the 95 per cent confidence level, an item was deemed as showing DIF. An additional criterion applied was that a difference in item difficulty between boys and girls had to be larger than 0.4 logits before the item was deemed to show large gender DIF.

In cases where items displayed a large gender DIF, content experts inspected the reasons for the observed bias. The items were flagged but not automatically removed simply based on statistical evidence of bias. Items were discarded only where there was an agreement between the psychometric evidence and the content experts' review. Two sample ICCs displaying gender DIF are illustrated in Figure 5 and Figure 6. Each graph includes an observed line for each gender group. When lines are more than 0.4 logits apart, the item was flagged for gender DIF.

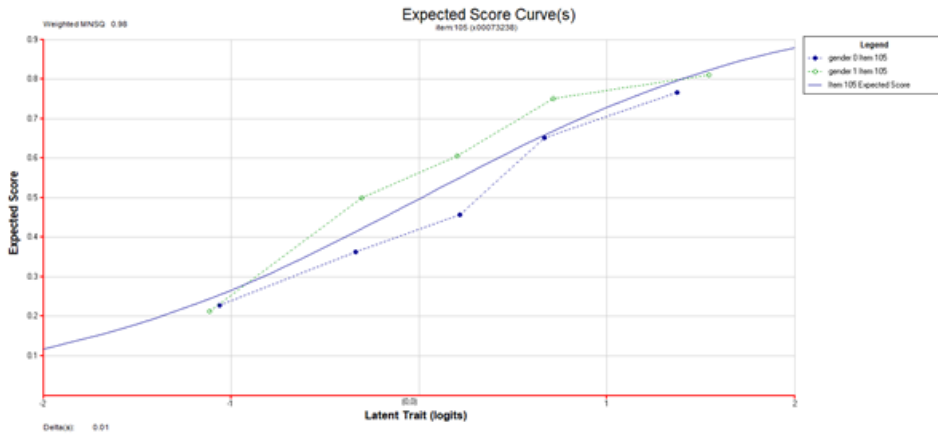


Figure 5. A sample ICC displaying gender DIF in favor of girls

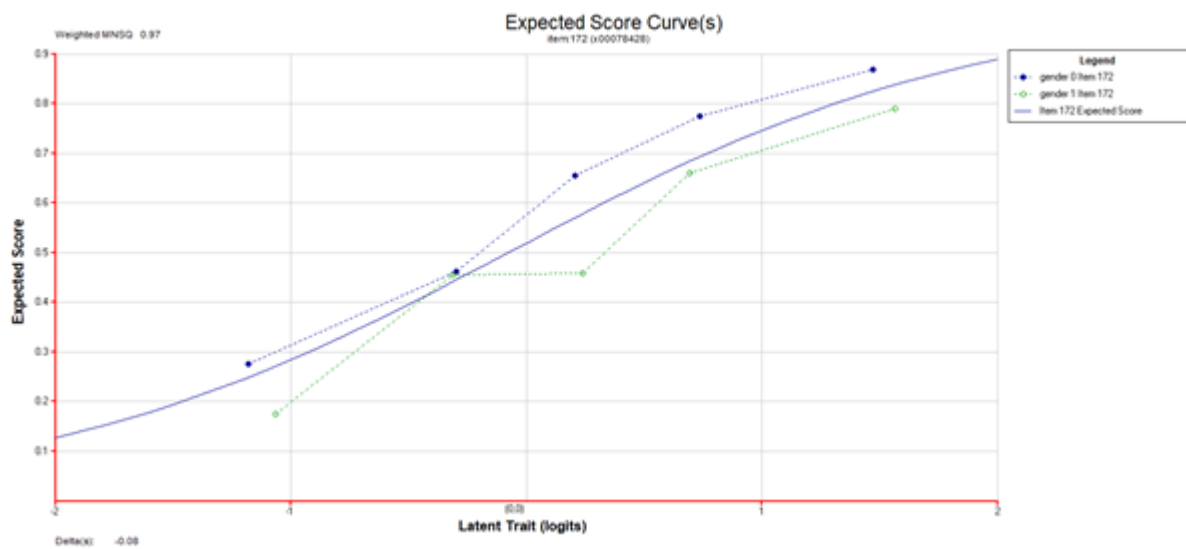


Figure 6. A sample ICC displaying gender DIF in favor of boys

Item selection for the 2021 NAPLAN tests

The results emerging from the psychometric analysis provided a pool of psychometrically sound items for test managers to select for inclusion in the final NAPLAN 2021 tests. Furthermore, results obtained from DIF analysis enabled test managers to exclude those items that displayed bias against students of a particular gender.

Chapter 3: NAPLAN test design

The aim of this chapter is to describe the NAPLAN 2021 test design. The first part of this chapter describes the test design for both online and paper tests. The branching method implemented in the NAPLAN multistage tailored test design is discussed in the second part.

Multi-stage, 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. Multi-stage tailored tests are considered a balanced compromise between non-adaptive paper-and-pencil and item-level adaptive tests (Hendrickson, 2007).

Some benefits of tailored testing are:

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

The multistage tailored test design for numeracy, and reading is illustrated in Figure 7. This figure shows a design with six nodes A, B, C, D, E and F. Each node comprises three testlets (e.g. A1, A2, A3), of which one is randomly allocated to the student. Each student completes three testlets in one of the following ordered combinations: ABC, ABE, ABF, ADC, ADE, ADF or ACB.

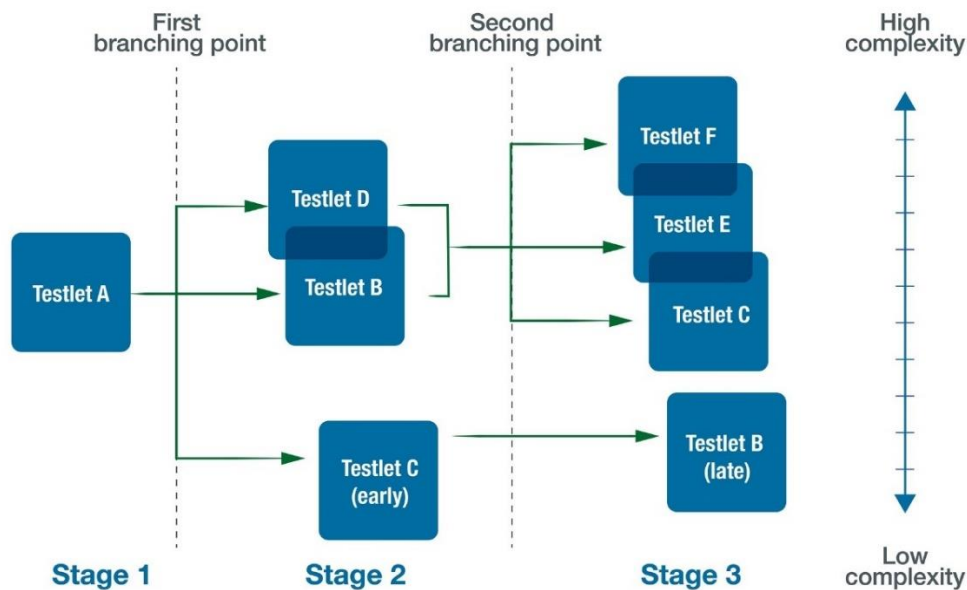


Figure 7. The multistage tailored test design for numeracy and reading

Students at each year level start with testlet A. Each student's answers to testlet A determines the testlet they will be branched to and, as such, the questions they see. These may be less complex (B) or more complex (D). The student's answers in the 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 Online results for each student are based on both the number of the questions the student answers correctly and the average difficulty of the items that were assigned to the student. A student who completes a more complex set of questions is more likely to achieve a higher score (and a higher band placement), while a student who answers the same number of questions correctly, but follows a less complex pathway, will achieve a lower score.

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

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

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

The grammar and punctuation section of the CoL test has the same multistage, multistream adaptive test design as numeracy and reading. The spelling test has a similar design but with only two testlets in the third stage (PD and PB). The graphical representation of the CoL test design is illustrated in Figure 8.

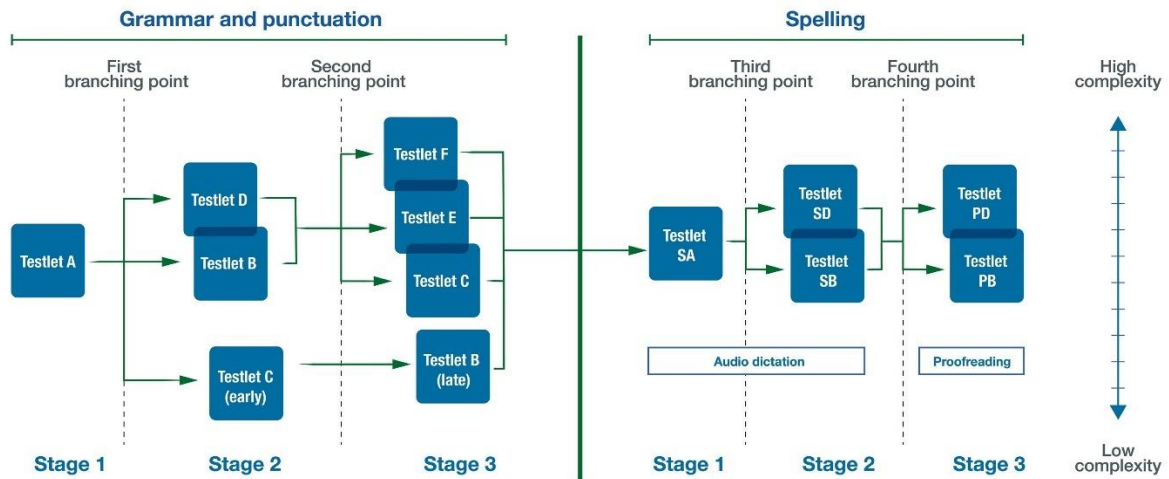


Figure 8: Online test design for conventions of language

As Figure 8 shows, the first two stages of the spelling section are focused on an audio component while the third stage is used to test proofreading. The spelling multistage design is discussed in more detail in the ‘Setting branching rules’ section.

Construction of NAPLAN Online tests

The tests were constructed for use in 2020, but because of the cancellation of NAPLAN 2020 due to COVID-19, the tests were used for 2021. Data from the trial and 2019 main study largely determined the placement of items within testlets. Skills, curriculum strands and proficiencies were balanced across nodes and testlets. When populating test designs, the choice and placement of link items were usually considered before other items, as they were vital to ensure comparability across vertical year levels and from calendar year to calendar year.

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

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

Test length

Table 9 to Table 11 outline the test lengths for each domain. The grammar and punctuation and spelling sections of the conventions of language tests are not delineated by year level as there were no differences in the specifications for each.

Table 9. NAPLAN Online Numeracy test: number of items and time available

Numeracy		Items per testlet	Total test items	Time available
Year 3		12	36	45 minutes
Year 5		14	42	50 minutes
Year 7	CA ³	16 items x ½ testlet (8 items)	48	65 minutes
	NC ⁴	16 items x 2 ½ testlets (40 items)		
Year 9	CA	16 items x ½ testlet (8 items)	48	65 minutes
	NC	16 items x 2 ½ testlets (40 items)		

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

Table 10. 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 11: NAPLAN Online Conventions of Language test: number of items and time available

Conventions of language	Items per testlet	Item per section	Total test items	Time available
Grammar and punctuation	9	27	52	45 minutes
Spelling	6 items per testlet (audio dictation) 9 items per testlet (audio dictation) 10 items per testlet (proofreading)	25		

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

³ CA – calculator-allowed

⁴ NC – non-calculator

from least to most complex. For reading, in general, the unit⁵ with the lower average difficulty was presented first in each testlet and the unit with the higher average difficulty was presented last.

Table 12 to Table 15 outline the predefined difficulty ranges in logits and average difficulty for the testlets in each test.

Table 12: NAPLAN Online numeracy: predefined difficulty parameters for each testlet

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

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

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

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

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

⁵ A reading unit comprises 1 stimulus text with 4-7 items related to that stimulus text.

Table 15. NAPLAN Online spelling: predefined difficulty parameters for each testlet

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

Item types for online tests

The distribution of item types across the NAPLAN numeracy tests was nominally set at 40 per cent multiple-choice(s) items, 15 per cent text entry (constructed response) and 45 per cent technology-enhanced items (TEI). The reading tests include multiple choice(s) and technology-enhanced items only.

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

Table 16 to Table 18 show the final distribution of item types in the suite of items at each year level.

Table 16. NAPLAN online numeracy: item types in the item pool by year level

Numeracy	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in suite
Year 3	88	30	38	156
Year 5	88	36	50	174
Year 7	114	51	51	216
Year 9	117	53	38	208

Table 17. NAPLAN online reading: item types in the item pool by year level

Reading	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in suite
Year 3	191	-	42	233
Year 5	239	-	34	273
Year 7	280	-	40	320
Year 9	247	-	41	288

Table 18. NAPLAN online conventions of language: item types in the item pool by year level

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

Curriculum coverage

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

For numeracy, the focus in Algebra is on pre-algebra concepts at Years 3, 5 and 7. At Year 9, after students have been introduced to variables in Year 7, the split between Algebra and Number is more pronounced. Therefore, the percentage split in Year 9 is for 40 per cent Algebra, and 15 per cent Number.

For grammar and punctuation, the focus is predominantly on the sentence-level grammar, word-level grammar and punctuation sub-domains with a smaller focus on editing, text cohesion and vocabulary. Spelling items make up around half of a conventions of language test. Curriculum coverage is summarised in Table 19 to Table 30.

Table 19. NAPLAN Numeracy Year 3 curriculum coverage by mode and pathway

Year 3	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	58%	56%	55%	56%	56%	57%
<i>Measurement and Geometry</i>	30%	28%	30%	30%	31%	31%	30%
<i>Statistics and Probability</i>	15%	14%	14%	16%	14%	12%	14%
Proficiencies							
<i>Fluency</i>	20%	19%	29%	35%	35%	26%	20%
<i>Understanding</i>	30%	31%	24%	29%	24%	24%	25%
<i>Problem-solving</i>	30%	31%	31%	26%	28%	32%	35%
<i>Reasoning</i>	20%	19%	16%	10%	13%	18%	19%
Item types							
<i>MC/MCS</i>	60%	72%	60%	55%	58%	59%	56%
<i>Text entry</i>	15%	28%	14%	19%	19%	20%	19%
<i>Interactive</i>	25%	-	26%	26%	22%	20%	25%

Table 20. NAPLAN Numeracy Year 5 curriculum coverage by mode and pathway

Year 5	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	55%	56%	56%	57%	56%	57%
<i>Measurement and Geometry</i>	30%	29%	29%	29%	29%	29%	29%
<i>Statistics and Probability</i>	15%	14%	15%	14%	14%	15%	15%
Proficiencies							
<i>Fluency</i>	20%	19%	31%	21%	18%	16%	23%
<i>Understanding</i>	30%	29%	33%	30%	25%	23%	25%
<i>Problem-solving</i>	30%	31%	25%	31%	33%	37%	31%
<i>Reasoning</i>	20%	21%	12%	18%	24%	24%	21%
Item types							
<i>MC/MCS</i>	60%	64%	53%	59%	58%	58%	56%
<i>Text entry</i>	15%	36%	15%	11%	13%	14%	14%
<i>Interactive</i>	25%	-	33%	30%	29%	28%	29%

Table 21. NAPLAN Numeracy Year 7 curriculum coverage by mode and pathway

Year 7	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	56%	53%	55%	53%	51%	51%
<i>Measurement and Geometry</i>	30%	31%	32%	31%	32%	34%	33%
<i>Statistics and Probability</i>	15%	13%	15%	15%	15%	15%	16%
Proficiencies							
<i>Fluency</i>	20%	21%	28%	23%	21%	21%	21%
<i>Understanding</i>	30%	29%	42%	34%	31%	23%	32%
<i>Problem-solving</i>	30%	29%	21%	30%	31%	33%	31%
<i>Reasoning</i>	20%	21%	10%	13%	17%	23%	16%
Item types							
<i>MC/MCS</i>	60%	67%	53%	52%	58%	53%	52%
<i>Text entry</i>	15%	33%	20%	19%	18%	26%	24%
<i>Interactive</i>	25%	-	27%	28%	24%	20%	24%

Table 22. NAPLAN Numeracy Year 9 curriculum coverage by mode and pathway

Year 9	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	54%	56%	57%	56%	56%	55%
<i>Measurement and Geometry</i>	30%	31%	31%	31%	31%	30%	32%
<i>Statistics and Probability</i>	15%	15%	13%	13%	13%	14%	13%
Proficiencies							
<i>Fluency</i>	20%	19%	26%	24%	19%	17%	23%
<i>Understanding</i>	30%	33%	39%	38%	35%	28%	32%
<i>Problem-solving</i>	30%	29%	19%	23%	28%	33%	27%
<i>Reasoning</i>	20%	19%	16%	15%	19%	22%	18%
Item types							
<i>MC/MCS</i>	60%	67%	56%	67%	65%	56%	59%
<i>Text entry</i>	15%	33%	22%	22%	21%	28%	23%
<i>Interactive</i>	25%	-	22%	12%	14%	16%	18%

Table 23. NAPLAN Reading Year 3 curriculum coverage by mode and pathway

Year 3	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Language</i>	5-15%	16%	27%	27%	21%	22%	23%
<i>Literature</i>	5-15%	13%	9%	11%	7%	3%	6%
<i>Literacy</i>	70-90%	71%	65%	62%	72%	75%	71%
Cognitive processes							
<i>Locating & identifying</i>	30-50%	34%	39%	47%	44%	39%	28%
<i>Integrating & interpreting</i>	30-50%	50%	49%	45%	47%	51%	53%
<i>Analysing & evaluating</i>	10-20%	16%	12%	8%	9%	10%	18%
Stimulus texts							
<i>Number of texts</i>		6	-	7	6	6	6
<i>Average word count</i>		191	169	115	163	191	205
Item types							
<i>MC</i>	90-100%	90%	76%	74%	82%	79%	78%
<i>MCs</i>	0-10%	5%	6%	3%	3%	8%	9%
<i>Other</i>	0-10%	5%	18%	24%	15%	13%	13%

Table 24. NAPLAN Reading Year 5 curriculum coverage by mode and pathway

Year 5	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Language</i>	5-15%	8%	19%	18%	22%	22%	19%
<i>Literature</i>	5-15%	8%	12%	11%	10%	13%	16%
<i>Literacy</i>	70-90%	84%	69%	72%	67%	66%	65%
Cognitive processes							
<i>Locating & identifying</i>	30-50%	31%	32%	50%	38%	27%	21%
<i>Integrating & interpreting</i>	30-50%	46%	48%	38%	45%	52%	54%
<i>Analysing & evaluating</i>	10-20%	23%	19%	12%	16%	21%	25%
Stimulus texts							
<i>Number of texts</i>		6	-	6	6	6	6
<i>Average word count</i>		238	227	169	208	230	249
Item types							
<i>MC</i>	90-100%	82%	81%	80%	83%	86%	85%
<i>MCs</i>	0-10%	10%	6%	4%	6%	7%	7%
<i>Other</i>	0-10%	8%	12%	16%	10%	7%	8%

Table 25. NAPLAN Reading Year 7 curriculum coverage by mode and pathway

Year 7	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Language</i>	10-20%	28%	27%	32%	29%	26%	31%
<i>Literature</i>	10-20%	18%	13%	8%	14%	14%	14%
<i>Literacy</i>	50-70%	54%	60%	60%	58%	60%	55%
Cognitive processes							
<i>Locating & identifying</i>	20-40%	22%	29%	40%	32%	27%	23%
<i>Integrating & interpreting</i>	40-60%	54%	50%	49%	54%	52%	48%
<i>Analysing & evaluating</i>	20-40%	24%	21%	10%	14%	21%	29%
Stimulus texts							
<i>Number of texts</i>		8	-	9	9	9	9
<i>Average word count</i>		280	290	232	281	317	323
Item types							
<i>MC</i>	90-100%	94%	83%	88%	88%	82%	80%
<i>MCs</i>	0-10%	2%	4%	1%	3%	6%	7%
<i>Other</i>	0-10%	4%	13%	11%	9%	13%	13%

Table 26. NAPLAN Reading Year 9 curriculum coverage by mode and pathway

Year 9	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Language</i>	10-20%	30%	22%	24%	22%	20%	26%
<i>Literature</i>	10-20%	16%	15%	13%	16%	17%	13%
<i>Literacy</i>	50-70%	54%	64%	63%	63%	63%	61%
Cognitive processes							
<i>Locating & identifying</i>	20-40%	20%	18%	26%	23%	16%	12%
<i>Integrating & interpreting</i>	40-60%	60%	50%	51%	55%	53%	46%
<i>Analysing & evaluating</i>	20-40%	20%	32%	22%	22%	31%	42%
Stimulus texts							
<i>Number of texts</i>		8	-	9	9	9	9
<i>Average word count</i>		319	317	265	323	322	346
Item types							
<i>MC</i>	90-100%	84%	78%	87%	86%	78%	69%
<i>MCs</i>	0-10%	10%	5%	3%	6%	10%	13%
<i>Other</i>	0-10%	6%	14%	10%	8%	11%	19%

Table 27. NAPLAN Conventions of Language Year 3 curriculum coverage by mode and pathway

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

Table 28. NAPLAN Conventions of Language Year 5 curriculum coverage by mode and pathway

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

Table 29. NAPLAN Conventions of Language Year 7 curriculum coverage by mode and pathway

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

Table 30. NAPLAN Conventions of Language Year 9 curriculum coverage by mode and pathway

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

Paper test design

Four paper-based tests were administered at each of Years 3, 5, 7 and 9 as in previous cycles. The four 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.

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

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

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

Table 31 to Table 33 outline the total number of items in each test at each year level and the time available to students to complete the tests.

Table 31. NAPLAN numeracy paper test number of items and time available

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

Table 32. NAPLAN reading paper test number of items and time available

Number of items		Time available	
Year 3	37	45 minutes	
Year 5	39	50 minutes	
Year 7	50	65 minutes	
Year 9	50	65 minutes	

Table 33. NAPLAN language conventions paper test number of items and time available

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

The numeracy, reading and language conventions paper tests were created from a selected subset of online test items. Tables outlining test specifications encompassing average difficulty (logits), alignment to the Australian Curriculum and item types, are included in Table 19 to Table 33.

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 asked 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. In 2021, all students were required to write a narrative text. Prior to the test, neither the student nor the teachers knew what the genre or topic was. Students completed the writing test either on paper (handwritten) or online (typed). All Year 3 students completed the test on paper regardless of whether they completed other test domains online.

In 2021, five writing prompts were used across years 3, 5, 7 and 9, and the paper and online modes. A further two prompts were kept in reserve in case of widespread technical issues or a security breach. No reserves were needed for 2021. Two of the five prompts were assigned to the Years 3 and 5 tests, and three to the Years 7 and 9 tests. The prompt that each student received depended on whether the test was taken on paper or online, and on which day of the writing test window the student sat the test (see Table 34). Each prompt has closely scripted scaffolding, or instructions. All prompts had been trialed to ensure that they were of similar difficulty.

Table 34. NAPLAN Writing prompt designation schedule according to test day

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

All students were given 40 minutes to respond to the prompt. An additional two minutes is allocated to the online tests to allow for listening to the audio recording of the prompt. It is recommended that students divide their time to three stages of writing: planning, writing and editing.

Table 35. 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. The same marking guide is used from year to year and to assess all students' writing, allowing for a national comparison of student writing capabilities across these year levels and over time.

The analytical, criterion-referenced marking guide consists of a rubric and exemplar scripts. The narrative rubric has ten criteria and a total of 47 score points. In each criterion, each score category is cumulative and hierarchical. Each criterion is analysed as a polytomous item. The 10 criteria with the associated number of score categories are shown in Table 36 and Table 37.

Table 36. 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 37. 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

Writing marking training and quality assurance

Students' writing is marked by people who have received intensive training in the application of the ten writing criteria. In 2021, almost 2,000 markers were employed nationally. Most markers were practicing or retired teachers. Test administration authorities in each state and territory were responsible for marking each script within their jurisdictions. In total there was over 1 million student scripts that needed to be marked nationally. See Table 38 below.

Table 38. Writing scripts marked for each jurisdiction

	ACT	NSW	NT	QLD	SA	TAS	VIC	WA	Total
Number of writing scripts	20 100	374 004	10 250	243 000	77 005	24 237	291 021	128 586	1 168 203

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

Each jurisdiction's marking centre⁶ was in operation for different windows of time and for varying durations. The dates of commencement and conclusion were contingent on the number of scripts, the availability of the facilities for training and marking, the contractors'

⁶ Victoria & Tasmania, and NSW & ACT have combined marking operations.

requirements and other factors. There was an overlap where all marking centres were in operation.

Table 39 shows the commencement and conclusion dates of each operation and the sum total of days each marking centre was in operation for.

Table 39. NAPLAN 2021 marking centre operational periods and duration by jurisdiction

	NSW & ACT	NT	QLD	SA	VIC & TAS	WA
Start of marking	16/05	26/05	17/05	22/05	16/05	19/05
Finish of marking	13/06	4/06	6/06	12/06	16/06	12/06
Days of marking	29	10	21	22	31	25

Table 40 shows the approximate number of markers in each jurisdiction each day of their marking operation.

Table 40. Approximate number of NAPLAN writing markers per day by jurisdiction.

	NSW & ACT	NT	QLD	SA	TAS	VIC	WA	Total
Number of markers	439	62	427	200	16	240	129	1900

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

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

A comprehensive online Writing Marker Training course was also provided to Test Administration Authorities for use in training new and experienced markers and leaders. The course was based on the face-to-face course used in previous years and delivered through a Learning Management System (LMS). Over 1600 markers successfully completed the course nationally. Other resources provided for use in preparation for and during the marking period included slideshow presentations, training sample scripts and national marking protocols.

The core component of training and quality assurance was the provision of pre-marked sample scripts with annotations called Training, Practice and Control (TPC) scripts. These scripts were originally selected from the pool of item trial scripts, given individual marks by members of the Marking Quality Team⁷ (MQT), then moderated to arrive at agreed consensus or expert scores for each criterion. Annotations were then written to support

⁷ the MQT is made up of chief markers from the jurisdictions

the reasoning behind the scores. Sixty TPC scripts were developed in total across the five prompts. A subset of these scripts (Training and Practice) was used in the training of new and experienced markers and for 'calibration' or 'benchmarking' scripts to ensure comparability to the assigned expert score.

Table 41. The number of Training, Practice and Control scripts developed for each prompt

	Prompt 1	Prompt 2	Prompt 3	Prompt 3	Prompt 5
Training	10	10	9	0	0
Practice	4	3	2	5	6
Control	10	10	9	0	0
Total	24	23	18	5	6

Control scripts were used to check individual marker accuracy and collect data on national consistency. On each day of the marking period, control script data from each jurisdiction was provided directly to ACARA's secure FTP site. This data was aggregated on a daily basis and a summary marking performance report was periodically provided to each TAA so they could ascertain their jurisdiction's marking accuracy compared with the expert scores for the daily control script and other markers in the nation marking the same control script on the same day. The first control script is issued when the first marking centre commences 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 to accommodate contextual factors, variability in the degree of comparative data is inevitable.

In addition to control scripts, quality assurance through check-marking (sometimes referred to as double marking, spot checking or backmarking) was undertaken for each marker by the group leader, centre leader or other senior person appointed by the Test Administration Authority. Within each marking group or team, check marking covered at least 10% of all scripts marked across the marking operation. Jurisdictions used a range of reports to locate discrepant scores and marking patterns. Following check-marking, centre leaders were open to several courses of action informed by national marking protocols (see Table 42 below).


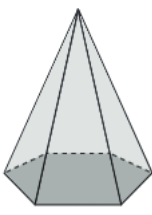
Table 42. National marking protocols

	Monitor	Discuss/ Re-train	Negotiate future marking
Total score	3 – 4 points discrepant	5 – 8 points discrepant	If 5 or more points discrepant on 3 occasions after retraining OR More than 8 points discrepant on 2 occasions
Criterion score	2 points discrepant	2 points discrepant on 3 or more occasions OR	If 2 or more points discrepant on 3 occasions after retraining

	Monitor	Discuss/ Re-train	Negotiate future marking
		3 or more points discrepant on 1 occasion	
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

Example items in reporting bands

Table 43. Numeracy example items in reporting bands

Band	NAPLAN scale score	Item	Key / key string
1	270	<p>7 Kay has saved \$3247 for a holiday. She spends \$2000 on airfares. How much of her savings does Kay have left?</p> <p>\$5247 \$3227 \$3047 \$1247</p> <p><input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/></p>	D
2	322	<p>6 Ning has this money in her money box.</p>  <p>In total, how much money does she have in her money box?</p> <p>\$2.15 \$6.10 \$6.60 \$7.10</p> <p><input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/></p>	D
3	374	<p>14 The base of this pyramid is in the shape of a hexagon.</p>  <p>How many faces of the pyramid are triangles?</p> <p>3 4 5 6 7</p> <p><input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/></p>	D

Band	NAPLAN scale score	Item	Key / key string														
4	426	<p>19 This table shows the number of students who prefer different after-school activities.</p> <table border="1"> <thead> <tr> <th rowspan="2">Activity</th> <th colspan="2">Number of students</th> </tr> <tr> <th>Girls</th> <th>Boys</th> </tr> </thead> <tbody> <tr> <td>Play computer games</td> <td>5</td> <td>3</td> </tr> <tr> <td>Play sport</td> <td>8</td> <td>10</td> </tr> <tr> <td>Read books</td> <td>4</td> <td>6</td> </tr> </tbody> </table> <p>How many more students prefer to read books than to play computer games?</p> <input type="text"/>	Activity	Number of students		Girls	Boys	Play computer games	5	3	Play sport	8	10	Read books	4	6	2
Activity	Number of students																
	Girls	Boys															
Play computer games	5	3															
Play sport	8	10															
Read books	4	6															
5	478	<p>14 Bindi takes the ferry from Darwin to Bathurst Island. She leaves Darwin at 11:15 in the morning and arrives at Bathurst Island at 1:45 in the afternoon.</p> <p>How long did Bindi take to get from Darwin to Bathurst Island?</p> <p> <input type="radio"/> 2 hours and 30 minutes <input type="radio"/> 2 hours and 45 minutes <input type="radio"/> 3 hours and 30 minutes <input type="radio"/> 3 hours and 45 minutes </p>	A														
6	530	<p>10 In Devonport, there are 30 604 people. Each day, the average person uses 173 litres of water.</p> <p>Which of these gives the best estimate for the total number of litres of water used in Devonport each day?</p> <p> <input type="radio"/> $30\,000 \times 200$ <input type="radio"/> $30\,000 \times 100$ <input type="radio"/> $30\,000 \div 200$ <input type="radio"/> $30\,000 \div 100$ </p>	A														
7	582	<p>4 In 2017, workers at an office recorded the amount of paper they each recycled.</p> <ul style="list-style-type: none"> The office had 40 workers. Each worker recycled 50 kilograms of paper. Every 1000 kilograms of recycled paper saves 24 trees. <p>In total, how many trees did these workers save in 2017?</p> <input type="text"/>	48														

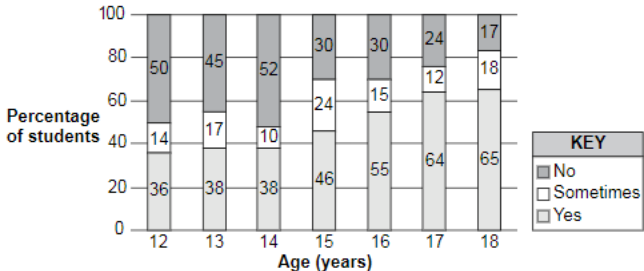

Band	NAPLAN scale score	Item	Key / key string																																
8	634	<p>35 Students at a high school were surveyed to find whether they slept with a phone near their bed.</p> <p>The graph below shows the results.</p>  <table border="1"> <caption>Data for Question 35: Percentage of students by age and response</caption> <thead> <tr> <th>Age (years)</th> <th>No (%)</th> <th>Sometimes (%)</th> <th>Yes (%)</th> </tr> </thead> <tbody> <tr> <td>12</td> <td>50</td> <td>14</td> <td>36</td> </tr> <tr> <td>13</td> <td>45</td> <td>17</td> <td>38</td> </tr> <tr> <td>14</td> <td>52</td> <td>10</td> <td>38</td> </tr> <tr> <td>15</td> <td>30</td> <td>24</td> <td>46</td> </tr> <tr> <td>16</td> <td>30</td> <td>15</td> <td>55</td> </tr> <tr> <td>17</td> <td>24</td> <td>12</td> <td>64</td> </tr> <tr> <td>18</td> <td>17</td> <td>18</td> <td>65</td> </tr> </tbody> </table> <p>There were 150 12-year-old students at the high school.</p> <p>How many 12-year-old students responded 'No'?</p> <p> <input type="radio"/> 21 <input type="radio"/> 50 <input type="radio"/> 54 <input type="radio"/> 75 <input type="radio"/> 100 </p>	Age (years)	No (%)	Sometimes (%)	Yes (%)	12	50	14	36	13	45	17	38	14	52	10	38	15	30	24	46	16	30	15	55	17	24	12	64	18	17	18	65	D
Age (years)	No (%)	Sometimes (%)	Yes (%)																																
12	50	14	36																																
13	45	17	38																																
14	52	10	38																																
15	30	24	46																																
16	30	15	55																																
17	24	12	64																																
18	17	18	65																																
9	686	<p>33 At the entrance to a harbour there are two lights.</p> <p>A red light flashes every 5 seconds.</p> <p>A green light flashes every 7 seconds.</p> <p>The red light and the green light both flash together at 7:00 am.</p> <p>How many more times will the lights both flash at the same time in the next 3 minutes?</p> <input type="text"/>	5																																
10	738	<p>38 Suki makes a regular hexagon from six identical triangular tiles.</p> <p>Each tile has an area of 3.9 cm^2.</p>  <p>Suki then adds more tiles to make a hexagon with double the side length of this hexagon.</p> <p>What will be the area of this larger hexagon?</p> <p> <input type="radio"/> 7.8 cm^2 <input type="radio"/> 23.4 cm^2 <input type="radio"/> 46.8 cm^2 <input type="radio"/> 93.6 cm^2 </p>	D																																

Table 44. Reading example items in reporting bands

Dingle's game

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

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

'I bags his front legs,' called Michael.

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

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

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

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



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

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

A great southern secret—*two views*

View 1

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

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

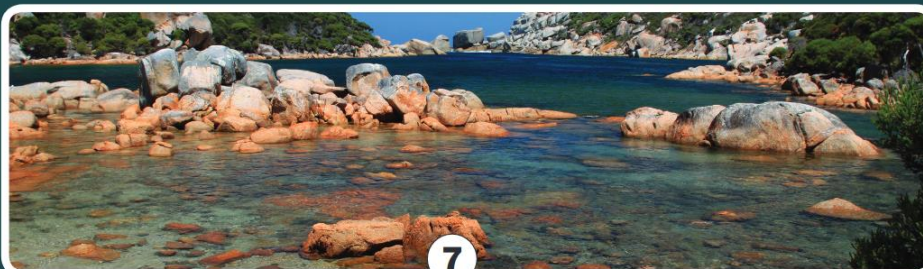
The Waychinicup is one of the few rivers on the south coast not to have a sand bar, and on either side of the river the steep slopes are carpeted in thick impenetrable coastal scrub. Scattered across this carpet rear enormous, smooth, bone-coloured boulders, so inexplicably smooth, they are like finely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

View 2

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

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

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



7

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

Band	NAPLAN scale score	Stimulus text	Item
10	727	A great southern secret – two views x00074163	In View 1, what is the main point of contrast between the river and the sea? <input checked="" type="radio"/> sound <input type="radio"/> depth <input type="radio"/> colour <input type="radio"/> beauty


Table 45. Grammar and punctuation example items in reporting bands

Band	NAPLAN scale score	Item	Key / key string
1	215	Place the correct word in the box to complete this sentence. <div style="border: 1px solid #ccc; padding: 5px; margin: 5px 0;"> or so for </div> I like baking cakes but I do not like cleaning up afterwards.	but
2	283.1	Place the correct ending in the box to complete this sentence. <div style="border: 1px solid #ccc; padding: 5px; margin: 5px 0;"> swimming with friends if she has time because it is hot </div> Every day after school, Jill helps her dad .	Jill helps her dad
3	328.8	Choose the word that describes how the man walked. <div style="border: 1px solid #ccc; padding: 5px; margin: 5px 0;"> Slowly the old man walked down the hall and then wearily climbed into bed. </div>	Slowly

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

Band	NAPLAN scale score	Item	Key / key string																									
8	618	<p>Choose one checkbox in each row of the table to show the correct word class for each word taken from this sentence.</p> <p>The chilly wind blows wildly.</p> <table border="1" data-bbox="480 450 1310 757"> <thead> <tr> <th></th> <th>adverb</th> <th>adjective</th> <th>verb</th> <th>noun</th> </tr> </thead> <tbody> <tr> <td>chilly</td> <td><input type="checkbox"/></td> <td><input checked="" type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> <tr> <td>wind</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input checked="" type="checkbox"/></td> </tr> <tr> <td>blows</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input checked="" type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> <tr> <td>wildly</td> <td><input checked="" type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> </tbody> </table>		adverb	adjective	verb	noun	chilly	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	wind	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	blows	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	wildly	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
	adverb	adjective	verb	noun																								
chilly	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																								
wind	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>																								
blows	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>																								
wildly	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																								
9	655	<p>Place the correct punctuation mark in each sentence.</p> <div data-bbox="488 871 1318 943" style="border: 1px solid #ccc; padding: 5px; margin: 10px 0;"> : ; </div> <p>Rover lost his collar <input type="text" value=";"/> he was swimming in the dam.</p> <p>Our fitness has improved <input type="text" value=";"/> it has taken many hours of training.</p> <p>I have finally learnt the secret to success <input type="text" value=":"/> believe in yourself.</p> <p>I love everything Dad cooks <input type="text" value=":"/> steak, pizza and chicken pasta.</p>																										
10	731.2	<p>Which adverb in this sentence describes when an action happens?</p> <p>Henry arrived early for training, dropped his bag hurriedly and ran quickly to the oval where his coach was waiting patiently for the rest of the team.</p>	early																									


Table 46. Spelling items in bands

Band	NAPLAN scale score	Item	Key / key string
1	256.0	<p>They were giving out apples for _____.</p> <p>Click on the play button to hear the missing word.</p>  <p>Type the correct spelling of the word in the box.</p> <div data-bbox="486 779 1045 884" style="border: 1px solid #ccc; padding: 5px; width: fit-content;"> <p>free</p> </div>	free
2	325.7	<p>The spelling mistake in this sentence is underlined.</p> <p>The toy began to spinn around.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div data-bbox="499 1227 1109 1339" style="border: 1px solid #ccc; height: 50px; width: 100%;"></div>	spin
3	362.6	<p>The spelling mistake in this sentence is underlined.</p> <p>He kickd the football through the goals.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div data-bbox="493 1715 1102 1827" style="border: 1px solid #ccc; padding: 5px; width: fit-content;"> <p>kicked</p> </div>	kicked

Band	NAPLAN scale score	Item	Key / key string
4	398.2	<p>The spelling mistake in this sentence is underlined.</p> <p>A dog is much bigga than a mouse.</p> <p>Type the correct spelling of the underlined word in the box.</p> <input data-bbox="501 629 1107 741" type="text"/>	bigger
5	430.0	<p>The spelling mistake in this sentence is underlined.</p> <p>One rool in our class is to raise your hand to ask for help.</p> <p>Type the correct spelling of the underlined word in the box.</p> <input data-bbox="494 1070 1075 1182" type="text"/>	rule
6	516.6	<p>The spelling mistake in this sentence is highlighted.</p> <p>The children saved the day and were heros .</p> <p>Type the correct spelling of the highlighted word in the box.</p> <input data-bbox="494 1509 1107 1621" type="text"/>	heroes

Band	NAPLAN scale score	Item	Key / key string
7	534.3	<p>The spelling mistake in this sentence is highlighted.</p> <p>A rock band often has a <u>gitar</u> player.</p> <p>Type the correct spelling of the highlighted word in the box.</p> <div data-bbox="501 638 1099 748" style="border: 1px solid black; padding: 5px; margin: 10px 0;">guitar</div>	guitar
8	611.2	<p>There is one spelling mistake in this sentence.</p> <p>The students had a very efficiant method for completing their homework.</p> <p>Type the correct spelling of the word in the box.</p> <div data-bbox="486 1032 938 1115" style="border: 1px solid black; padding: 5px; margin: 10px 0;">efficient</div>	efficient
9	654.6	<p>There is one spelling mistake in this sentence.</p> <p>The performance was given spontaineous applause.</p> <p>Type the correct spelling of the word in the box.</p> <div data-bbox="486 1451 1109 1565" style="border: 1px solid black; padding: 5px; margin: 10px 0;">spontaneous</div>	spontaneous
10	716.3	<p>The spelling mistake in this sentence is underlined.</p> <p>The mouse was a <u>nuscence</u> when it chewed through the electricity cord.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div data-bbox="496 1832 944 1915" style="border: 3px double black; padding: 5px; margin: 10px 0;">nuisance</div>	nuisance

Table 47. Example writing prompt

YEAR 3 AND YEAR 5	
<h2 style="text-align: center;">Following tracks</h2> <p>Write a narrative (story) about footprints, tracks or a trail.</p> <p>The tracks in your story may be left by a person, an animal, a vehicle or something odd.</p> <p>Perhaps the tracks are clues or lead your characters to something exciting or difficult.</p> <p>You can use an idea on this page or you can use your own idea about following tracks.</p> <p>Think about:</p> <ul style="list-style-type: none"> • the characters and where they are • the complication or problem to be solved • how the story will end. <p>Remember to:</p> <ul style="list-style-type: none"> • plan your story before you start • choose your words carefully • write in sentences • pay attention to your spelling, punctuation and paragraphs • check and edit your writing. 	

Setting branching rules

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

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

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

intersect at the point that will route 50 per cent of students to each of these testlets, which was the mean of the student ability distribution.

However, the student ability and item difficulty means are not always aligned; therefore, in translating the intersection of the test information curves on to the student ability scale, care was taken to account for such mistargeting. The investigation showed that the empirical distributions of the ability estimates did not differ significantly across year level and domains, when the measurement scale was case-centred within year level (that is, when the mean of student ability was set to zero). Consequently, the same set of item difficulty estimates for NAPLAN online testlets could be used across year levels for the grammar and punctuation, numeracy and reading domains. The final testlet boundaries and parameters were developed and empirically investigated in a series of simulations to establish feasibility and robustness of such overall NAPLAN online test parameters for reading and numeracy tests.

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

Branching rules for numeracy, reading and grammar and punctuation tests

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

All students at each year level and domain started with testlet in node A (Stage 1). Once this testlet was completed, a decision was made to branch a student to either an easier testlet (node B) or a harder testlet (node D), which was the *first branching point*. Assuming that a student was sent to a testlet in node D and completed this testlet, then another decision was made to branch this student to a testlet in node C (low complexity items), a testlet in node E (items with average complexity) or a testlet in node F (high complexity items), which was the *second branching point*. If a student was branched to node E, pathway ADE (shown in Figure 7) was completed. As discussed earlier, students with very low performance on testlet in node A were first assigned the easiest testlet in node C as a second testlet before finally being assigned testlet B as the third testlet (pathway ACB). This allowed low-performing students to demonstrate their knowledge with items that matched their test performance and to engage more efficiently through the test.

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

The information functions for testlets in nodes C, B and D are illustrated in Figure 9. As this figure shows, the peak of the information function for testlets in nodes B and D was about -1 and 1 logits, respectively. This means that the items were allocated to B and D so that D was more suited to more able students and B was more suited to less able students. In fact, given that the curves intersect at about 0.0 logits, these information functions show that if a student's ability was below 0.0 logits, then testlet B was expected to work best for them; whereas if a student's ability was above 0.0 logits, then testlet D was expected to work best for them. Similarly, this figure shows that testlet C (green curve) provides more information for students with an ability less than -1.5 logits. Given that the testlets C and B curves intersect at about -1.6 logits, if a student's ability was below -1.6 logits, then testlet C was expected to work best for that student.

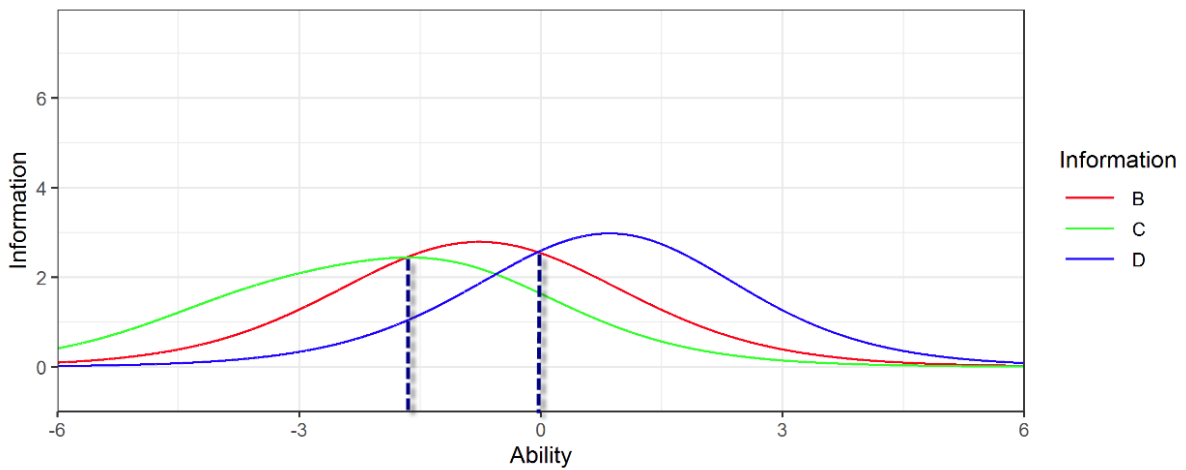


Figure 9. Test information functions: curves for testlets C, B and D

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

Figure 10 illustrates how the test characteristic curve for one testlet (in node A) can be used to find the raw scores that correspond to the cut-points between testlet information functions. The test characteristic curve for testlet A is shown on the same axis as the information functions for testlets C, B and D. If a student has a raw score of 4 or less on testlet A, then their ability estimate is in a region for which testlet C provides most precision; whereas if a student has a raw score greater than 4 and less than 9 on testlet A, then their ability estimate is in a region for which testlet B provides most precision. Similarly, students with a raw score of 9 or more will be assigned testlet D that provides most precision.

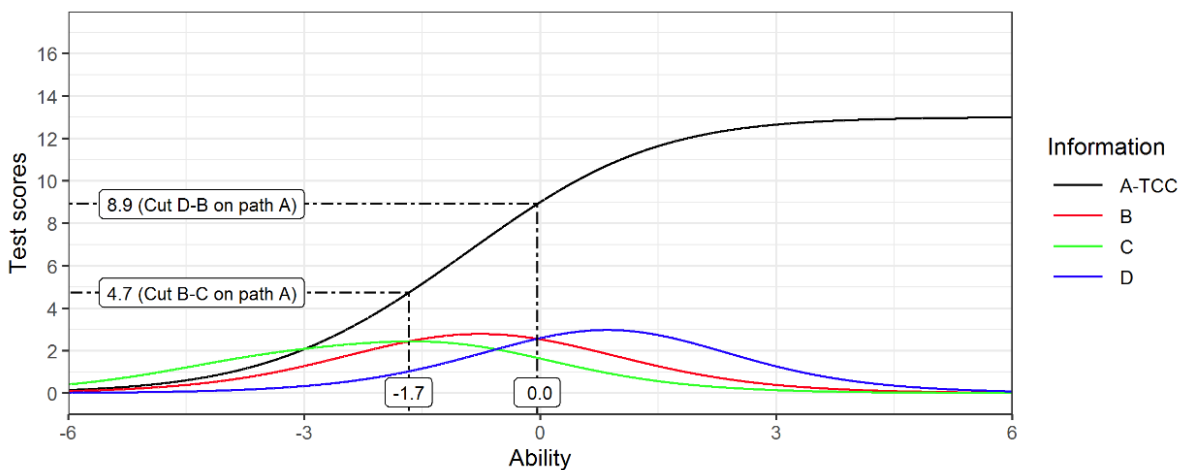


Figure 10. Stage 1. Testlet A-C|B|D cut scores

The branching rules for the first branching point discussed above are presented in Table 48 .

Table 48. Stage 1 cut scores (Testlet A to C|B|D)

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

The same approach was taken to set the rules (cut scores) for the second branching point (Figure 11 and Table 49).

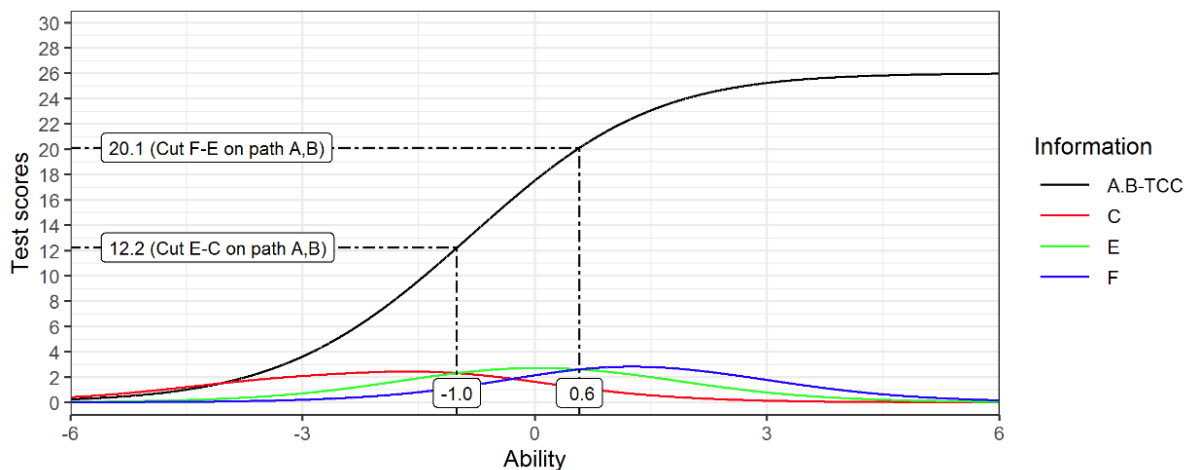


Figure 11. Stage 2. Testlet AB-C|E|F cut scores

In Figure 11, the test characteristics curve for testlet AB is shown on the same axis as the information functions for testlets C, E and F. If a student had a cumulative raw score of 12 or less on testlets A and B, then their ability estimate was in a region for which testlet C provided most precision; whereas if a student had a cumulative raw score greater than 12 but less than 21 on testlets A and B, then their ability estimate was in a region for which testlet E provided most precision. Finally, students with a cumulative raw score of 21 or more were assigned Testlet F, which was designed for high-performing students. The branching rules for the second branching point after students completed testlets A and B are presented in Table 49.

Table 49. Stage 2 cut scores (testlet AB to C|E|F)

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

In Figure 12, the test characteristics curve for testlet AD is shown on the same axis as the information functions for testlets C, E and F. If a student had a cumulative raw score of 8 or less on testlets A and D, then their ability estimate was in a region for which testlet C provided most precision; whereas if a student had a cumulative raw score greater than 8 but less than 17 on testlets A and D, then their ability estimate was in a region for which Testlet E provided most precision. Finally, students with a cumulative raw score of 17 or

more were assigned Testlet F, which contained the most challenging items. The branching rules for the second branching point after students completed testlets A and D are presented in Table 50.

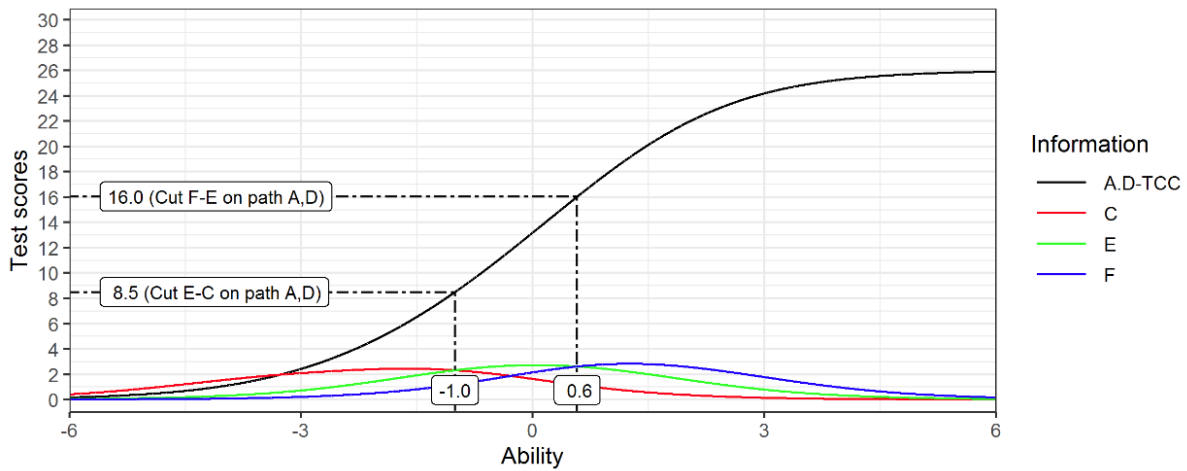


Figure 12. Stage 2. Testlet AD-C|E|F cut scores

Table 50. Stage 2 cut scores (testlet AD-C|E|F)

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

Branching rules for spelling

The right-hand side of Figure 8 illustrates a three-stage tailored test design (1–2–2) for spelling with one testlet in Stage 1, two testlets in Stage 2, and two testlets in Stage 3. These five testlets formed four pathways (SA–SD–PD, SA–SD–PB, SA–SB–PD, SA–SB–PB).

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

Figure 13 shows that two decisions were made before branching students to the final stage in the multistage tailored tests: 1) after completion of testlet SA, and 2) after completion of testlets SA–SB or SA–SD. These decisions were made before the multistage test was administered. The same rationale, applied to setting branching rules for reading and numeracy tests, was utilised in spelling. The branching rules for spelling are illustrated in Figure 13, Figure 14 and Figure 15.

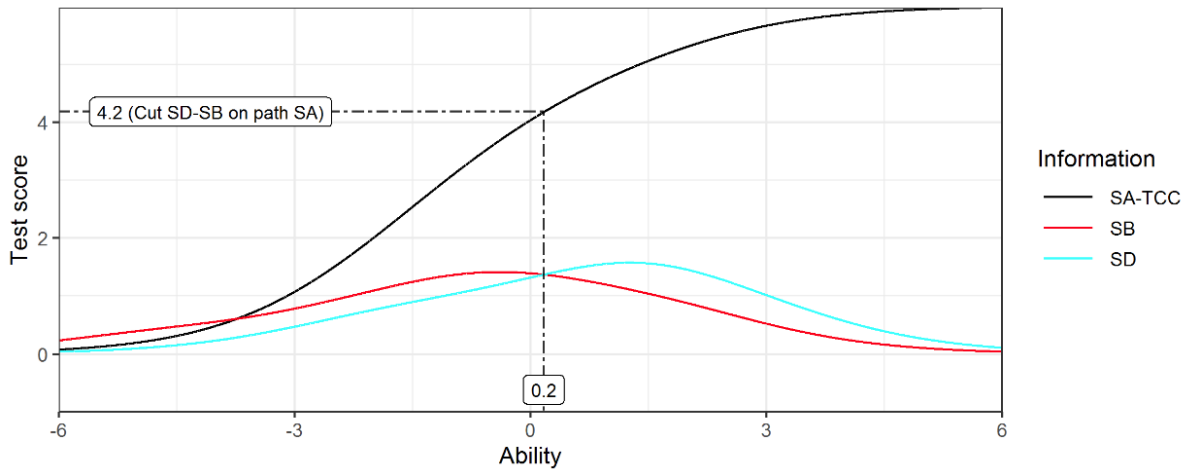


Figure 13. Stage 1. Testlet SA–SB|SD cut scores

In Figure 13, the test characteristics curve for testlet SA is shown on the same axis as the information functions for testlets SB and SD. If a student had a raw score of 4 or less on testlet SA, then their ability estimate was in a region for which testlet SB provided most precision; whereas if a student had a raw score greater than 4 on testlet SA, then their ability estimate was in a region for which testlet SD provided most precision. The branching rules for the first branching point in spelling is presented in Table 51.

Table 51. Stage 1, Testlet SA–SB|SD cut scores

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

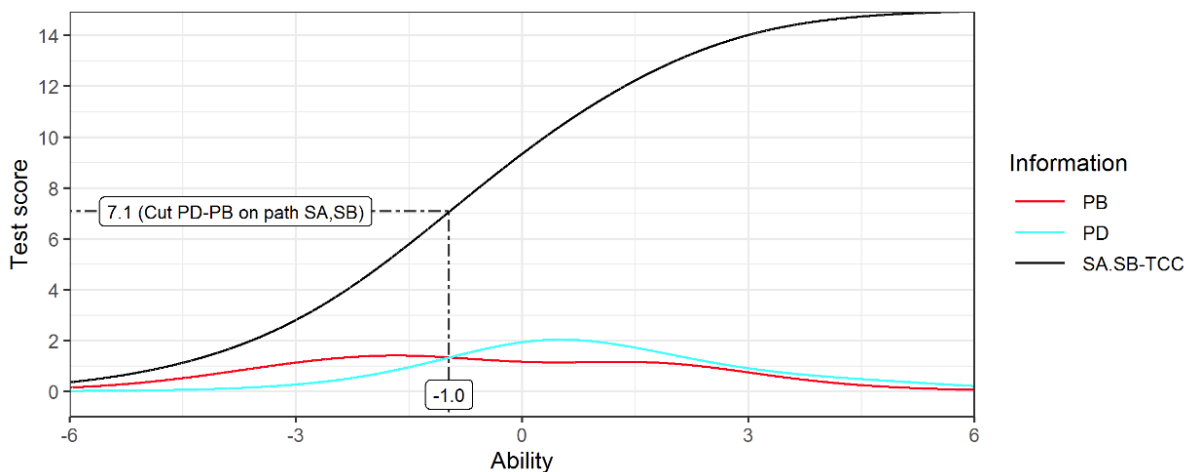


Figure 14. Stage 2. Testlet SA–SB to PB|PD cut scores

In Figure 14, the test characteristics curve for testlet SA–SB is shown on the same axis as the information functions for testlets PB and PD. If a student had a cumulative raw score of 7 or less on testlets SA and SB, then their ability estimate was in a region for which testlet PB provided most precision; whereas if a student had a cumulative raw score greater than 7 on testlets SA and SB, then their ability estimate was in a region for which

testlet PD provided most precision. The branching rules for the second branching point in spelling is presented in Table 52.

Table 52. Stage 2, Testlets SA–SB to PB|PD cut scores

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

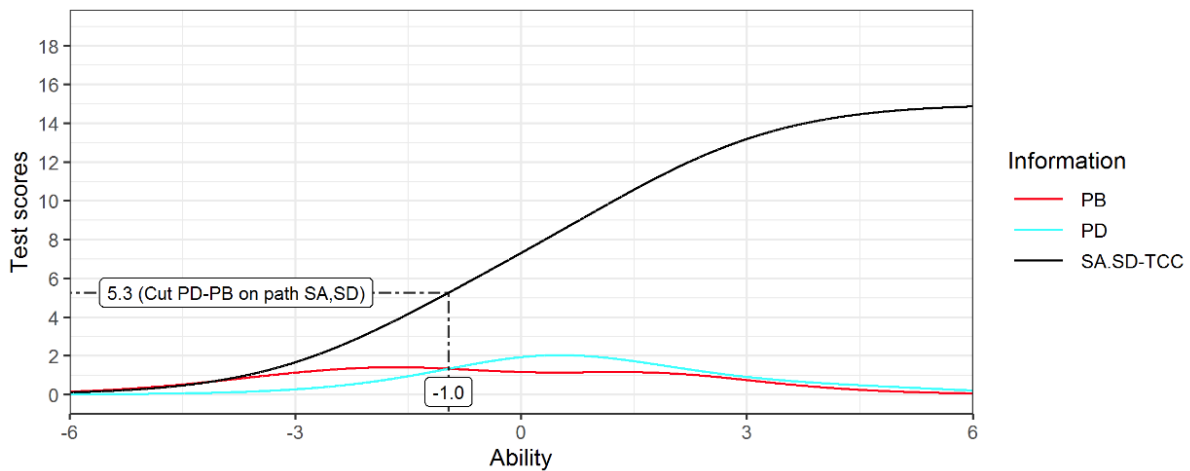


Figure 15. Stage 2, Testlets SA–SD to PB|PD cut scores

In Figure 15, the test characteristics curve for testlet SA–SD is shown on the same axis as the information functions for testlets PB and PD. If a student has a cumulative raw score of 5 or less on testlets SA and SD, then their ability estimate is in a region for which testlet PB provides more precision; whereas if a student has a cumulative raw score greater than 5 on testlets SA and SD, then their ability estimate is in a region for which testlet PD provides more precision. The branching rules for the second branching point in spelling is presented in Table 53.

Table 53. Stage 2, Testlet SA–SD to PB|PD cut scores

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

Pathway utilisation

This section describes how different pathways were utilised in NAPLAN 2021 online tests using Year 3 numeracy as an example. The results for other year levels and domains are presented in Appendix A.

The percentage of students assigned to each pathway, and ability distributions at each stage for Year 3 numeracy are shown in Figure 16 and Figure 17.

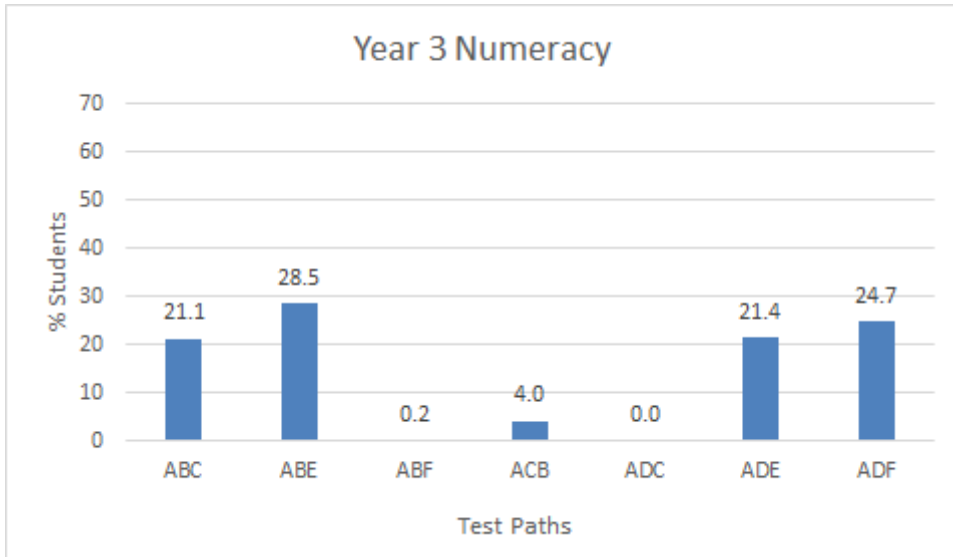


Figure 16. Percentage of students assigned to each pathway in Year 3 numeracy

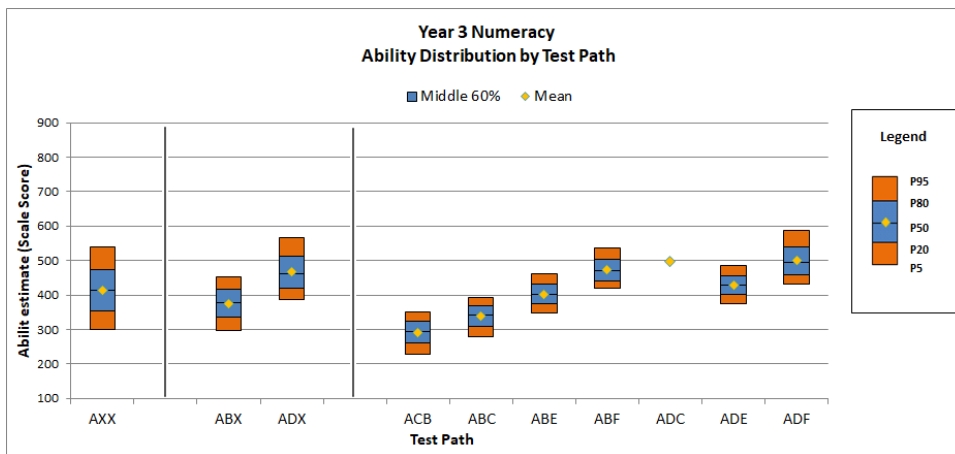


Figure 17. Ability distribution by pathway for Year 3 numeracy

As Figure 16 shows, the ideal separation of the testlet information curves for testlets B and D has been achieved thus approximately 50 per cent of students have been sent to each of these two testlets. The number of students assigned to each path varied from 0 per cent for ADC pathway to approximately 29 per cent in ABE pathway. To some extent, this was expected since, for example, going through the ADC pathway would require high performance on testlet A followed by very poor performance on testlet D. Similarly, a very low percentage (0.2) for ABF pathway was expected since it would require low performance on testlet A followed by high performance on testlet B. This chart also shows only 4 per cent students were sent to Testlet C immediately after completing Testlet A.

Ability distributions by pathway are illustrated in Figure 17. Patterns of ability distributions across pathways were roughly as expected. That is, students ending in testlet F had the highest ability distribution and students who were administered testlet C, immediately after completing Testlet A (ACB), had the lowest ability distributions. Furthermore, the ability distribution in second stage shows that high- and low-performing students were sent to testlet D and testlet B, respectively. Figure 17 also shows that pathways overlapped in abilities.

Chapter 4: Data collection and preparation

This chapter describes data collection and delivery, data validation and data preparation for NAPLAN 2021. The first part of the chapter focuses on how data for paper and online tests are collected by test administration authorities (TAAs) from each jurisdiction and delivered to ACARA. The second part of the chapter describes how data are validated and prepared by the contractor before performing the analysis.

Data collection and delivery

Test administration authorities (TAAs) are responsible for:

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

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

The purpose of the Data Standards Manual: Student Background Characteristics⁸ is to provide guidance to schools and school systems in the collection of information on student background characteristics, using the nationally agreed standard measures of the characteristics. The manual is to be used by schools and school systems when enrolling students for the first time in the school year, or when collecting information, via special data collection forms, on those students participating in national assessments.

The nationally agreed student background characteristics collected are:

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

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

- staggered delivery of online test data including both scored and raw response data (used for item calibration)
- delivery of the merged paper-based horizontal equating data from equating samples from the jurisdictions by domain for reading, spelling, grammar & punctuation, and numeracy for both paper and online schools (used for horizontal equating)
- delivery of the second version of the Student Master File (SMF) and Item Response File (paper calibration sample for those jurisdictions that sat NAPLAN tests on paper and online).

⁸ www.acara.edu.au/reporting/data-standards-manual-student-background-characteristics

- delivery of the fourth version of the SMF, IRF and online test data (NAEs), previously called stage 1 census data, for analysis to produce the NAPLAN 2021 summary results.
- delivery of the final SMF / IRF / NAEs, previously called stage 2 complete census data, to produce the NAPLAN 2021 National Report

Paper tests

Data collection for paper tests was undertaken by the test administration authorities (TAAs) in the jurisdictions. There were three rounds of data delivery for the central data analysis and a final round for the preparation of the national report. The first round involved delivery of data from the *equating samples* and the second round involved the delivery of the second version of SMF / IRF (paper calibration samples). The third round involved the fourth delivery of the SMF/IRF, nearly complete stage 1 full cohort NAPLAN paper-based test data and mixed mode data of Years 3, 5, 7 and 9 students in mid-July 2021. These data were used for the generation of the NAPLAN 2021 summary results. The complete (with background data) full cohort data used for the production of the national report were delivered in September 2021. With each round of data delivery, ACARA has performed a comprehensive quality assurance on the data and provided TAAs with an exception report. TAAs then resolved all the issues included in the exception report and resubmitted their data, if required. The datasets then were cleaned and sent to the contractor for analysis. A systematic process involving data checking was used to ensure that each dataset was consistent with national code frames and data dictionaries. There are several types of exception rules implemented in the NAPLAN QA scripts such as structural, show-stopper, advisory, statistical etc. A sample of the exception rules is included in Appendix M.

Online tests

The Education Services Australia (ESA) managed the online national assessment platform (platform) on which the NAPLAN 2021 online tests were delivered. The Australian Council for Educational Research (ACER) received the online test data extracted from the platform directly from ACARA by domain as those became available. With the tight timeline between the online assessments and the delivery of school and student summary reports (SSSRs), quality assurance checks of online data extracted from the platform along with the SMF and IRF started in late May. The preparation for online data checking and management and for the analysis of online data followed the quality assurance check. Data integrity checking included verification that online data files conformed to their data dictionary and coding conventions (supplied by ACARA) and that item responses in the data files conformed to the valid codes specified in the code frames.

Data cleaning validation process

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

Data preparation

The recoding of test data was conducted prior to data analysis.

In 2021, responses to multiple-choice items were indicated by the number of the chosen response option for each item; that is, 1, 2, 3, 4, or 5. Responses for students not participating on a particular test or testlet were recoded to 'R' and treated as *not administered*. Multiple responses ('7') were treated as *incorrect*. Embedded missing responses were coded as '9' and treated as *incorrect*. Trailing missing responses were also coded as '9' for the first unanswered item and treated as *incorrect*, while the remaining trailing missing items were recoded as 'M' and treated as *not reached*. These not-reached items were treated as *not administered* items for item calibration to obtain an appropriate estimate of the item difficulty (for students who had a chance to respond). However, these not-reached responses were treated as *incorrect* for the final estimation of student abilities. In summary:

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

Data for partial-credit items were indicated by ordered categories starting with 0 up to the maximum possible value. Short-answer items were given scores of 0 or 1. The rules for data coding are provided in Table 54.

Table 54: Rules for data coding

Participation code	Data recoding rule
P – present	Data string (i.e. item responses) expected. Any embedded missing responses are indicated with a 9, invalid responses with a 7. The first trailing missing response is to be kept as a 9; subsequent trailing missing responses are retained as trailing-missing responses, and are to be recoded as an M. Any embedded missing responses within the data string are kept as a 9. Students who are present but do not attempt any question ('non-attempts') will have a string of Ms. Additionally, for the online tailored test data, responses for items in those testlets that were not administered to the students are coded as an R.
A – absent	A data string of all 8s for that test was expected from the TAA. Item response data are recoded as a string of Rs (this is like 'not-administered').
S – sanctioned abandonment	Response data are coded as an R. This is specifically used to indicate students who unexpectedly abandon the test due to illness or injury. See National Protocols for Test Administration, section 5.5.
W – withdrawn	A data string of all 8s for that test. See National Protocols for Test Administration, section 5.4. Response data are coded as an R.
E – exempt	A data string of all 8s for that test. See National Protocols for Test Administration, section 5.2. These students are not included in the calibration or in the calculation of means. Item data are recoded as a string of Rs.

Students who did not reach the last testlet 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 55. For example, students who did not attempt any numeracy or reading items were assigned pathway ACB. Students who only attempted some items in testlet A were assigned pathway ABE. Students who aborted the test testlet B or D during stage 2 were assigned testlet E in stage 3.

Table 55: Pathway assignment rules to incomplete online tests

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

Distribution of not reached items

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

Not reached items in online tests

Figure 18 to Figure 21 show the percentage of trailing missing responses by year level and test pathway in numeracy, reading, spelling and grammar & punctuation for the online tests. In these charts, the trailing missing responses were shown for one set of parallel testlets (for example, testlets A1 to F1 for numeracy, reading and grammar & punctuation, and testlets SA1 to PD1 for spelling). Across domains, grammar & punctuation had the lowest trailing missing rates. In numeracy and spelling, trailing missing responses started to appear from the third testlet of a test, and increased towards the end of a test. Across test paths, the most difficult test path A1-D1-F1 had the highest trailing missing rates in Years 5 and 7 numeracy. In spelling, the easiest test path SA1-SB1-PB1 had the highest trailing missing rates in Years 3, 5, 7 and 9. In Year 5, 7 and 9 reading, Year 9 Numeracy and Years 3, 5, 7 and 9 grammar & punctuation, the test path A1-C1-B1 had the highest trailing missing rates. This is consistent with students branching to the easiest testlet (C) from A and subsequently branching to a harder testlet (B). Similar patterns of trailing missing responses were found in other parallel testlets.

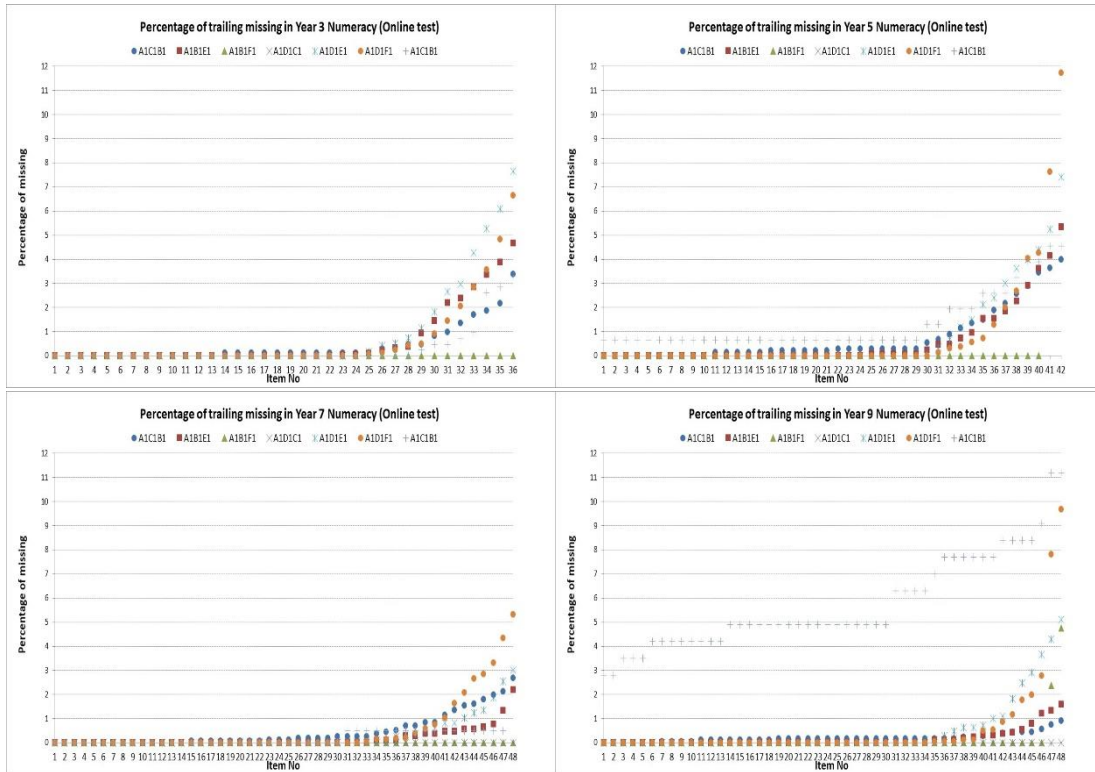


Figure 18. Trailing missing percentages in numeracy online test

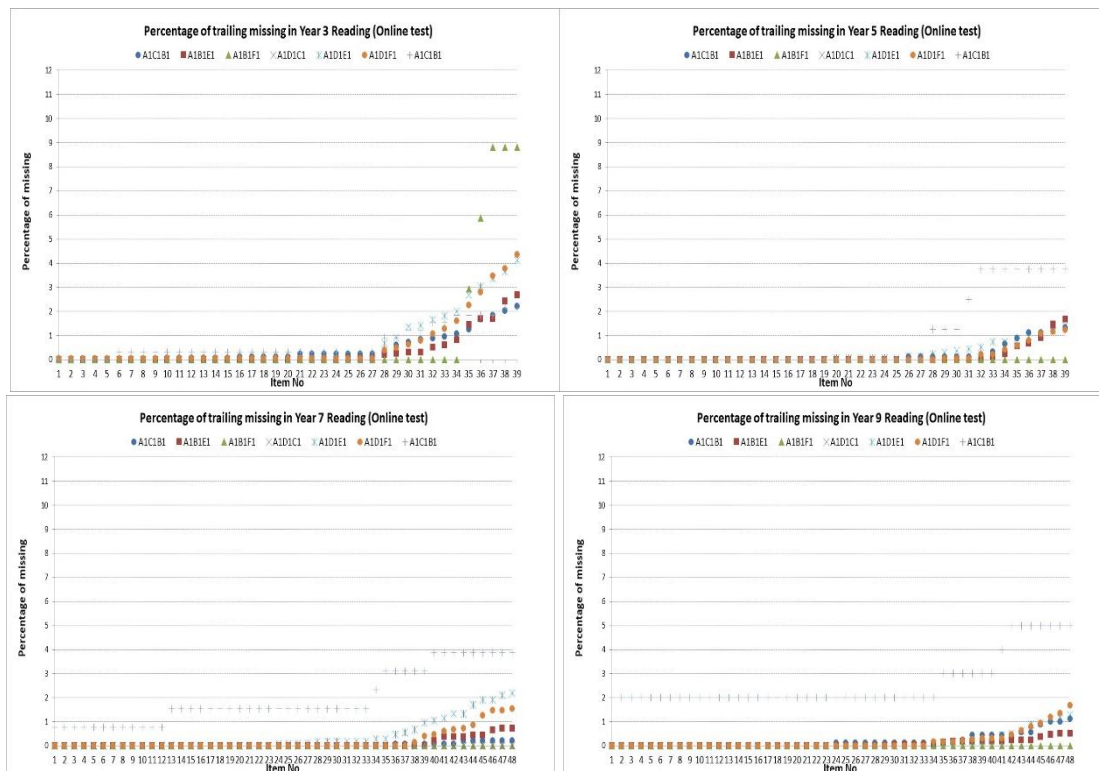


Figure 19. Trailing missing percentages in reading online test

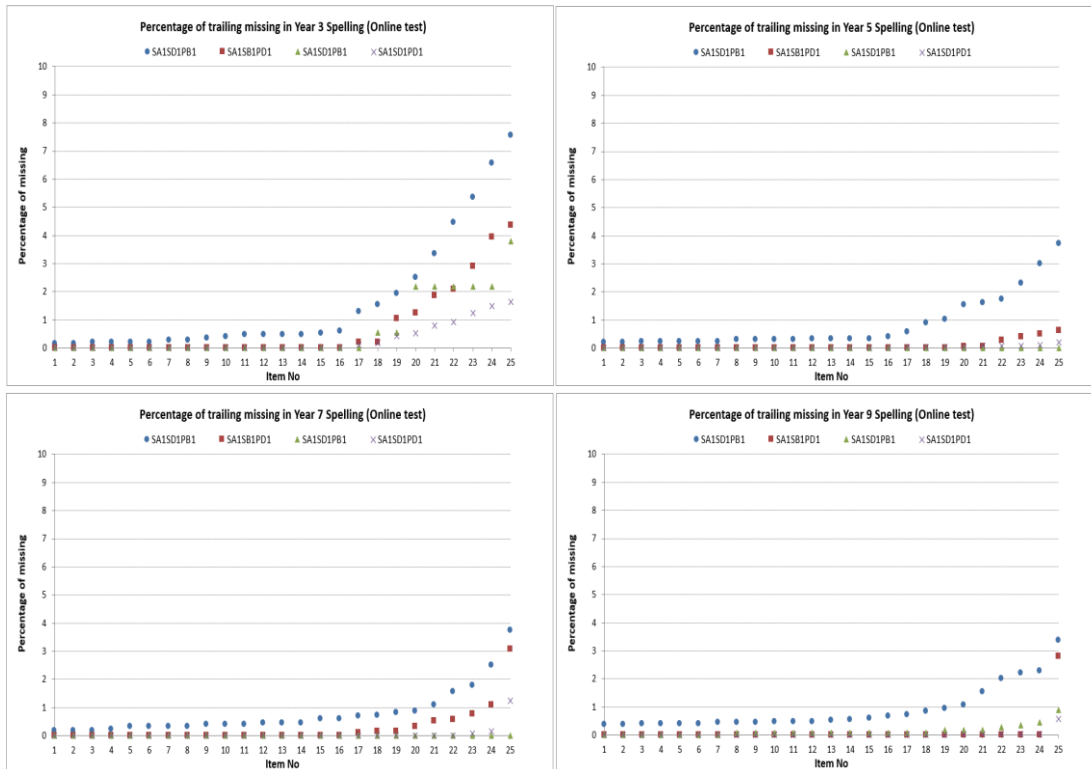


Figure 20. Trailing missing percentages in spelling online test



Figure 21. Trailing missing percentages in grammar & punctuation online test

Not reached items in paper tests

Figure 22 shows the percentage of trailing missing responses in each year level in numeracy, reading, spelling, and grammar & punctuation for the paper tests. It reveals that trailing missing responses started to appear around the middle of a test paper and increased towards the end of a test, as expected. Across domains, numeracy and spelling had the highest trailing missing rates, and grammar & punctuation had the lowest trailing missing rates. Within a domain, lower year levels tended to have a higher trailing missing rate, and higher grade levels tended to have lower trailing missing rates, except for Year 9 spelling. The proportions of trailing missing responses were all below 10 per cent except for the last item in year 5 numeracy, which suggests that the current test lengths for the paper test were appropriate. The last eight items in the numeracy Year 7 and Year 9 test papers were ‘non-calculator’ items, meaning that students were not permitted to use a calculator when responding to these items. No steep increase in the proportion of trailing missing responses was observed amongst the non-calculator items.

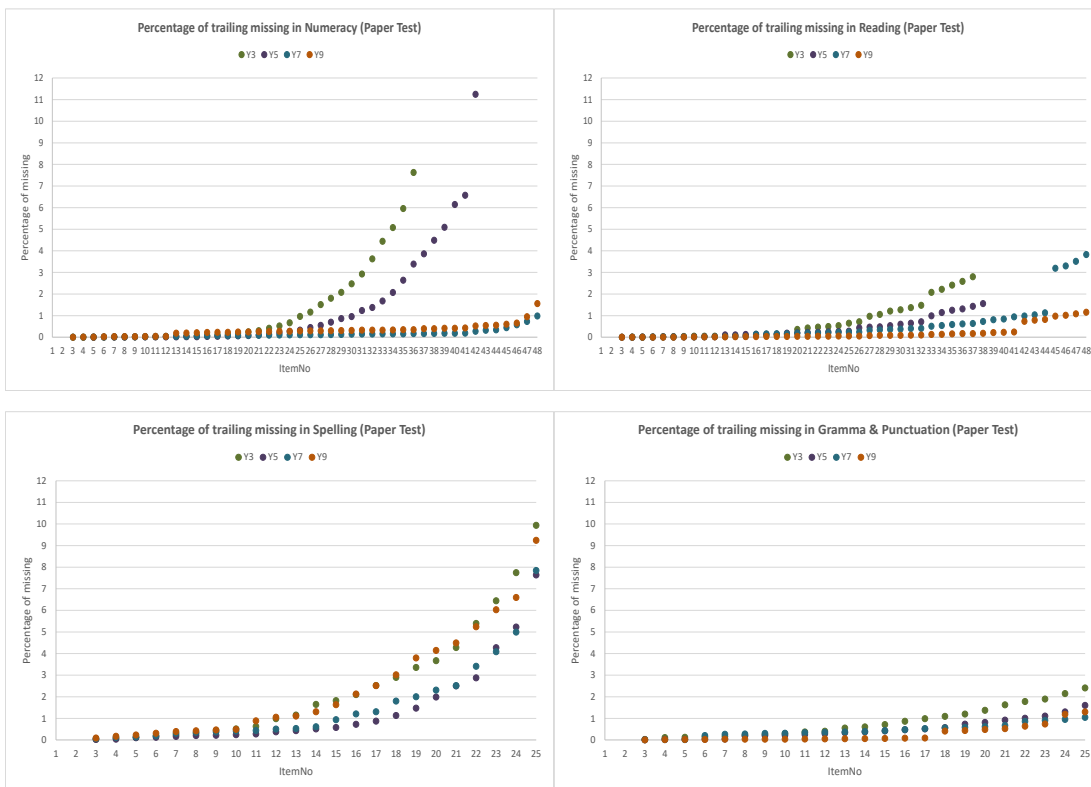


Figure 22. Trailing missing percentages in numeracy, reading, spelling and grammar & punctuation paper tests

Final student participation rates

Final student participation rates of NAPLAN 2021 are recorded in Table 56 below. The participation rate standard was 90 per cent at national and jurisdictional level to ensure unbiased population statistics. Results in the National Report were annotated if the response rate standard was not met. These percentages are coloured red in Table 56.

Table 56: Student participation rates

TAA	Year level	Numeracy (%)	Reading (%)	Spelling (%)	Grammar and punctuation (%)	Writing (%)
NSW	3	96.5	97.1	96.9	96.9	96.5
Vic.	3	94.2	95.0	94.7	94.7	94.0
Qld	3	92.4	93.3	92.9	92.9	92.6
WA	3	95.7	96.4	95.9	95.9	95.6
SA	3	94.5	95.2	94.8	94.8	94.0
Tas.	3	94.3	95.8	95.1	95.1	94.5
ACT	3	93.8	94.5	93.9	93.9	92.7
NT	3	81.9	83.1	82.6	82.6	88.2
Aus.	3	94.6	95.4	95.0	95.0	94.6
NSW	5	96.6	97.4	97.2	97.2	97.1
Vic.	5	94.4	95.3	95.1	95.1	95.0
Qld	5	92.1	93.4	92.9	92.9	93.0
WA	5	95.9	96.8	96.3	96.3	96.5
SA	5	94.1	95.3	94.6	94.6	95.1
Tas.	5	95.0	96.1	95.6	95.6	96.0
ACT	5	94.7	95.5	94.8	94.8	95.3
NT	5	81.4	82.6	81.6	81.6	83.8
Aus.	5	94.6	95.6	95.2	95.2	95.3
NSW	7	94.9	96.0	95.6	95.6	96.0
Vic.	7	93.2	94.5	94.3	94.3	94.2
Qld	7	88.0	89.7	89.0	89.0	89.8
WA	7	94.0	95.7	94.7	94.7	95.3
SA	7	93.0	94.5	93.5	93.5	93.9
Tas.	7	92.2	94.8	93.2	93.2	94.0
ACT	7	93.3	94.9	94.0	94.0	94.6
NT	7	78.9	80.8	79.3	79.3	80.7
Aus.	7	92.5	93.9	93.4	93.4	93.7
NSW	9	91.2	92.7	92.2	92.2	92.8
Vic.	9	88.9	90.4	90.4	90.4	90.4
Qld	9	80.9	82.9	82.4	82.4	83.2
WA	9	92.0	93.5	92.6	92.6	93.3
SA	9	88.4	90.2	88.8	88.8	89.8
Tas.	9	87.3	90.0	88.1	88.1	89.3
ACT	9	87.0	89.0	88.3	88.3	89.3
NT	9	70.5	73.3	70.3	70.3	73.2
Aus.	9	87.9	89.6	89.1	89.1	89.6

Chapter 5: Scaling methodology and outcomes

This chapter describes the processes and methodologies used in the NAPLAN 2021 central analysis, as well as the outcomes of the scaling analysis. The psychometrics and scaling methods used are methods that have been widely utilised in many large scale assessment programs, including the Programme for International Student Assessment (PISA).

Scaling model

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

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

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

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

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

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

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

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

Software used for analyses

For the Rasch scaling analysis, the software *ACER ConQuest 5* (Adams et al.; 2020). was used. *ACER ConQuest 5* provides tools for the estimation of a variety of item response models and regression models. It was used for test calibrations, for generating weighted likelihood estimates (WLEs) used for the score-equivalence tables, and for drawing plausible values (PVs) based on a multidimensional item response model with latent regression. The marginal maximum likelihood (MML) estimation method was used for test calibrations and for generating the plausible values. When calibrating items from multistage adaptive test designs, it has previously been shown that MML estimation produces unbiased estimates (Eggen & Verhelst, 2011; Adams & Lazendic; 2013).

Item calibration

Item response data for the online calibration was extracted as soon as data was collected for 50 per cent of students within each jurisdiction for all year levels. In total, the number of students included in the estimation of each domain was between 120,000 and 170,000 by year level. For the paper item calibration, a sample was drawn so that these schools could be prioritised for processing by New South Wales, Victoria, Queensland and Western Australia (other jurisdictions did not have sufficient paper schools to be include in the paper item calibration) and the analysis could commence before all paper data was collected. For each jurisdiction, a minimum of 3000 students were required by year level. The sample was broadly representative of the paper sub-population of each of these jurisdictions. The number of students in the calibration of the paper items within a domain ranged between 12,500 and 14,000 by year level.

For 2021 NAPLAN tests, the numeracy, reading, spelling, and grammar and punctuation tests were calibrated separately by domain, year level and test mode (e.g., either online or paper), resulting in 16 separate calibrations for each test mode. For each of the four non-writing online tests, items from all testlets within a domain and a year level were calibrated in a concurrent analysis.

For 2021 writing, the resulting scripts from students who responded on paper or online from different tasks were scored using the same marking rubric based on the ten criteria. The scored writing data from Years 3, 5, 7 and 9 were calibrated concurrently, based on the partial credit model with the latent distribution conditioned on year level, separately by test mode. The reason for the concurrent calibration was that some scores did not occur for some year levels. The calibration results were compared with parameters from previous NAPLAN cycles.

In the estimation of parameters, unreached-missing (M) and responses from an absent student (8, including *absent*, *withdrawn* and *exempt*) were treated as *not administered*, and embedded-missing (9) and invalid response (7 in paper tests) were treated as *incorrect* responses. Non-attempts (students who were present for the test but did not answer any items) have only Ms, no 9s. Online items that were not included in a student's pathway and therefore not presented to students (R) were treated as *not administered* in all analyses.

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

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

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

The student weight is equal to 1 for each students. This means for each jurisdiction, the sum of the senate weights was equal to the sum of the senate weight for the jurisdiction with the largest student population, NSW.

Given the small proportion of students participating in the paper tests, no weights were applied for the paper test calibration.

Review of test and item characteristics

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

Any summary statistics (e.g. Coefficient Alpha) shown at the end of the item analysis results for the online numeracy, reading, spelling, and grammar and punctuation tests are to be ignored as these were not for any one test form but were for the whole item pool at each year level. Traditional test reliability, quantified using the Coefficient Alpha internal consistency index, is presented at the end of the item analysis results for each of the paper-based tests.

The Rasch item parameter estimates and statistics are summarised in Appendix C for the online items in each of the 16 item pools for the numeracy, reading, spelling, and grammar and punctuation tests, and for each of the 16 paper tests (numeracy, reading, spelling and grammar & punctuation) across four year levels. The item parameters shown in these tables are delta-centred for each test (that is, the mean of item difficulties is set to zero). The 95 per cent confidence interval from *ACER ConQuest 5* output for the expected value of the infit mean square is also provided for each item.

Item Characteristic Curves (ICCs) for all items (online and paper-based) are shown in Appendix D. The ICC plot shows a comparison of the empirical ICC based on observations from 10 ability groupings (broken line joining 10 dots) and the expected model-based ICC (smooth line). Equal-distance grouping was used for each test node (generic testlet) for online tests with different ability range, and equal-size ability grouping was used for each paper tests. The two 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 10 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.

Test reliability

Table 57 shows the IRT-based reliabilities (WLE and EAP/PV) of each online test and each paper test.

For the online tests, the reliabilities were between 0.87 and 0.93 for the numeracy tests, between 0.81 and 0.90 for the reading tests, between 0.88 and 0.93 for the spelling tests, and between 0.77 and 0.84 for the grammar and punctuation tests. The reliabilities for the writing test were 0.95 and 0.91 for WLE reliability and EAP/PV reliability, respectively.

For the paper tests, the reliabilities were between 0.80 and 0.91 for the numeracy tests, between 0.79 and 0.87 for reading, between 0.83 and 0.90 for spelling, and between 0.67 and 0.76 for grammar and punctuation. The reliability for the writing test were the same as the online writing tests, 0.95 and 0.91 for WLE reliability and EAP/PV reliability, respectively. In general, the WLE reliability is higher than the EAP/PV reliability, and the reliability of online tests was somewhat higher than the reliability of the paper tests, except for the writing tests, where it was identical.

Table 57. Reliability (WLE) for NAPLAN 2021 paper tests

Test mode	Year level	Numeracy		Reading		Spelling		Grammar and punctuation		Writing*	
		WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV	WLE	EAP/PV
Online	3	0.90	0.87	0.90	0.87	0.93	0.92	0.84	0.83	0.95	0.91
	5	0.92	0.87	0.88	0.81	0.92	0.88	0.80	0.77		
	7	0.93	0.90	0.90	0.84	0.91	0.90	0.80	0.79		
	9	0.93	0.91	0.90	0.84	0.90	0.88	0.81	0.77		
Paper	3	0.87	0.80	0.86	0.81	0.90	0.85	0.75	0.71	0.95	0.91
	5	0.89	0.84	0.84	0.79	0.90	0.85	0.76	0.70		
	7	0.91	0.85	0.87	0.82	0.89	0.85	0.73	0.68		
	9	0.91	0.80	0.87	0.79	0.89	0.83	0.74	0.67		

*For Years 3, 5, 7 and 9 together

Test targeting and item spread

The purpose of the item-person map (or Wright map) is to compare the distribution of student locations (on the left side of the map) and the item thresholds (on the right side of the map). Item, step and person parameters are plotted on a common scale on a map. Appendix E provides the maps for each domain at each year level for the paper tests and online tests. It is important to note that for the online tests, 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 percent 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 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 23 displays the map for Year 3 numeracy online test. That map indicates that the current tests targeted the average numeracy achievement level of the student group quite well. The distribution of student abilities (each X represents approximately 279 students) matched up well with the distribution of item difficulties.

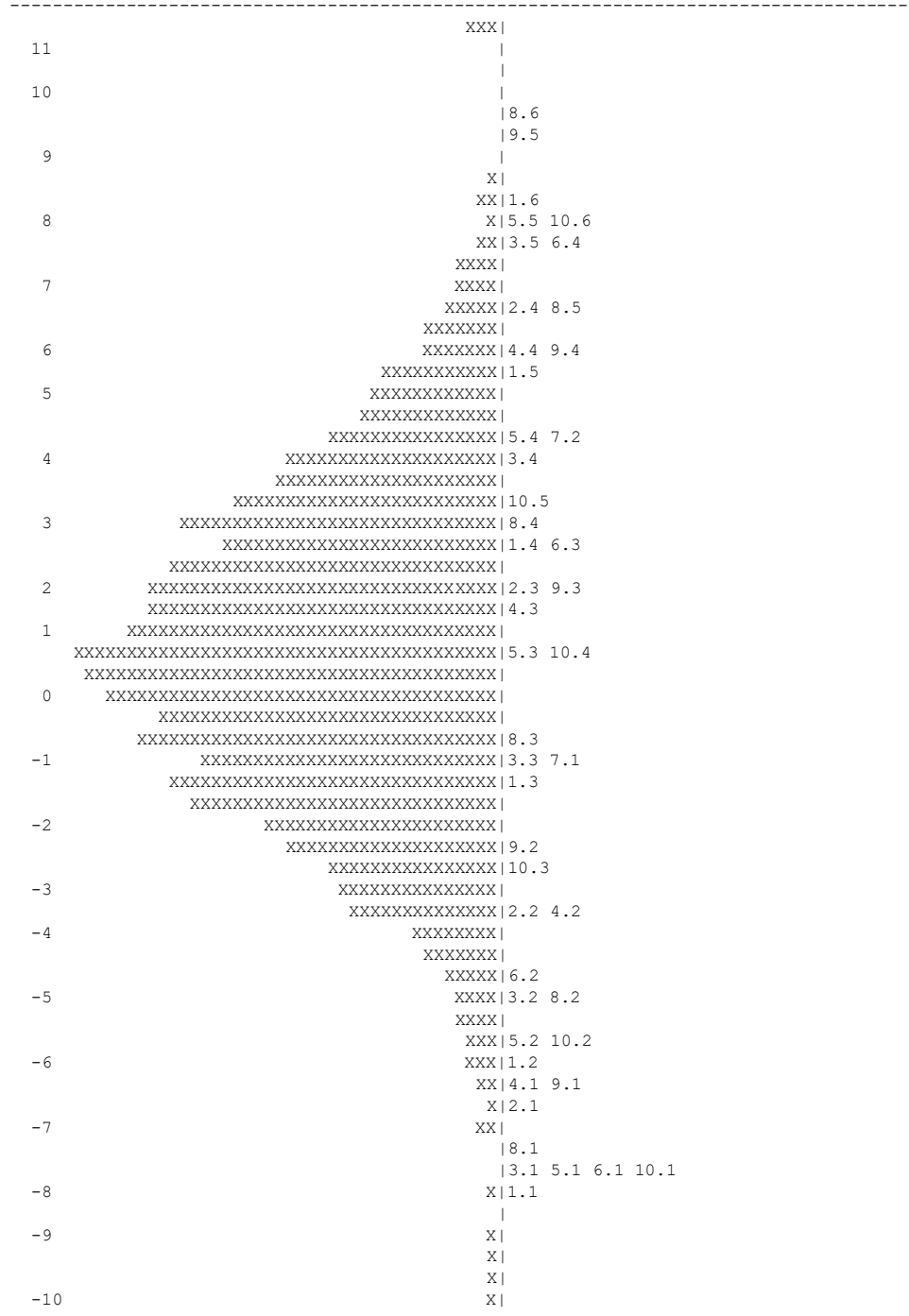
For the polytomously scored writing tests, the criterion difficulty of each of the 10 rating criteria is plotted in Figure 24 with the latent ability distribution on the left-hand side. Figure 25 shows locations of the Thurstonian thresholds of each item and again with the latent ability distribution on the left-hand side. The notation *a.b* indicates threshold *b* of criterion *a*. The location of the threshold indicates the ability level required for a student to have 50 per cent chance of achieving category *b* on criterion *a*. The maps show that the thresholds are well spread out and well separated.

Other item-person maps are included in Appendix E.

Chapter 5: Scaling methodology and outcomes

```

=====
NAPLAN 2021 Writing - Item Calibration Online Test      Tue Jul 06 21:31 2021
MAP OF LATENT DISTRIBUTIONS AND THRESHOLDS
=====Build: Aug 31 2020=====
Generalised-Item Thresholds
  
```



```

=====
Each 'X' represents 1144.8 cases
The labels for thresholds show the levels of
criteria, and category, respectively
=====
  
```

Figure 25. Thurstonian thresholds for online writing test

Item fit

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

As mentioned above, the ICC of each item shows a comparison of the empirical ICC based on observations from 10 ability groupings (broken line joining 10 dots) and the expected model-based ICC (smooth line), and the two curves should display small or no disparities for an item that has a good fit to the model. The ICCs for all items can be found in Appendix D.

Item fit to the Rasch model was closely examined for numeracy, reading, spelling, and grammar and punctuation at each of the four year levels. As all items were trialed and examined previously, few items should show misfit. Because of the large size of the calibration sample, the confidence intervals for the infit mean squares were rather narrow.

Table 58 and Table 59 present summaries of item statistics in the NAPLAN 2021 online tests and paper tests, respectively. 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, and the number of items in paper tests with discrimination less than 0.25 is also presented.

As seen from Table 58, there were 30 out of 2,763 items from 16 non-Writing online tests having infit greater than 1.20. There were 105 items with facility higher than 0.90 and 66 items with facility less than 0.10. Table 59 shows that there were 7 items across 547 items from 16 non-Writing paper tests having infit greater than 1.20. Regarding classical test statistics, there was a total of 72 items across the 16 tests with discrimination less than 0.25. There were 61 items with facility higher than 0.90 and 13 items with facility less than 0.10. Figure 26 shows the ICC of one online numeracy Year 3 item (item x00133872) with an infit statistic close to 1.00. In contrast, Figure 27 shows the ICC of one reading online item (item x00116955) with an infit statistic (1.29) higher than the criterion value (1.20) for evaluating the goodness of fit of each item. The item parameter estimates and statistics are included in Appendix C for each of the 17 online tests calibration (include writing) and 17 paper test calibrations (also include writing).

The evaluation of goodness of fit to the Rasch model for individual writing items was also based on the weighted mean square statistics. For both online and paper writing, the criteria paragraphing and punctuation exhibited misfit to the Rasch partial credit model (that is, infit are between 1.38 and 1.66). None of the other criteria exhibited misfit to the Rasch partial credit model. Inspection of the ICCs did not reveal large differences between the empirical

and the expected curves for each of the ten criteria. The ICCs of the 10 writing criteria for both paper and online writing are included in Appendix D.

Table 58. Summary of item statistics in NAPLAN 2021 online tests

Domain	Year level	Total number of items	Number of items with Infit > 1.2	Number of items with	
				Facility > 0.90	Facility < 0.10
Numeracy	3	155	2	2	1
	5	169	1	9	1
	7	216	3	5	1
	9	208	1	6	1
Reading	3	233	2	1	0
	5	273	1	8	0
	7	320	1	6	1
	9	288	4	9	2
Spelling	3	119**	6	2	15
	5	116	1	10	6
	7	118	1	9	10
	9	119	4	9	7
Grammar and punctuation	3	107	1	6	2
	5	108	0	9	3
	7	107	2	9	9
	9	107	0	5	7
Writing	3,5,7 & 9	10*	2	n/a	n/a

* Item in Writing is criterion.

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

Table 59. Summary of item statistics in NAPLAN 2021 paper tests

Domain	Year level	Total number of items	Number of items with item-rest correlation < 0.25	Number of items with Infit > 1.2	Number of items with	
					Facility > 0.90	Facility < 0.10
Numeracy	3	36	5	2	2	1
	5	42	2	1	5	2
	7	48	5	0	3	0
	9	48	1	1	4	0

Domain	Year level	Total number of items	Number of items with item-rest correlation <0.25	Number of items with Infit > 1.2	Number of items with	
					Facility > 0.90	Facility < 0.10
Reading	3	37	6	1	4	0
	5	38	3	0	7	1
	7	49	8	0	4	0
	9	49	11	0	9	2
Spelling	3	25	0	1	1	0
	5	25	0	0	2	1
	7	25	0	1	1	0
	9	25	0	0	1	0
Grammar and punctuation	3	25	8	0	5	1
	5	25	8	0	5	1
	7	25	9	0	4	2
	9	25	6	0	4	2
Writing	3,5,7 & 9	10*	0	2	n/a	n/a

* Item in Writing is criterion.

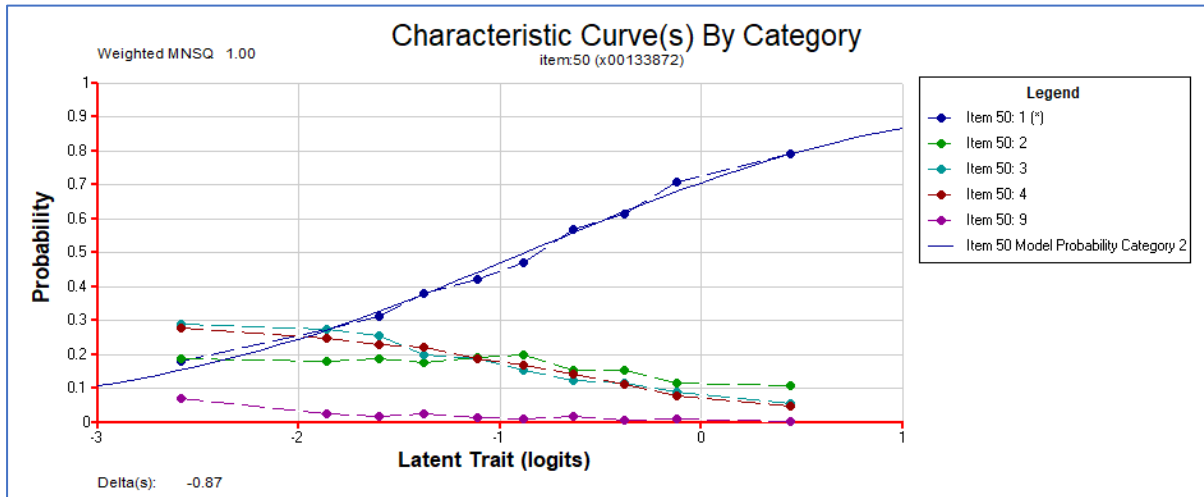


Figure 26. Item characteristic curves for an item with *infit* = 1.00

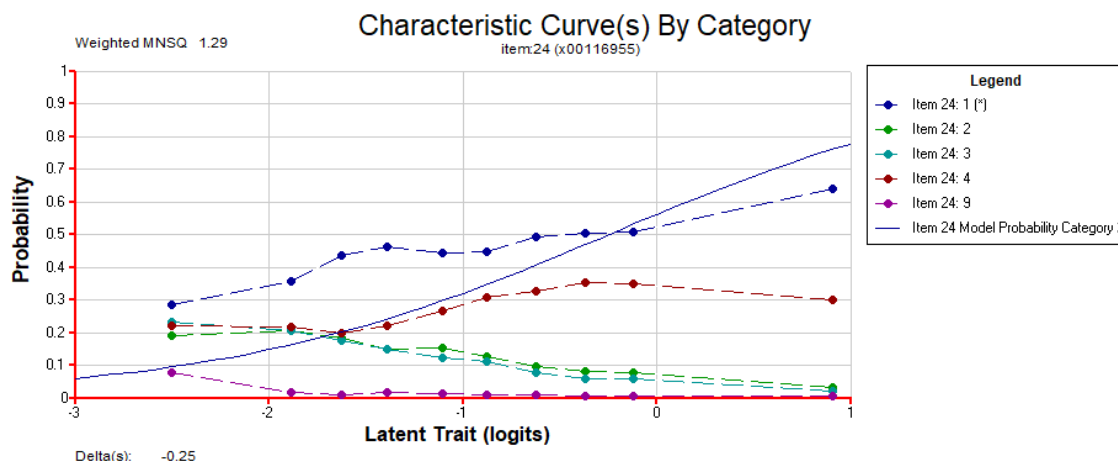


Figure 27. Item characteristic curves for an item with $infit = 1.29$

Differential Item Functioning (DIF) Analyses

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

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

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

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

For visual depiction of DIF, item locations of the reference group are plotted against those of the focal group as seen from appendices F, G and H (that is, gender, LBOTE and Indigenous status, respectively). Each item is represented by one point on the plot. An identity line ($y=x$) is plotted as the reference line. If the relative item difficulty for an item is not different between the two groups after taking their relative performance on the test into account, the point representing the item is on the reference line. The distance of a point

from the diagonal reflects the magnitude of DIF. Due to the large sample sizes, confidence bands were very narrow and were not plotted on the charts.

Gender DIF

Appendix F presents the scatter plots for examining gender DIF in the five domains for both paper and online tests. The plots for numeracy, reading, spelling, and grammar and punctuation are presented by year levels. The writing gender DIF was performed by combining all four grades together. On the whole, the plots indicate that there are few items that exhibit gender differences in the adjusted item estimates and that any differences are not large and thus were not of great concern.

Table 60 identifies the number of items (out of the total number of items) that show gender DIF with an absolute difference of 0.50 or greater for numeracy, reading, spelling, grammar and punctuation and writing⁹. Figure 28 shows as an example, one Year 3 numeracy online test item (Item x00116607) with an absolute difference of 0.50 or greater. This item with a positive difference indicates that the item was relatively easy (difference = 0.79) for male students. Appendix F includes DIF plots that show for each of the items the observed curves by gender group compared with the expected ICC.

Table 60. Number of items showing gender DIF by domain by year level

Test mode	Year level	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
Online	3	13/155	2/233	10/119	1/107	0/10
	5	13/169	9/273	11/116	1/108	
	7	22/216	15/320	20/118	3/107	
	9	11/208	24/288	19/119	5/107	
Paper	3	1/36	0/37	1/25	0/25	0/10
	5	2/42	1/38	3/25	0/25	
	7	1/48	0/49	6/25	0/25	
	9	0/48	2/49	1/25	1/25	

⁹ For writing, item referred is marking criterion. This is applied throughout the report.



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

Figure 28. Example of item characteristic curves displaying gender DIF†

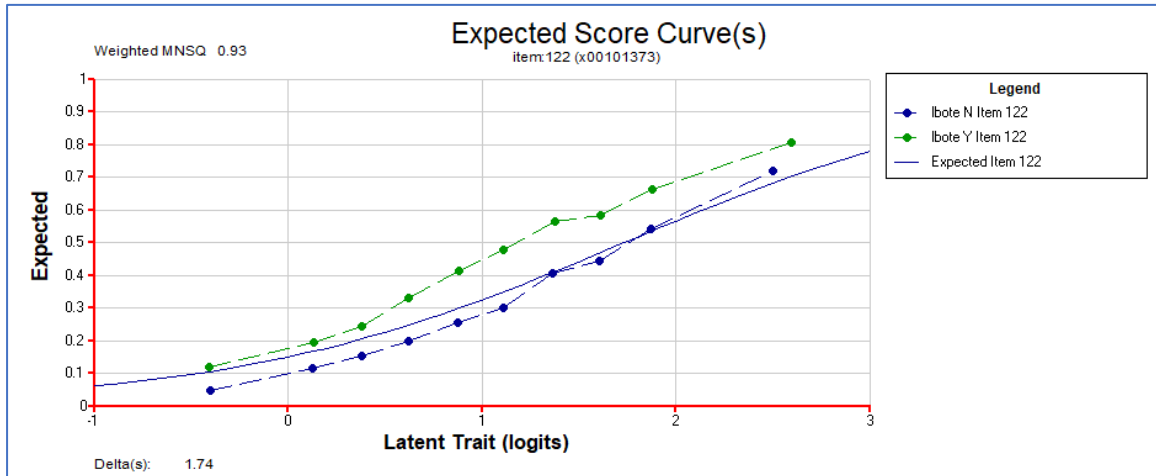
Language background DIF

Appendix G shows scatter plots for examining DIF due to language background in the five domains by the four year levels for both paper and online tests. Writing LBOTE DIF was performed by combining all four grades. These plots indicated that there were not many items that showed notable differences in the relative item difficulties.

Table 61 indicates the number of items that show DIF with an absolute adjusted difference of 0.50 or greater for reading, spelling, grammar and punctuation, and numeracy. Figure 29 depicts one Year 5 numeracy online test item (item x00101373) with an absolute mean difference of 0.50 or greater. This item was relatively easy (mean difference = -0.71) for LBOTE students.

Table 61. Number of Items Showing LBOTE DIF by Domain by Year Level

Test mode	Year level	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
Online	3	3/155	2/233	4/119	9/107	0/10
	5	6/169	1/273	10/116	11/108	
	7	6/216	2/320	10/118	11/107	
	9	13/208	14/288	14/119	14/107	
Paper	3	0/36	0/37	1/25	0/25	0/10
	5	0/42	0/38	0/25	1/25	
	7	1/48	0/49	0/25	2/25	
	9	3/48	0/49	2/25	2/25	



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

Figure 29. Example of item characteristic curves displaying LBOTE DIF†

Indigenous status DIF

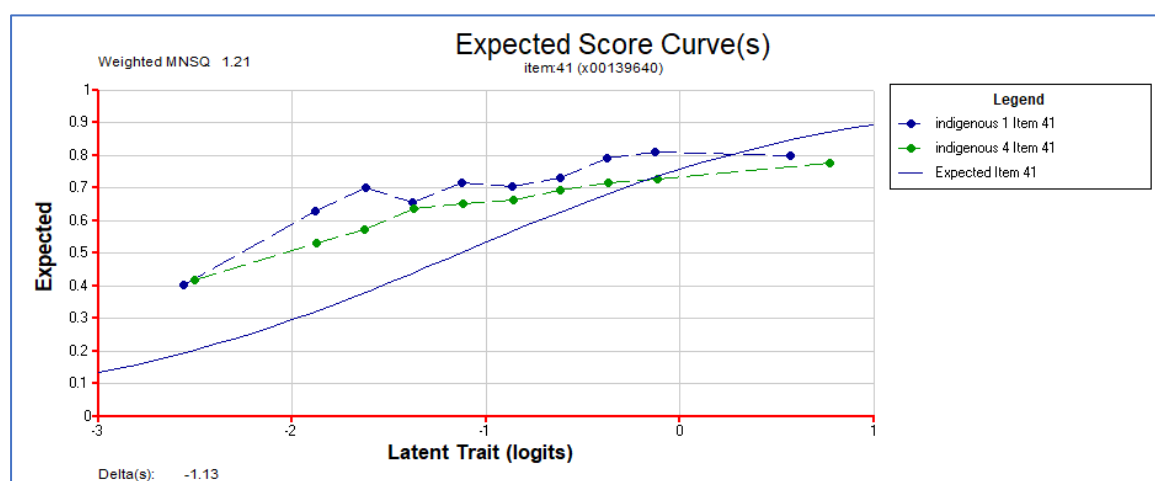
Appendix H includes scatter plots for examining Indigenous DIF in the five domains for both paper and online tests. Writing Indigenous DIF was performed by combining all four grades. These plots showed that there were not many items that showed notable differences in the relative item difficulties for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

Table 62 lists the number of items that show Indigenous DIF with an absolute adjusted difference of 0.60 or greater for reading, spelling, grammar and punctuation, and numeracy. The larger threshold (that is, 0.60 instead of 0.50) was used in order to identify only the items that showed larger DIF. Figure 30 depicts one online reading item (item x00139640) with an absolute mean difference of 0.60 or greater. This item was relatively easy (mean difference = -0.83) for Indigenous students.

Appendix H 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 62. Number of items showing Indigenous DIF by domain by year level

Test mode	Year level	Numeracy	Reading	Spelling	Grammar and punctuation	Writing
Online	3	2/155	2/233	0/119	7/107	0/10
	5	2/169	7/273	2/116	6/108	
	7	2/216	9/320	0/118	6/107	
	9	5/208	6/288	0/119	3/107	
Paper	3	1/36	5/37	0/25	7/25	0/10
	5	5/42	4/38	0/25	6/25	
	7	1/48	6/49	0/25	1/25	
	9	7/48	1/49	0/25	2/25	



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

Figure 30. Example of item characteristic curves displaying Indigenous DIF†

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

Jurisdictional DIF

In order to determine whether state/territory DIF exists, all tests were calibrated independently by state/territory by year level by mode. The relative item difficulties (or criterion difficulties for writing) were compared to the average item difficulty of eight states/territories for the online tests or four states/territories for the paper tests. The following procedures were applied:

- Items were calibrated by test mode, by jurisdiction, by domain and year level; item parameters were then delta-centred.
- The national item parameter for each item was calculated by averaging the states/territories item parameters.
- The parameter difference for item(i) between a state/territory and national average was calculated as:

$$Difference(i) = \frac{Item\ Parameter(i) - National\ Average\ of\ Item(i)}{2 \times Standard\ Error\ of\ Item(i)} \quad (4)$$

The differences were compared with Bonferroni Corrected Index (BCI) for all possible comparisons (3.11 for 28 pairs of comparisons from 8 jurisdictions administered the online tests, and 2.63 for 6 pairs of comparison from four jurisdictions administered the paper tests). If the difference for an item between a state/territory and national average was greater than the BCI and the item parameter difference is greater than 0.5 logit, then the item was deemed harder for the state/territory. If the difference was less than the BCI and the item parameter difference is greater than -0.5 logit, then the item was deemed easier for the state/territory.

The number of items showing statistically significant state/territory related DIF in online and paper numeracy, reading, spelling, grammar and punctuation, and writing are 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. For online tests, there were only two items in numeracy, seven items in spelling, nine items in grammar and punctuation and one writing criterion showing potential DIF, across all four year levels across the eight jurisdictions. For paper tests, based on the criteria described above, there were three items, one in each of reading, spelling, and grammar and punctuation, showing potential DIF across among the four year levels and four jurisdictions delivering the test on paper. Table 63 can be read in conjunction with Appendix J, which contains item DIF plots for items showing state/territory related DIF. For example, from Table 63, there was one item in Year 5 numeracy showing DIF in Qld when compared with the national level, with this item (x00137377) being easier for Qld, as seen in Figure 31 and from Appendix J.

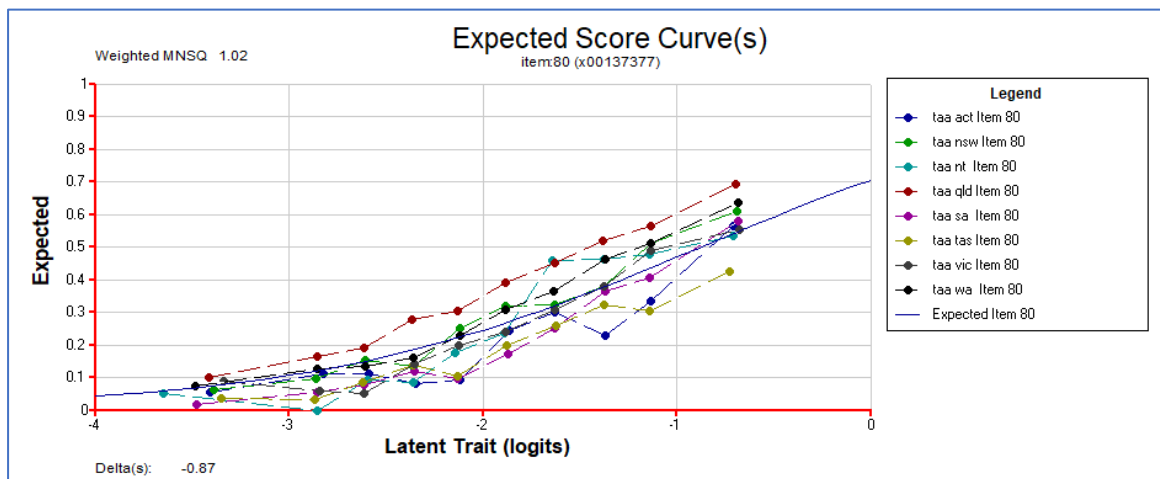


Figure 31. Example of item characteristic curves displaying jurisdictional DIF

Table 63. Number of items showing state/territory DIF by domain by year level

a) Online tests

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	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	9	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
Reading	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Spelling	3	-	-	-	-	-	-	-	-	-	1	-	-	-	1	1	-
	5	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-
	7	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1	-
	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
Grammar and punctuation	3	-	-	-	-	-	1	2	-	-	-	1	-	-	-	-	-
	5	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-
	7	-	-	-	-	-	-	3	-	-	-	-	-	-	-	-	-
	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Writing	3, 5, 7 & 9	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-

b) Paper tests

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	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Reading	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Spelling	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Grammar and punctuation	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	7	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Writing	3, 5, 7 & 9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Note. 'E' indicates that the item is relatively easier for the jurisdiction, and 'H' indicates that the item is relatively harder for the jurisdiction.

Device DIF

For online tests, a device DIF analysis was also carried out for non-writing domains¹⁰ as there were different devices used by different students. There were four different types of devices: Chromebook, iOS, Mac and Windows device. 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 pooled online item parameters. An item parameter demonstrating a significant difference greater than 0.25 logits was deemed as exhibiting DIF. A summary of device DIF is shown in Table 65. Table 65 shows that Mac and

¹⁰ Device DIF was not investigated for writing as some students completed the test on paper while others completed the test online

Windows devices had the most items demonstrating DIF, with Chromebook having only two items and iOS devices having only one item with potential DIF.

Table 64. Number of students by device

Domain	Year level	Chromebook	iOS	Mac	Windows
Numeracy	3	30670	52433	3344	81911
	5	30749	40781	5549	84440
	7	13525	13324	24261	92947
	9	11093	10808	26707	82768
Reading	3	30055	52455	3157	83110
	5	30275	40135	5910	86136
	7	11525	12139	22009	86367
	9	9384	8887	24154	77116
Spelling	3	28044	48189	2973	75131
	5	28814	37540	5296	78222
	7	11955	11789	23201	86933
	9	9673	10005	24905	76558
Grammar and punctuation	3	28350	48555	3003	76547
	5	28851	37585	5299	78478
	7	11979	11808	23222	87202
	9	9701	10022	24921	76813

Table 65. 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	-	-	-	-	-	-	-	-
	7	-	-	-	-	-	2	-	-
	9	1	-	-	-	2	3	-	-
Reading	3	-	-	-	-	-	-	-	1
	5	-	-	-	-	-	-	-	1
	7	-	-	-	-	2	-	-	2
	9	-	-	-	-	-	1	-	-
Spelling	3	-	-	-	-	-	1	-	-
	5	1	-	-	-	-	1	-	-
	7	-	-	-	-	-	-	-	-
	9	-	-	-	-	-	2	-	3
Grammar and punctuation	3	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	1	2
	7	-	-	-	-	-	3	-	-
	9	-	-	-	1	-	1	-	-

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 five plausible values was drawn for each domain.

Scoring and the generation of score-equivalence tables based on WLEs in logits were generated for each test path of the online tests by domain by year level or for each of the

paper tests based on delta-centred item parameters. Transformations were applied to the logit scores for conversion to NAPLAN reporting scale scores on the historic NAPLAN scales as was done in previous years.

For the estimation of population statistics, rather than using the WLE estimates, five sets of PVs of student latent proficiency estimates were drawn using *ACER ConQuest 5* based on imputation techniques and a multidimensional item response model (partial credit model) with latent regression (Wu et al., 2007) 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 regression model used in 2021 was the same as that used in previous NAPLAN cycles. 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, and the school reading WLE average score (adjusted for the student's own score) as a measure of average proficiency at the school level. A diagrammatic representation of the multidimensional model is shown in Figure 32.

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

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

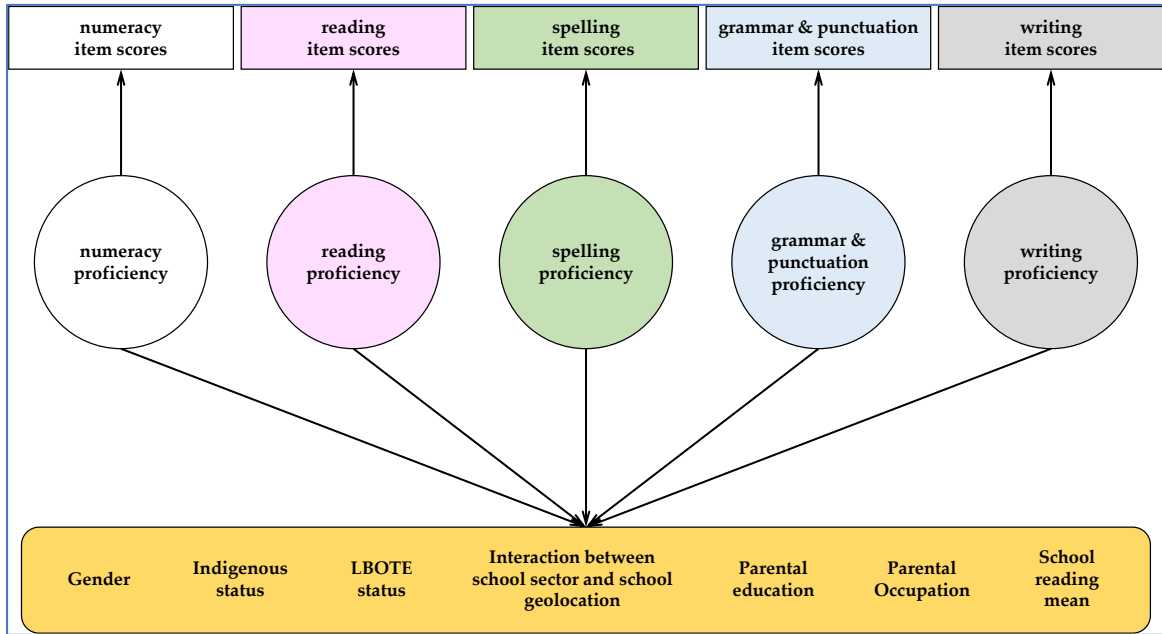


Figure 32. Conditioning variables for the multidimensional item response model with latent regression model

Chapter 6: Equating procedures

In 2021, about 70 per cent of students sat the online tests and another 30 per cent of students sat the paper tests. This chapter describes the process of equating the 2021 tests onto the NAPLAN historic scales for both the paper tests and the online tests in turn.

This chapter first describes equating procedures for numeracy, reading, spelling, and grammar & punctuation, and finishes with a description of the equating procedures for writing. For writing, a different equating design and methodology was applied.

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

NAPLAN results are reported using five national achievement scales, one for each of the assessed domains of literacy – reading, writing, spelling, and grammar and punctuation – and one for numeracy. The vertical and horizontal equating design for both online and paper tests is represented schematically in the data matrix in Table 66.

The 2021 year level NAPLAN tests were linked to each other by a set of common items between adjacent year levels. Due to the pandemic, there was no NAPLAN test administered in 2020. The 2021 online tests were linked to the historical scale by a set of items used in the 2019 tests and the majority items included in the paper tests were also in the online tests. The 2021 online and paper tests were also linked to the historical scale by a secure equating test that had been administered since 2009 to an equating sample selected from each cohort. The equating test is a paper test administered to equating samples from both assessment modes. Therefore, vertical equating is based on a common item equating design, while horizontal equating can be based on either common items or a common student equating design.

Table 66. Equating design for both assessment modes

NAPLAN test items (paper or online) – vertical links							
Students	Y3	Y3&5	Y5	Y5&7	Y7	Y7&9	Y9
Y3 population	█						
Y5 population		█					
Y7 population				█			
Y9 population						█	
Equating test items (paper and online) – horizontal links							
Students	Y3		Y5		Y7		Y9
Y3 equating sample	█						
Y5 equating sample			█				
Y7 equating sample					█		
Y9 equating sample							█
NAPLAN 2021 test items (online) – horizontal links							
Items	Y3		Y5		Y7		Y9
Y3 2019 test	█						
Y5 2019 test			█				
Y7 2019 test					█		
Y9 2019 test							█

The NAPLAN scale was established in 2008 by placing all year levels on the same scale using vertical link items. For the purpose of monitoring student achievement over time, the NAPLAN 2021 scale for each domain needs to be horizontally equated to the historic NAPLAN reporting scale. Although online tests can be equated to the NAPLAN historical scale using the common-person equating design, the horizontal links between the NAPLAN 2021 online tests and NAPLAN 2019 online tests included a large number of common items administered to the whole population. This provided direct and more stable links. Therefore, and following recommendations from the Measurement Advisory Group, common item equating was used as the final horizontal equating method to bring the NAPLAN 2021 scale onto the NAPLAN historical scale for online tests.

While the online and the paper test had many items in common, explorations of different equating methods revealed that direct equating between an adaptive and a non-adaptive test was not liable (also confirmed by the Measurement Advisory Group). Therefore, the paper test was equated to the historical using the same method as in the past; that is, using a common-person approach and a secure equating test.

A sample of students from Years 3, 5, 7 and 9 were administered the secure paper equating tests at their year level two weeks prior to the NAPLAN 2021 tests. A minimum of 600 students was required for each domain, from each year level. The number of students per jurisdiction was proportional to the size of paper sub-population in each jurisdiction.

The response data on the equating test were used to equate the 2021 paper tests onto the existing NAPLAN reporting scales. First, the response data of the equating test was

merged with the response data on the NAPLAN test. Second, the items in the equating test were freely estimated while anchoring the NAPLAN items to their official estimates. The difference in average difficulty of the equating test items in 2021 and in 2009 was the equating shift from 2021 to the historical scale.

In theory, no vertical link items were needed after 2008, when all year levels were placed on the same historical scale, because each year level could be shifted onto the historical scale by common student equating using the equating tests. However, vertical link items were used in all subsequent years to check and adjust the horizontal shifts for each year level. This method was labelled the horizontal–vertical regression (HVR) equating method and will be described in detail below.

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

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

Stage 1. Initial cross-test form scatterplots with all items were examined to ascertain the overall correlation and to note any patterns and outliers.

Stage 2. Each item was checked for misfit at each test form based on how well items discriminate between high- and low-performing students. Discrimination was checked by inspection of the ICC and graphical fit, infit statistics and the item-rest correlations. Items that showed pronounced misfit in either test form were omitted from the linking set.

Decisions to omit items due to misfit were not based on any one indicator in isolation; rather, decisions were based on all available evidence concerning the functioning of each item. Items that fail some criteria are normally excluded from the linking set but may have been retained if the total number of functioning links was relatively small, especially for vertical links in the paper test.

Stage 3. Items were omitted if they showed cross test form DIF. To evaluate test form DIF, difficulties of the set of common items were centred around zero for each test form. For each pair of adjacent tests, one set of item difficulties (e.g., of 2021 Year 3 link items) was then plotted against the other set of item difficulties (of 2019 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 review and selecting good link items. On the plots, each dot represents a common item. Links were broken in two steps. Outliers (absolute difference larger than 0.9 of a logit) were broken first. Any other items with an absolute difference of more than 0.5 were broken in the second step. For each set of adjacent test scales, mean item difficulties of the link items were calculated for each of the two test forms. The equating shift (either horizontal or vertical) is the difference between the two means.

After each stage, the scatterplot 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 adjacent test form (that is, 2021 Year 3 SD / 2019 Year 3 SD). Ideally the ratio should fall between 0.9 and 1.1.

This link-item review procedure was the same for NAPLAN paper tests and online tests, and the same for horizontal and vertical links.

Horizontal equating shifts of the online tests

There were two steps involved in equating the NAPLAN 2021 online tests to the NAPLAN historical tests. First, the 2021 NAPLAN online tests were equated to the NAPLAN 2019 online tests. This placed the NAPLAN 2021 online tests onto the NAPLAN 2019 online delta centered scale. Second, the equating parameters that were previously applied in 2019 to place the 2019 online test scales onto the NAPLAN historical test scales were applied to the NAPLAN 2021 online tests. This step resulted in the NAPLAN 2021 online tests being placed onto the historical NAPLAN scale. The bottom section of Table 66 shows the horizontal equating design for each of numeracy, reading, spelling, and grammar and punctuation at each year level.

Figure 33 to Figure 48 show the comparisons of the 2019 item parameter estimates with the 2021 item parameter estimates, for each of the 16 online tests. For link items that did not change in relative item difficulty, the bivariate points were on the identity line (a green dotted line on each graph). A thin solid line on each figure shows the linear line of best fit through the dots in each scatterplot.

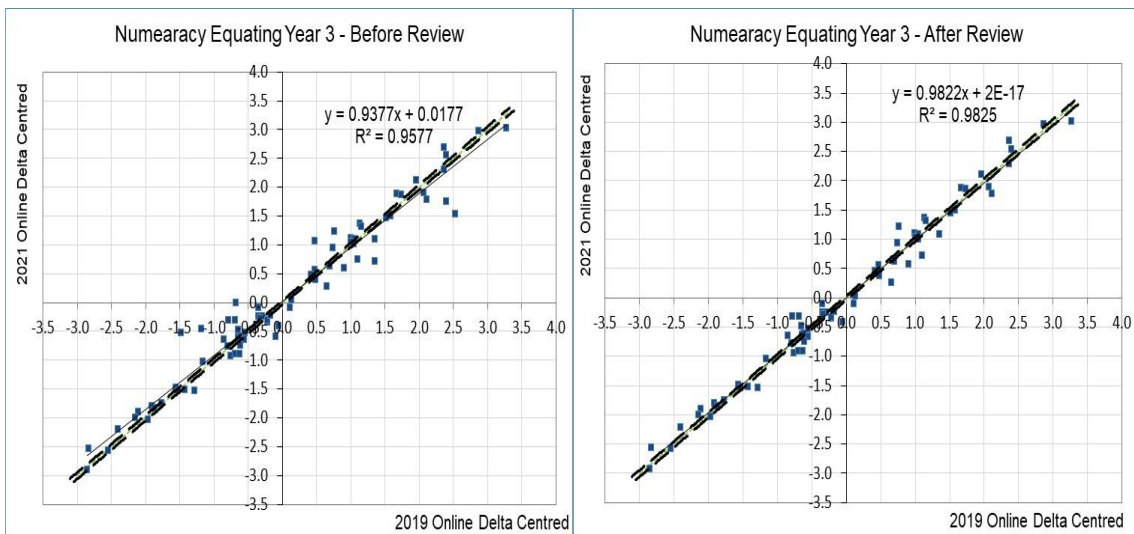


Figure 33. Scatterplot of numeracy, horizontal equating items between 2021 and 2019 for Year 3 online students

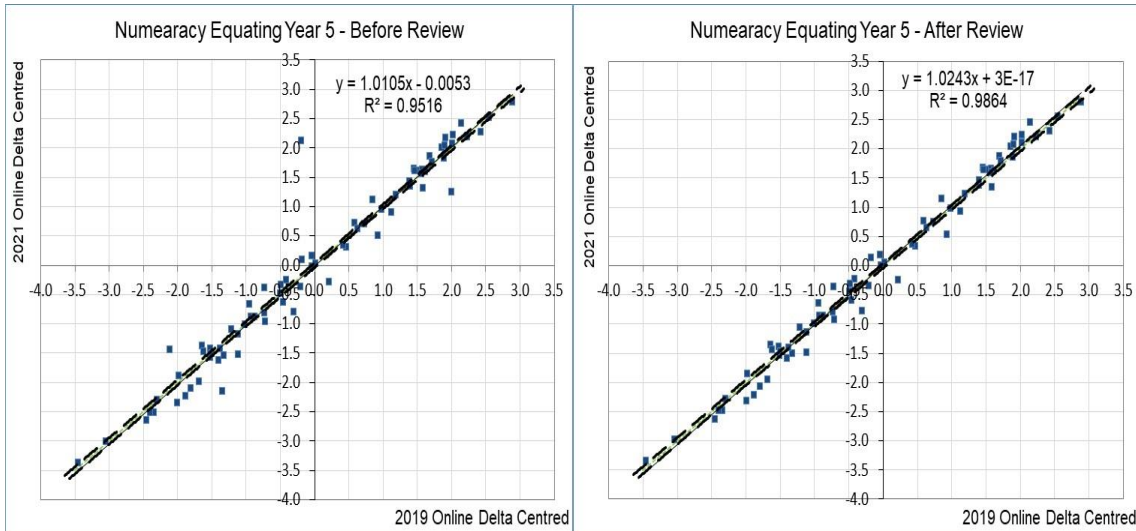


Figure 34. Scatterplot of numeracy, horizontal equating items between 2021 and 2019 for Year 5 online students

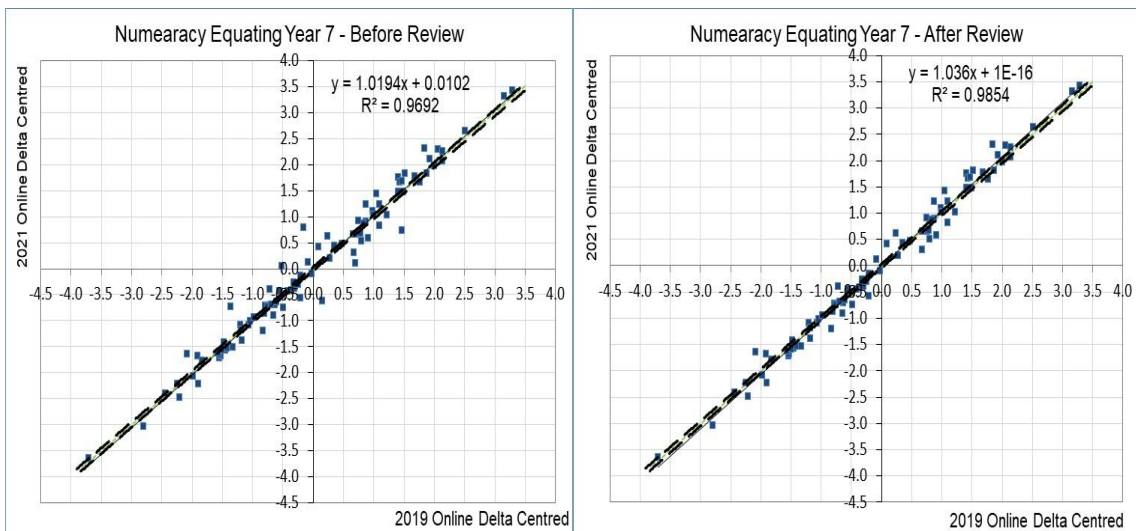


Figure 35. Scatterplot of numeracy, horizontal equating items between 2021 and 2019 for Year 7 online students

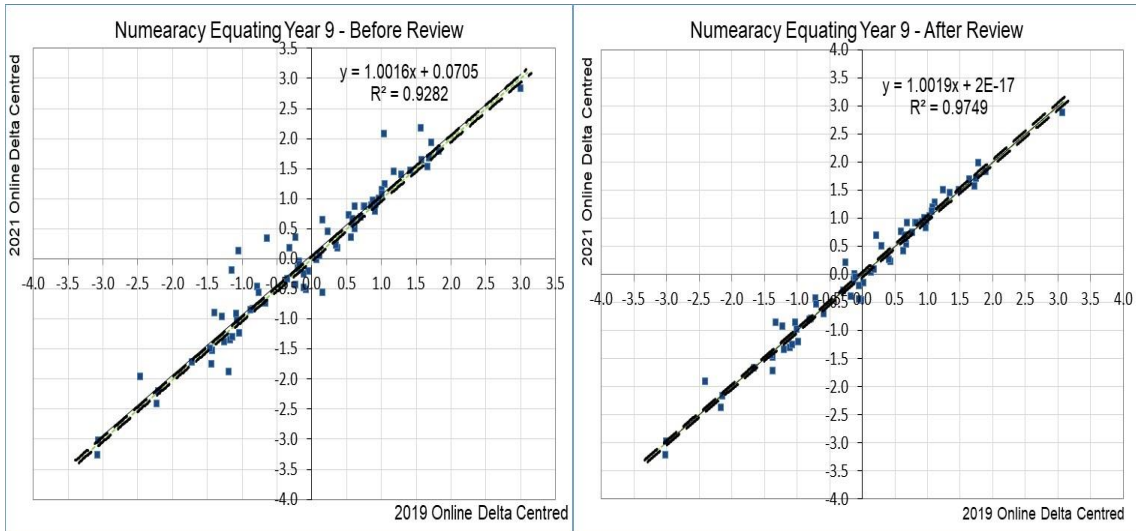


Figure 36. Scatterplot of numeracy, horizontal equating items between 2021 and 2019 for Year 9 online students

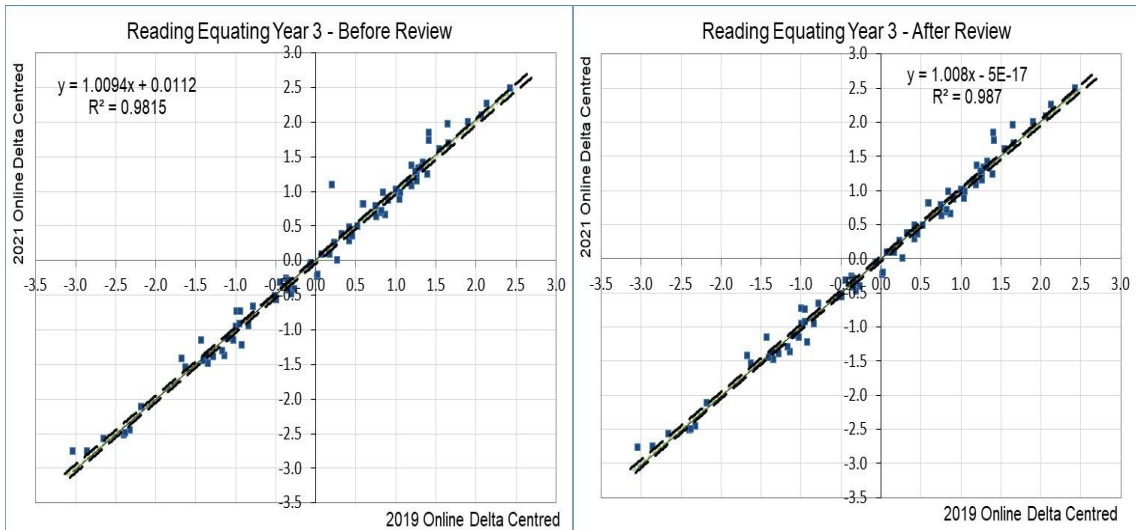


Figure 37. Scatterplot of reading, horizontal equating items between 2021 and 2019 for Year 3 online students

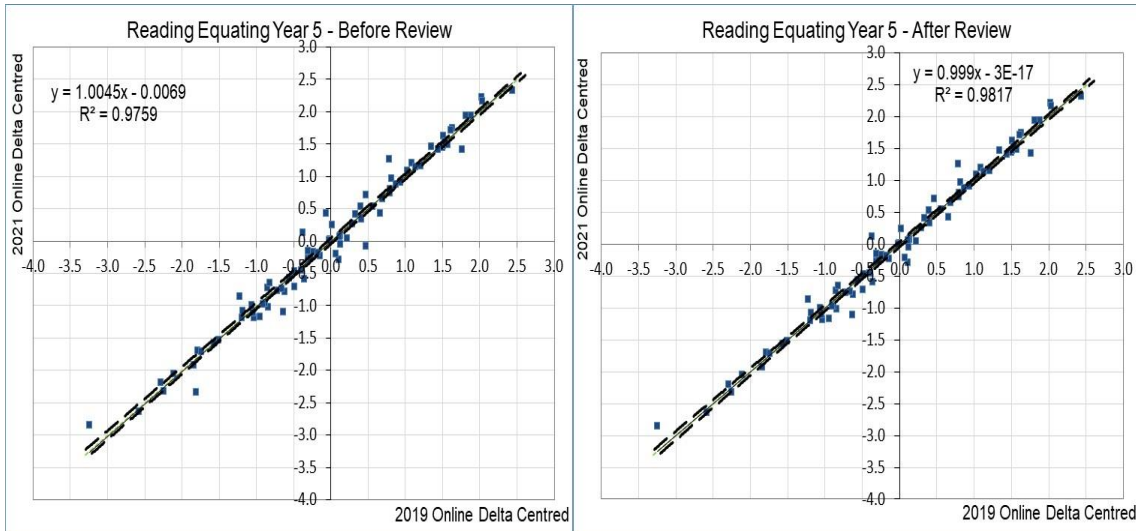


Figure 38. Scatterplot of reading, horizontal equating items between 2021 and 2019 for Year 5 online students

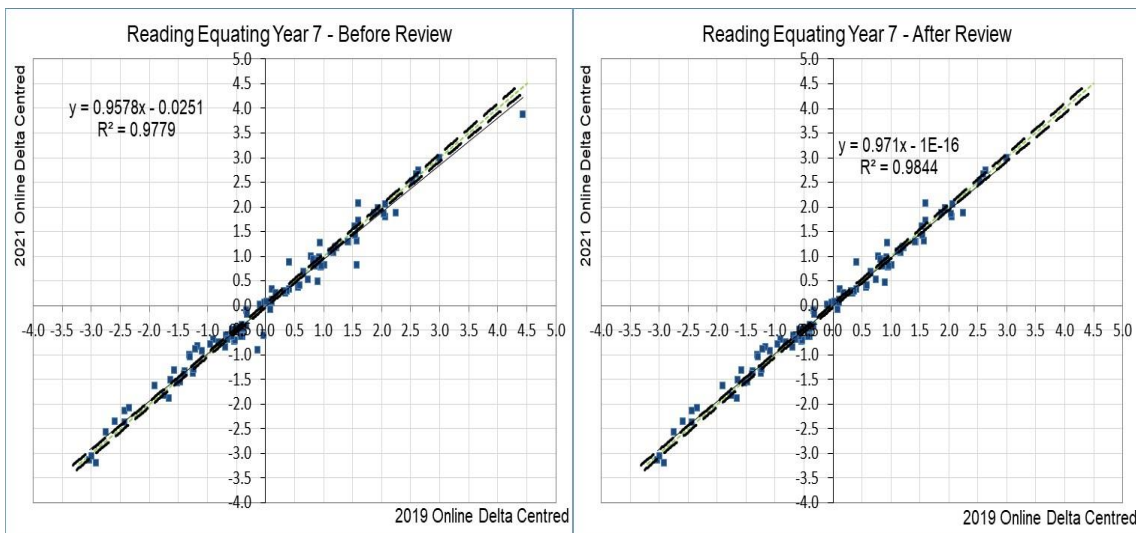


Figure 39. Scatterplot of reading, horizontal equating items between 2021 and 2019 for Year 7 online students

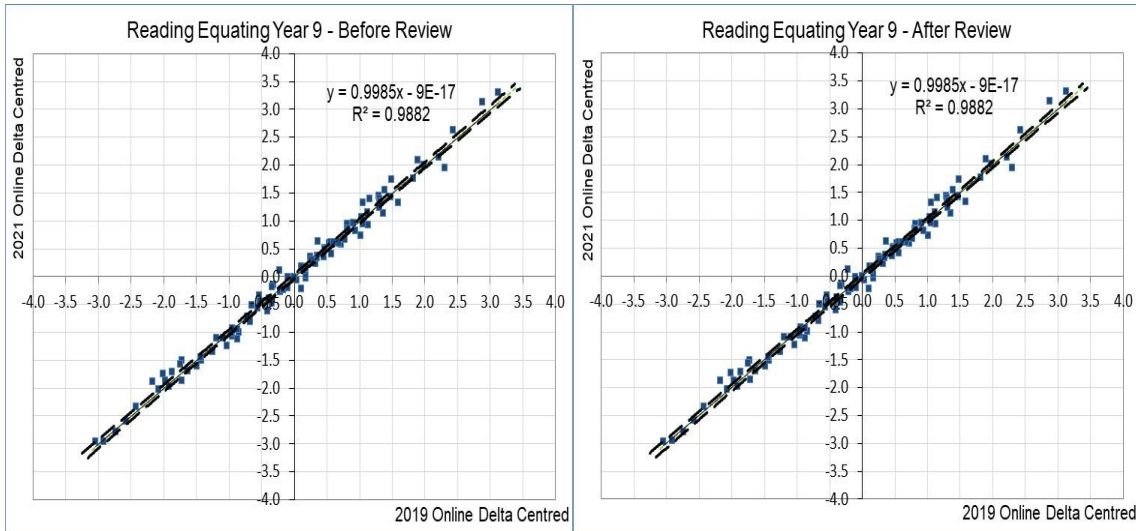


Figure 40. Scatterplot of reading, horizontal equating items between 2021 and 2019 for Year 9 online students

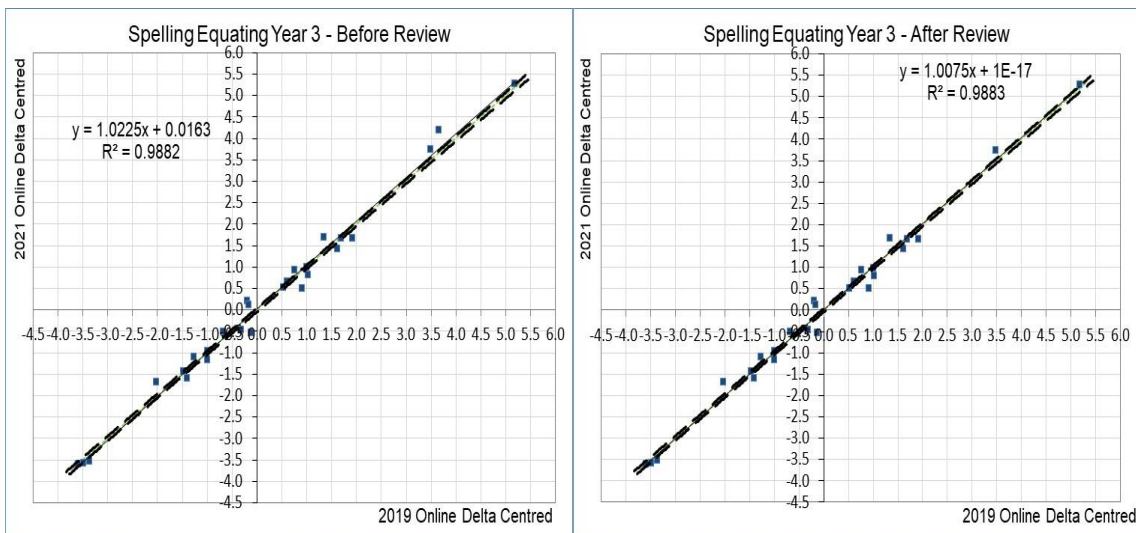


Figure 41. Scatterplot of spelling, horizontal equating items between 2021 and 2019 for Year 3 online students

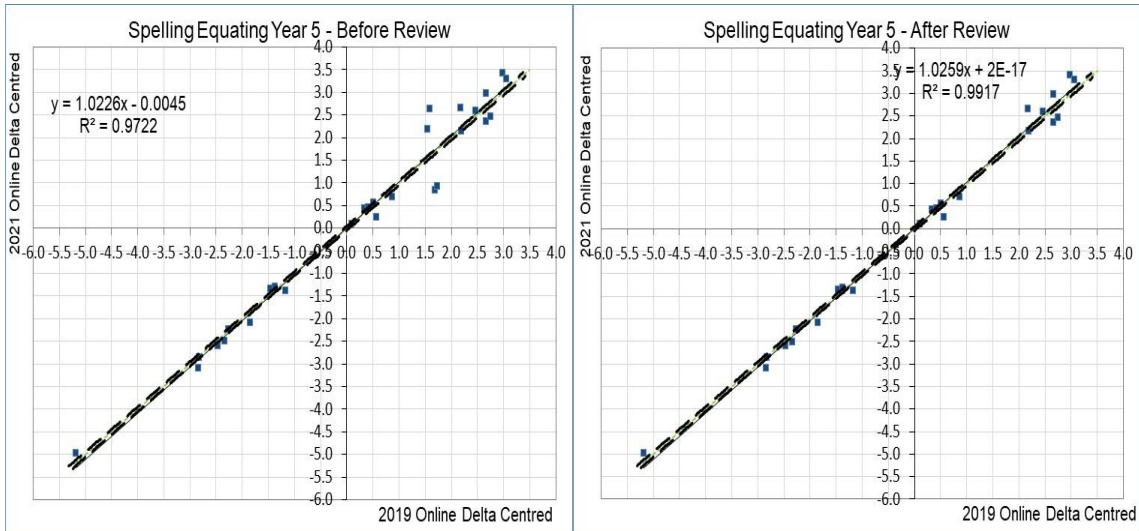


Figure 42. Scatterplot of spelling, horizontal equating items between 2021 and 2019 for Year 5 online students

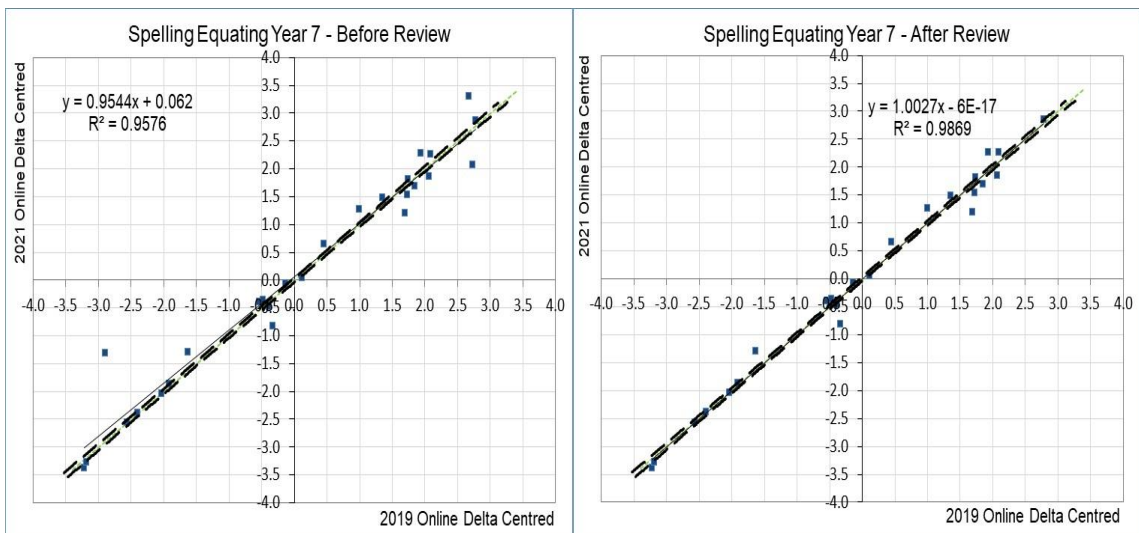


Figure 43. Scatterplot of spelling, horizontal equating items between 2021 and 2019 for Year 7 online students

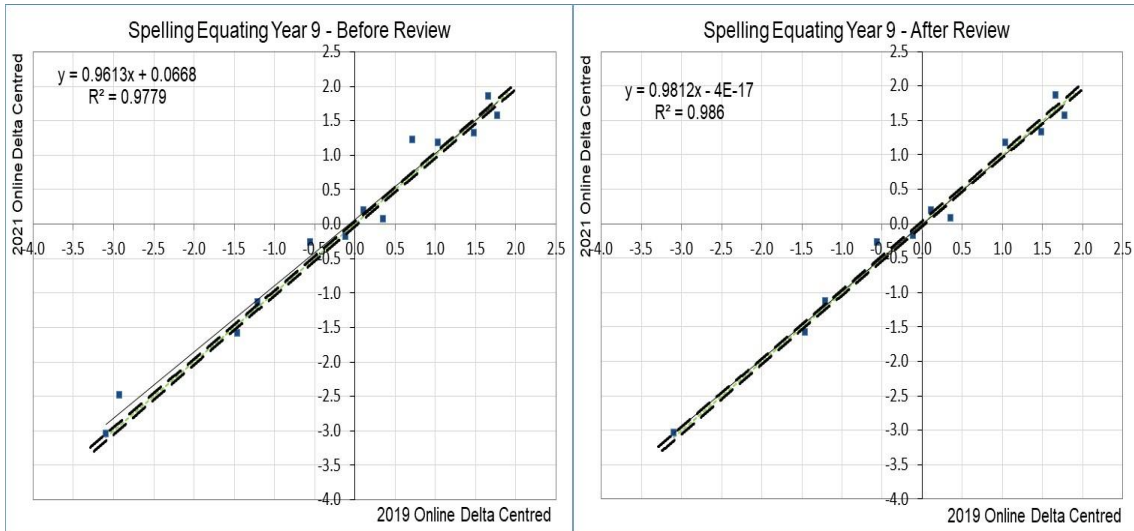


Figure 44. Scatterplot of spelling, horizontal equating items between 2021 and 2019 for Year 9 online students

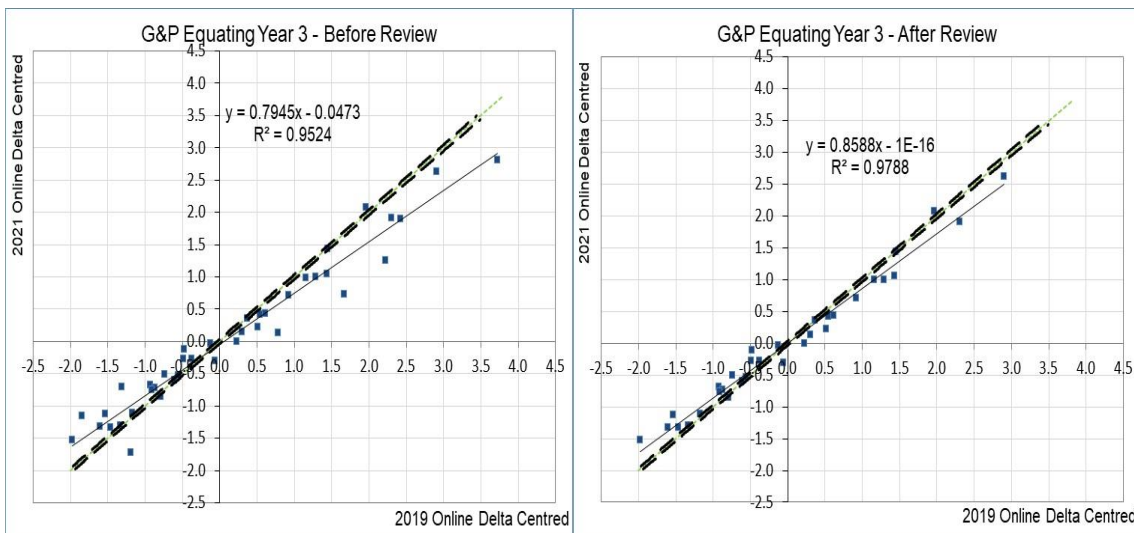


Figure 45 Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2019 for Year 3 online students

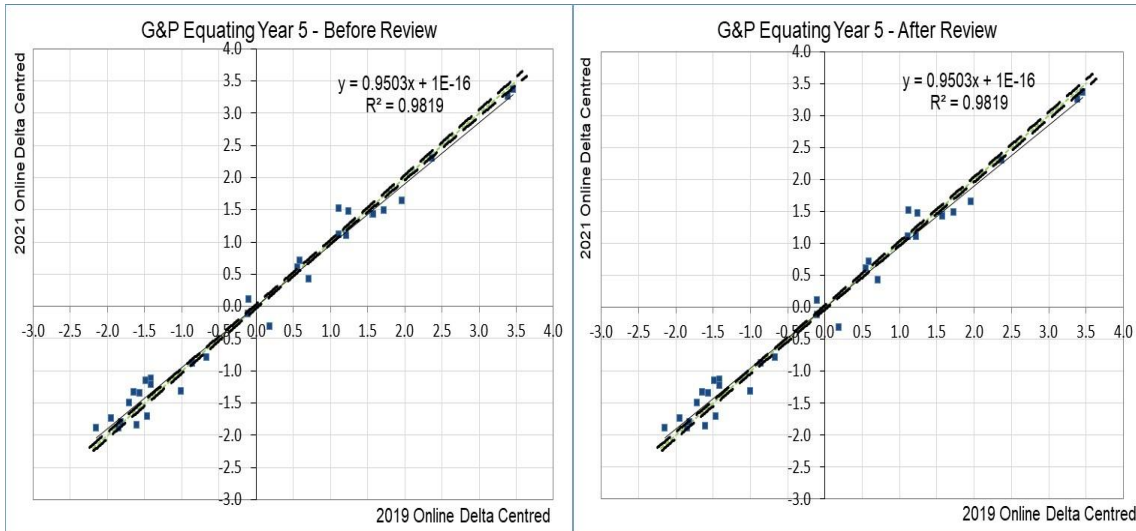


Figure 46 Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2019 for Year 5 online students

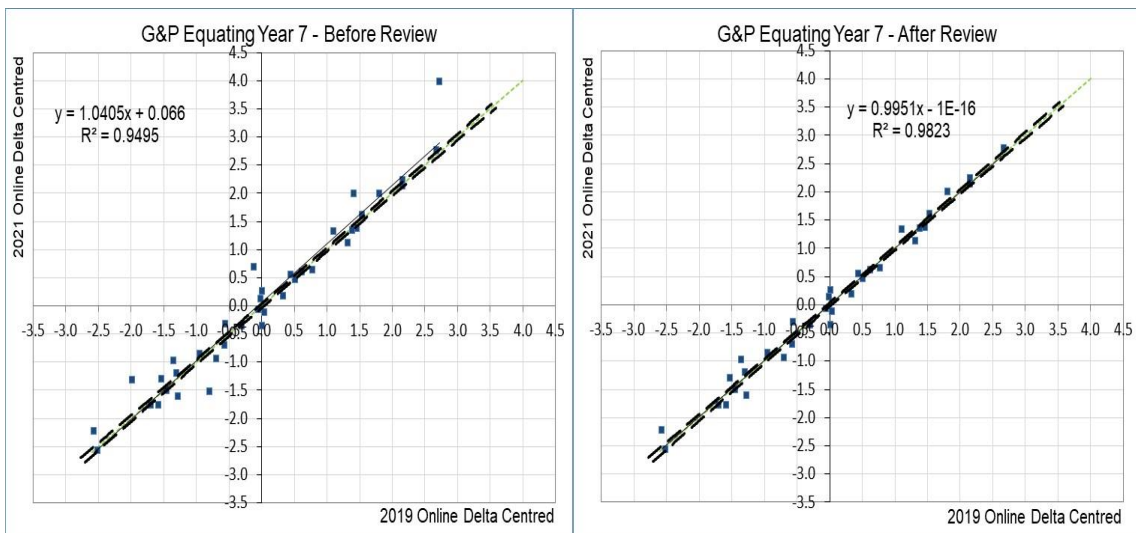


Figure 47. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2019 for Year 7 online students

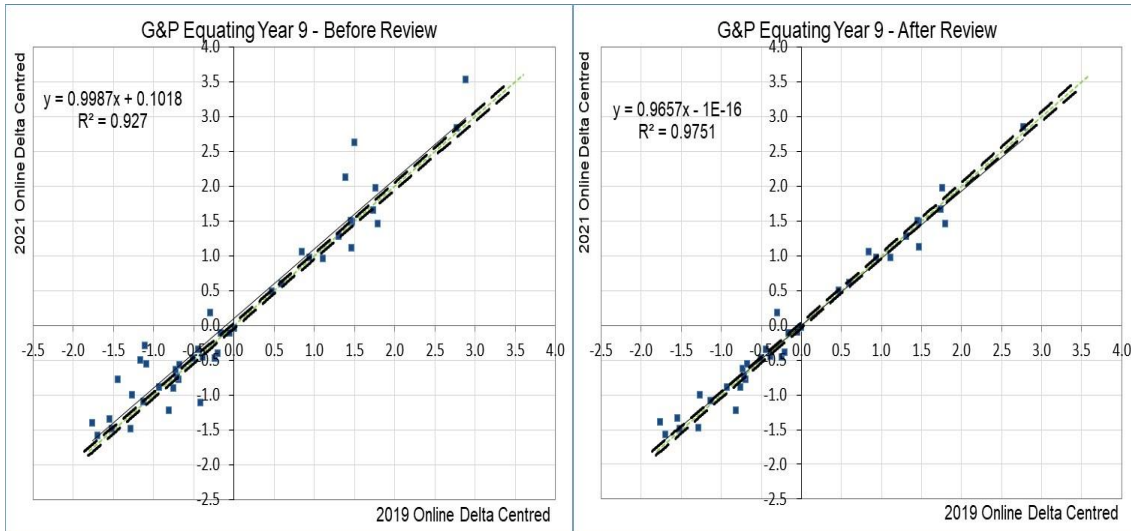


Figure 48. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2019 for Year 9 online students

After the review and evaluation of the equating items between the 2021 and 2019 online tests, a final set of link items was identified for each domain and year level. The final sets of link items were used to calculate the preliminary horizontal shifts from 2021 to 2019. After review, HVR adjustments were not needed for the online scales, so these preliminary horizontal shifts were also the final shifts. These final horizontal shifts equated the 2021 online tests onto the 2019 online tests' delta centered scales. Then, the parameters that used to equate the 2019 online tests to the NAPLAN historical scale will be applied, this would place the 2021 online tests onto the NAPLAN historical scales. The numbers of horizontal links used and retained for each online test are shown in Table 67 and the horizontal shift-constants for each domain at each year level are summarised in Table 68.

Appendix K presents the 2021 horizontal link item locations (Rasch difficulty parameters), standard errors, and differences in the item locations by domain and year level.

Table 67. Horizontal link review summary for online tests

Year level	Numeracy	Reading	Spelling	Grammar and punctuation
3	66/74	77/78	26/27	32/40
5	72/76	80/83	24/28	32/32
7	86/92	93/97	24/27	33/38
9	64/72	98/98	11/13	36/44

Table 68. Horizontal equating shifts between 2021 item locations and 2019 item locations by year level for online tests

Year level	Numeracy	Reading	Spelling	Grammar and punctuation
3	0.089	-0.128	0.503	1.105
5	0.029	-0.028	0.603	0.860
7	-0.305	-0.093	0.254	0.669
9	-0.324	-0.292	0.236	0.614

Horizontal equating shifts for paper tests

As described above, the common-person equating method was used for the paper test. This involved administering a secure paper equating test that has been administered since 2009. Some items were altered and were not used as link items.

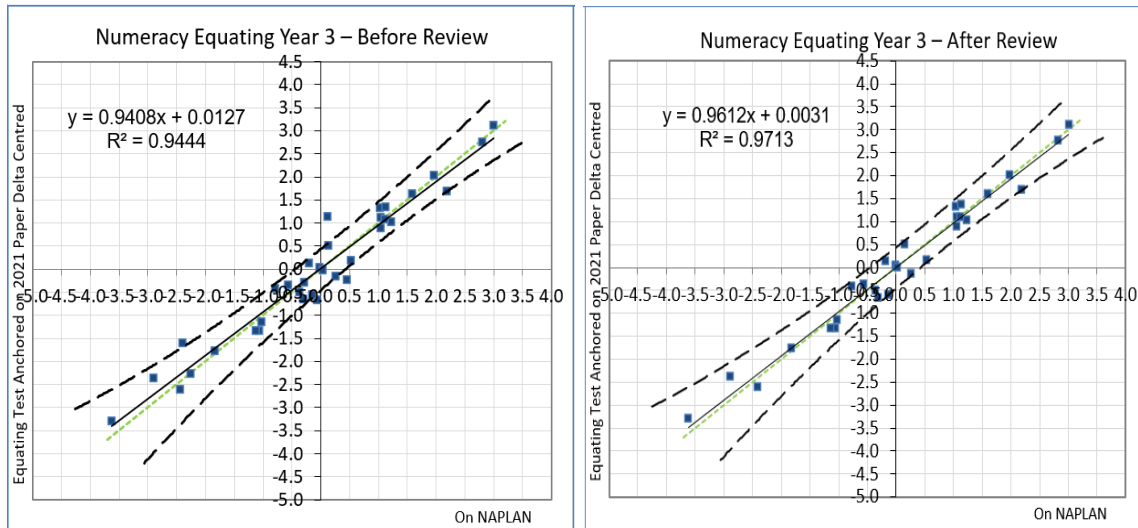


Figure 49. Scatterplot of numeracy, horizontal equating items between 2021 and 2009 for Year 3 paper students

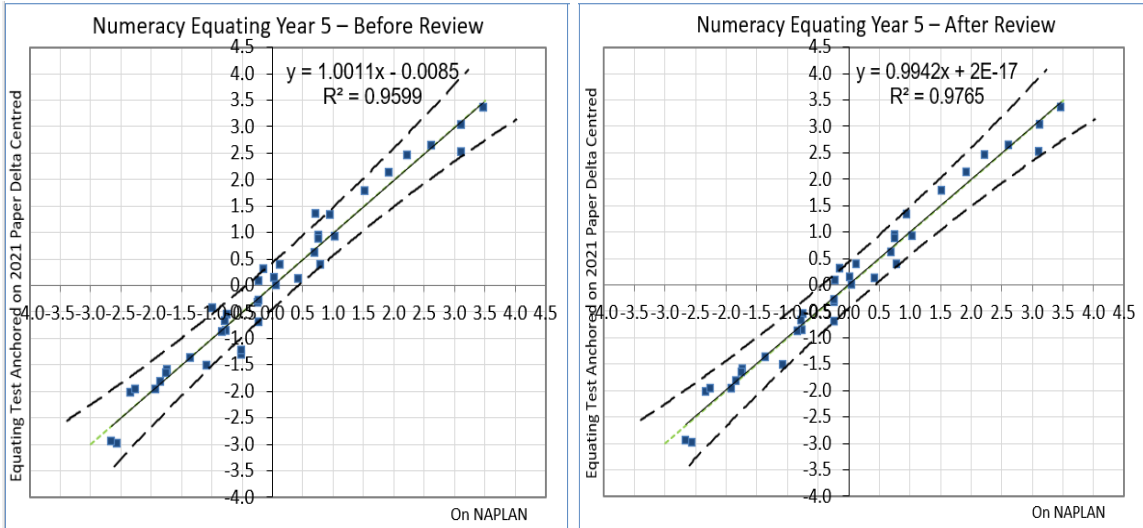


Figure 50. Scatterplot of numeracy, horizontal equating items between 2021 and 2009 for Year 5 paper students

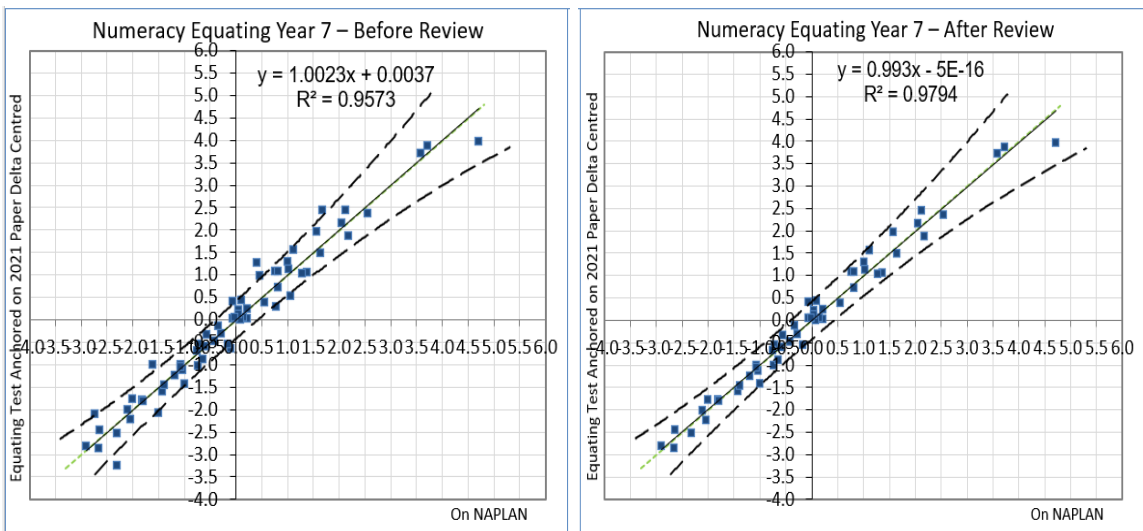


Figure 51. Scatterplot of numeracy, horizontal equating items between 2021 and 2009 for Year 7 paper students

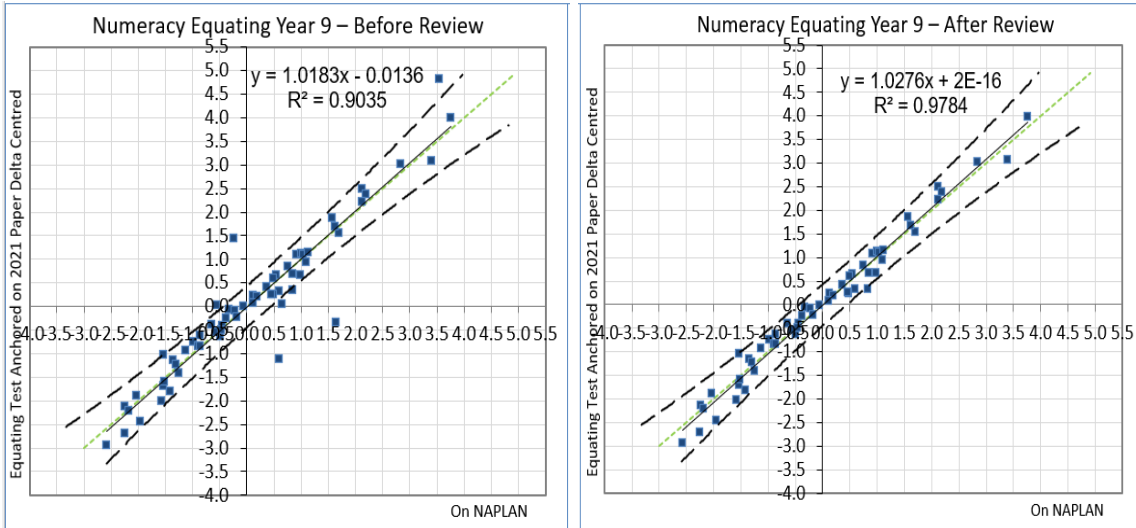


Figure 52. Scatterplot of numeracy, horizontal equating items between 2021 and 2009 for Year 9 paper students

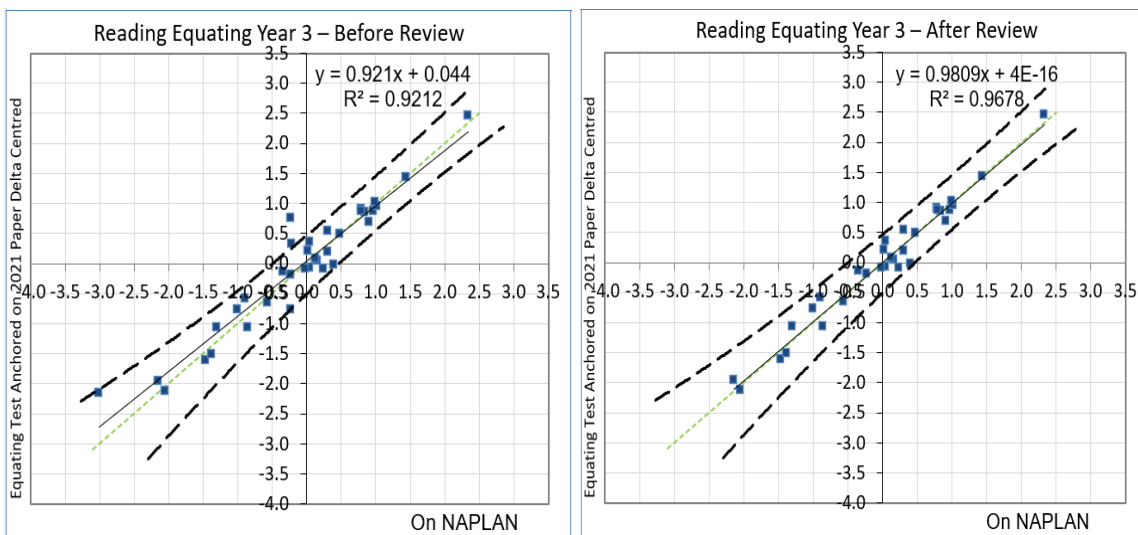


Figure 53. Scatterplot of reading, horizontal equating items between 2021 and 2009 for Year 3 paper students

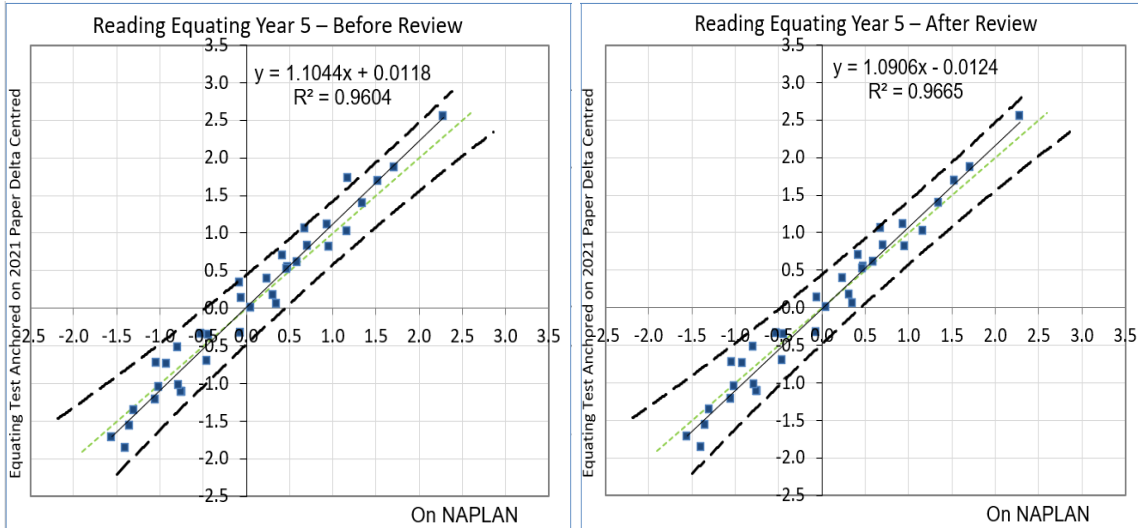


Figure 54. Scatterplot of reading, horizontal equating items between 2021 and 2009 for Year 5 paper students)

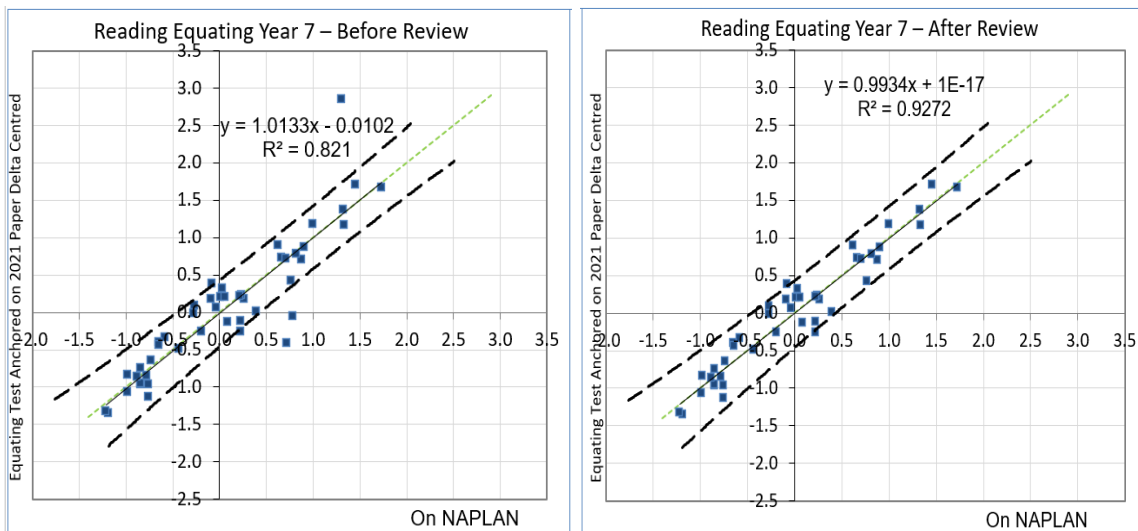


Figure 55. Scatterplot of reading, horizontal equating items between 2021 and 2009 for Year 7 paper students

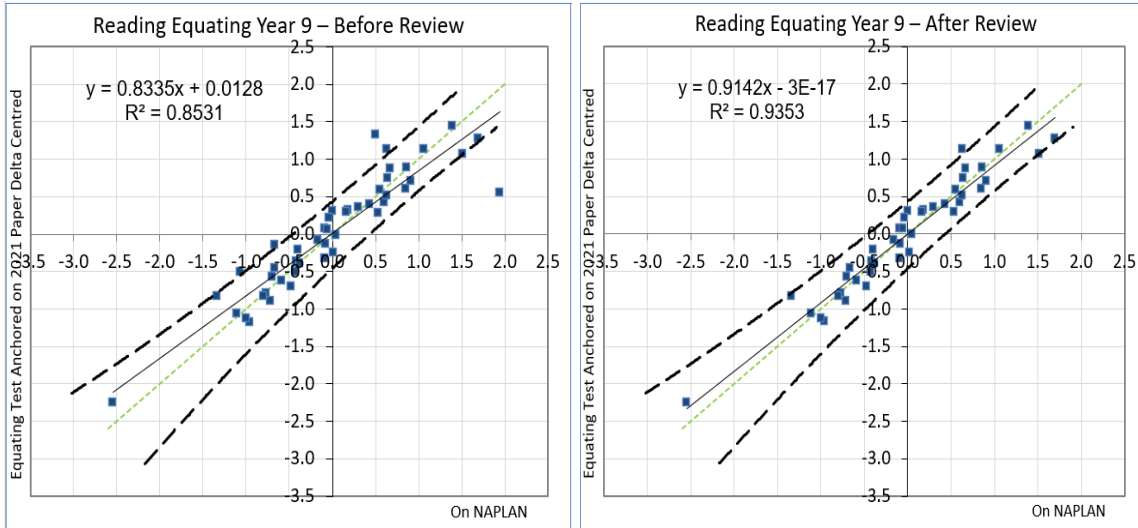


Figure 56. Scatterplot of reading, horizontal equating items between 2021 and 2009 for Year 9 paper students

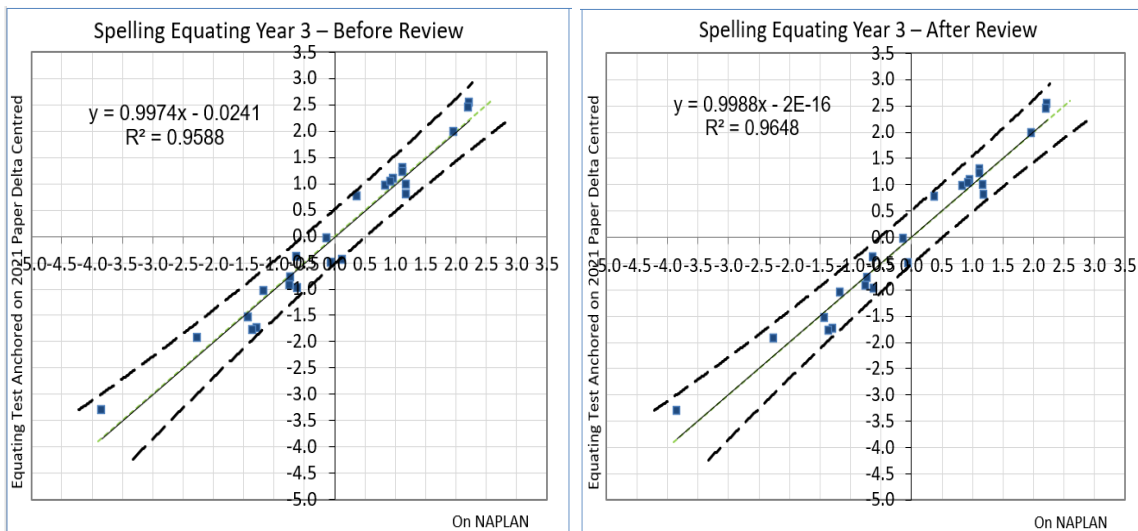


Figure 57. Scatterplot of spelling, horizontal equating items between 2021 and 2009 for Year 3 paper students

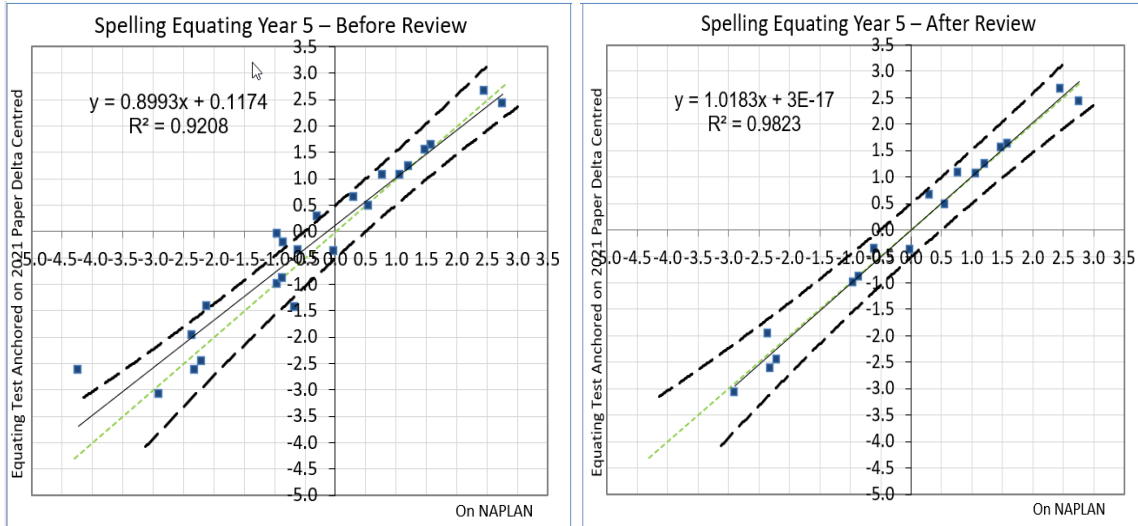


Figure 58. Scatterplot of spelling, horizontal equating items between 2021 and 2009 for Year 5 paper students

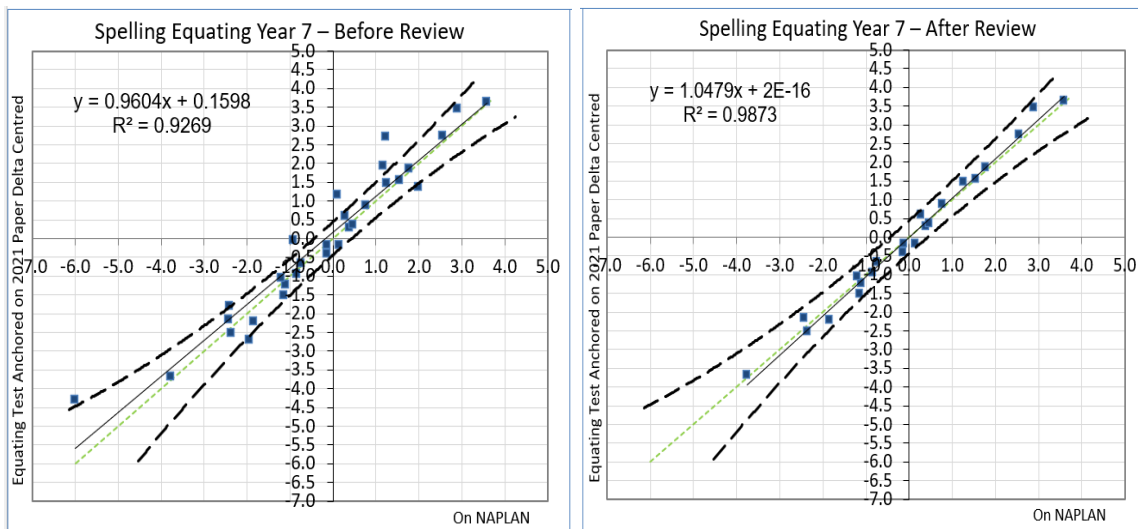


Figure 59. Scatterplot of spelling, horizontal equating items between 2021 and 2009 for Year 7 paper students

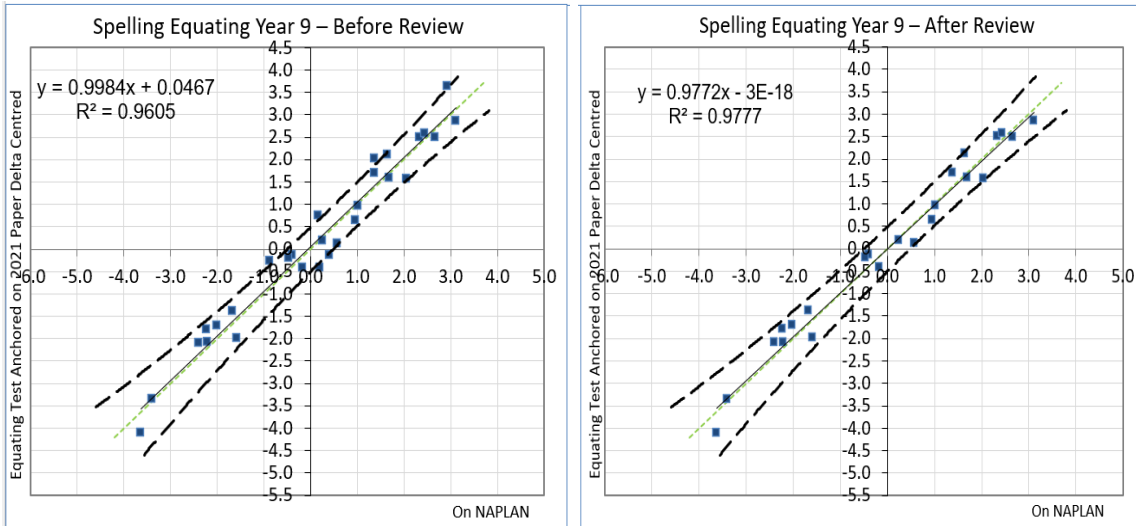


Figure 60. Scatterplot of spelling, horizontal equating items between 2021 and 2009 for Year 9 paper students

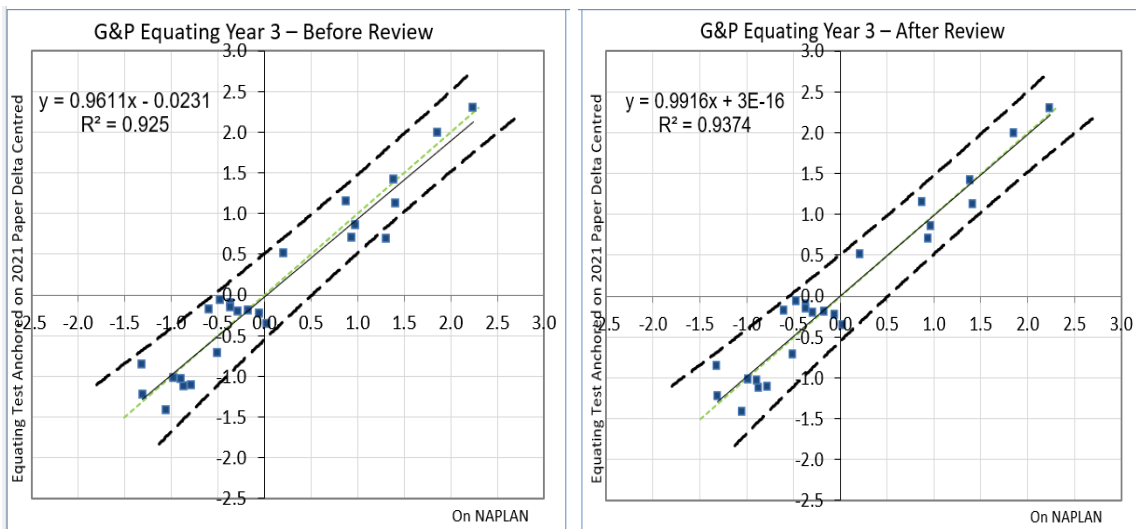


Figure 61. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2009 for Year 3 paper students

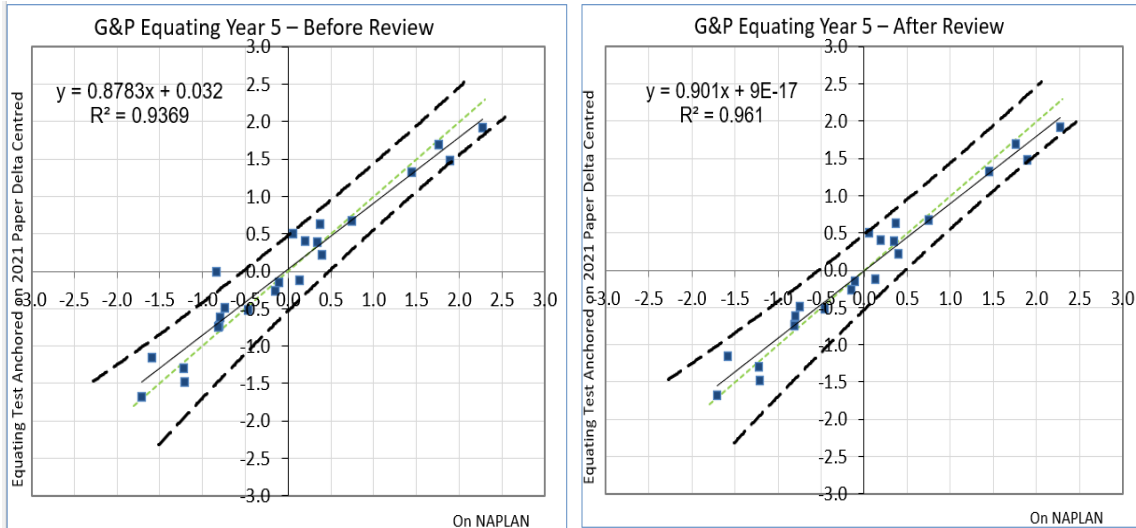


Figure 62. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2009 for Year 5 paper students

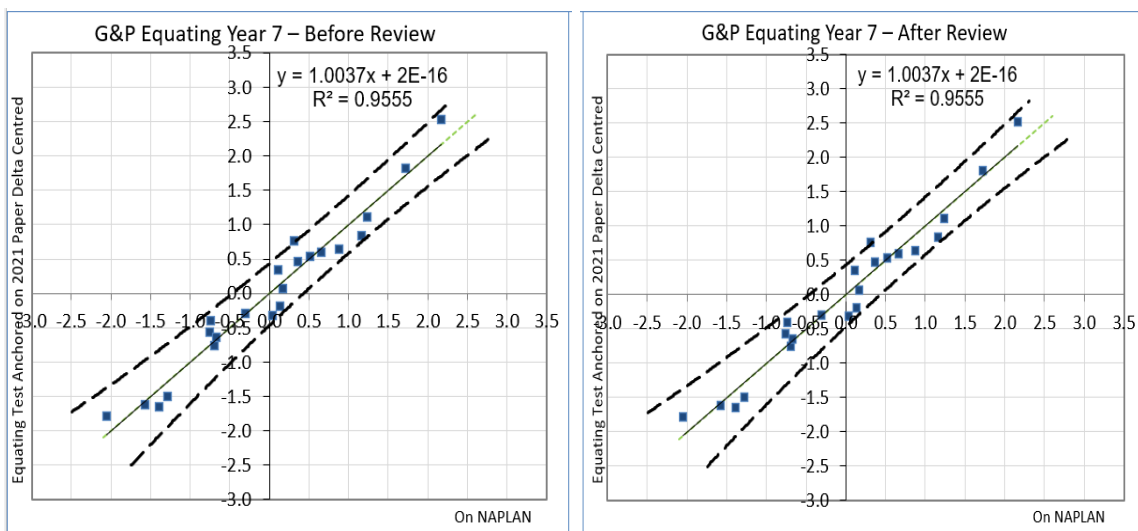


Figure 63. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2009 for Year 7 paper students

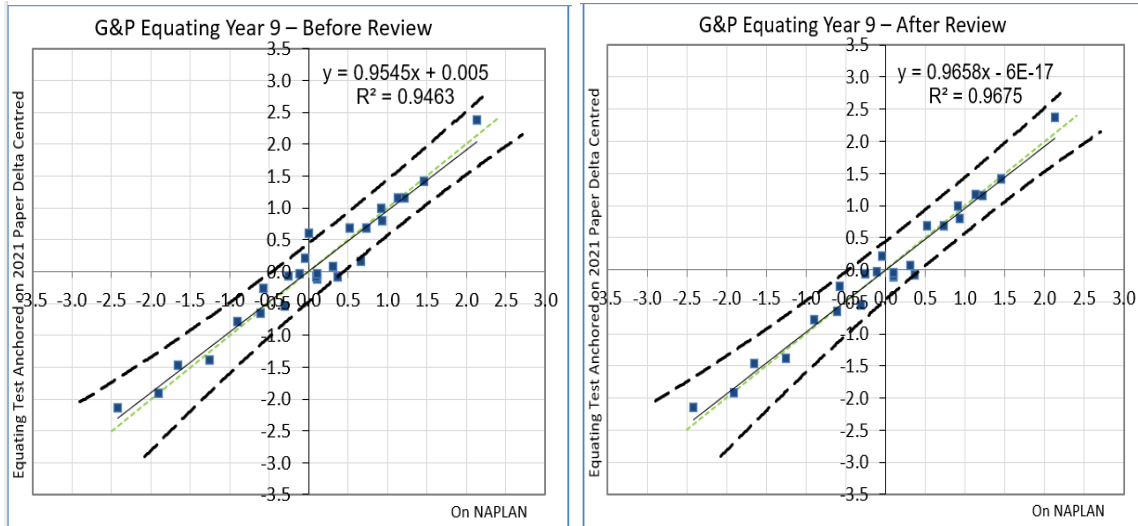


Figure 64. Scatterplot of grammar and punctuation, horizontal equating items between 2021 and 2009 for Year 9 paper students

The numbers of horizontal links used and retained for each test are shown in Table 69.

Table 70 shows the horizontal shift-constants for each domain at each year level by test mode. However, these were not the final shifts to equate the 2021 results onto the historical scale. Instead, these horizontal shifts were subsequently adjusted, using the vertical equating shifts, resulting in the final HVR shifts (see after the vertical equating sections). Appendix K presents the 2021 horizontal link item locations (Rasch difficulty parameters), standard errors, and differences in the item locations by domain for each adjacent pair of year levels.

Table 69. Horizontal link review summary for paper tests

Year level	Numeracy	Reading	Spelling	Grammar and punctuation
3	31/35	31/35	23/24	24/25
5	36/40	35/37	17/23	22/23
7	54/64	44/47	22/30	22/22
9	56/62	43/47	23/29	24/26

Table 70. Horizontal equating shifts between 2021 item locations and item locations on the historical NAPLAN scale for paper tests

Year level	Numeracy	Reading	Spelling	Grammar and punctuation
3	-0.722	-0.381	-0.940	0.104
5	0.516	0.700	1.205	1.096
7	1.162	1.511	1.995	1.462
9	2.053	1.656	3.415	1.787

Vertical equating shifts of the online tests

As in previous years of testing, the NAPLAN 2021 numeracy, reading, spelling, and grammar and punctuation tests were vertically linked across Years 3, 5, 7 and 9 by common items embedded in tests in adjacent year levels; that is, Year 3 and Year 5, Year 5 and Year 7, and Year 7 and Year 9 in both the online tests and paper tests.

The vertical scales were originally established in 2008. In each new calendar year, common items are included in the tests for adjacent year levels and new vertical equating shifts are estimated using the common items that work well as link items (that is, common items that show equivalent psychometric properties across year levels). While the vertical equating shifts are not strictly necessary for placing the NAPLAN 2021 results on the historical scale – because the horizontal shifts place each year level onto the common historical scale for all year levels – the vertical shifts are used to check and improve the horizontal shifts. In 2021, although vertical links were embedded in the online test forms across four non-writing domains, only horizontal equating was used for the online tests (i.e. no HVR adjustments). HVR adjustments, however, were applied to the horizontal paper equating shifts.

The 12 plots of the vertical equating for the online tests are shown in Figure 65 to Figure 76.



Figure 65. Scatterplot for vertical link item review for numeracy between Year 3 and Year 5 online tests

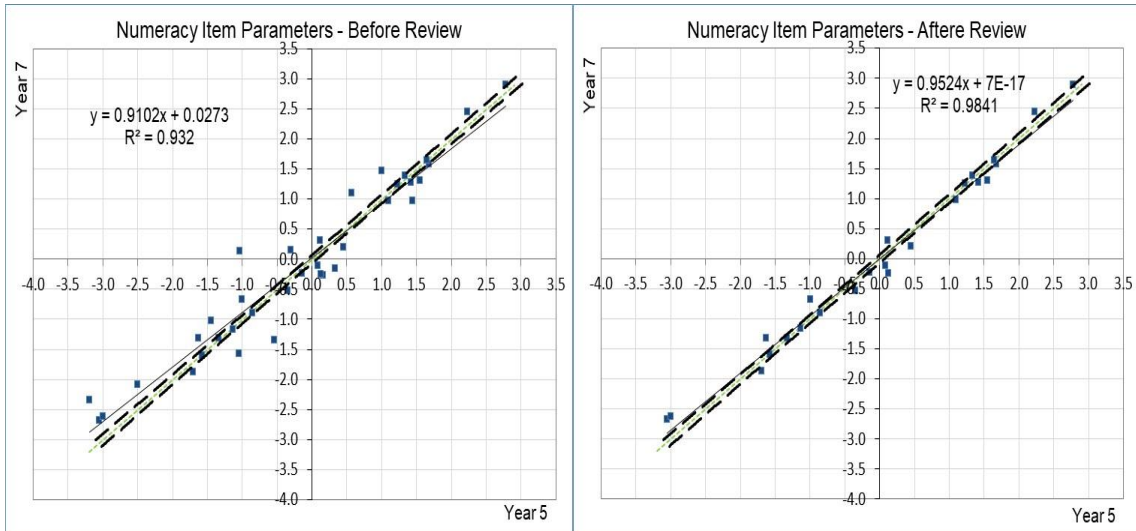


Figure 66. Scatterplot for vertical link item review for numeracy between Year 5 and Year 7 online tests

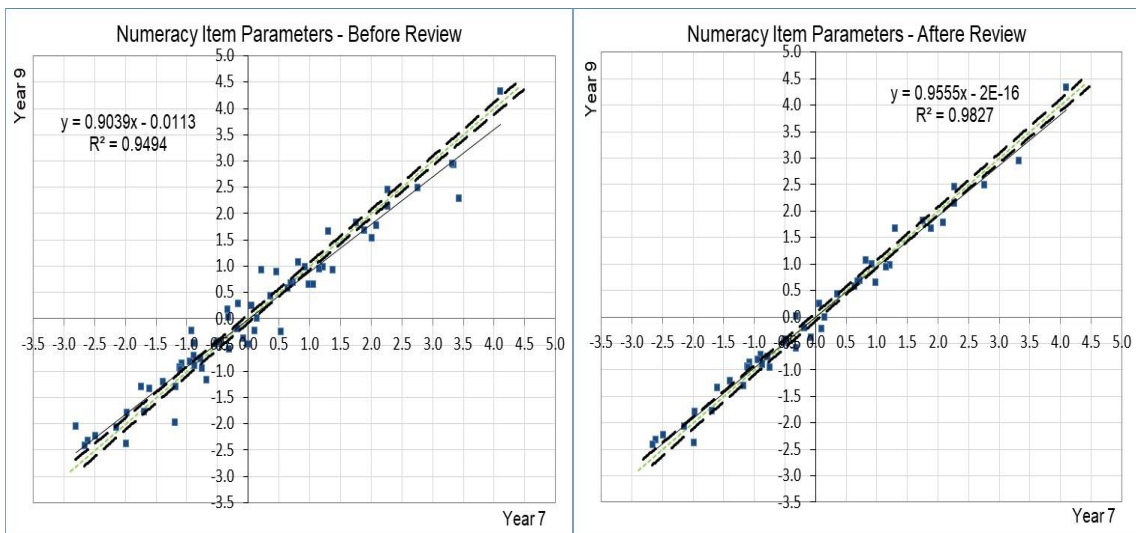


Figure 67. Scatterplot for vertical link item review for numeracy between Year 7 and Year 9 online tests

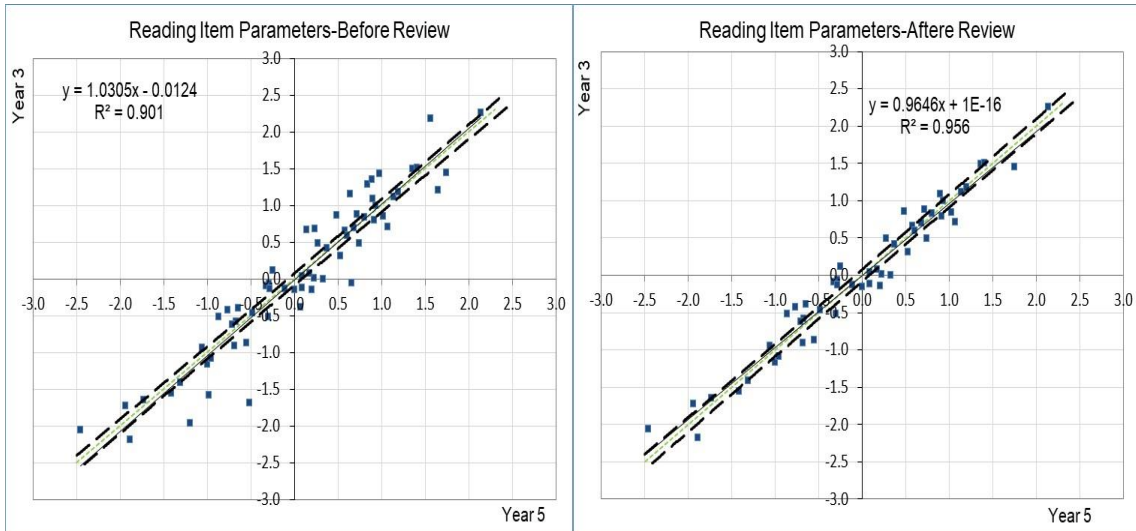


Figure 68. Scatterplot for vertical link item review for reading between Year 3 and Year 5 online tests

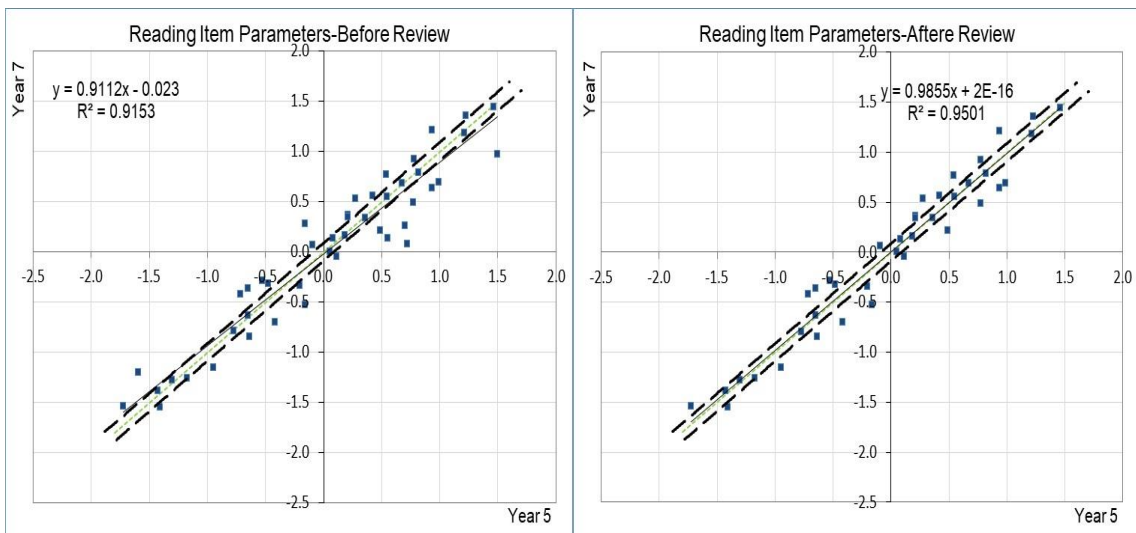


Figure 69. Scatterplot for vertical link item review for reading between Year 5 and Year 7 online tests

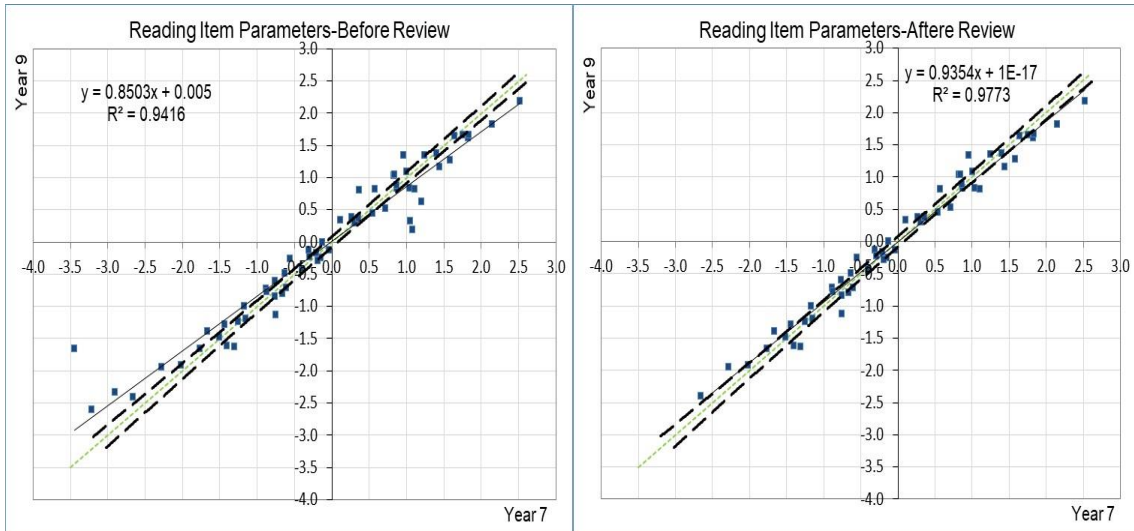


Figure 70. Scatterplot for vertical link item review for reading between Year 7 and Year 9 online tests

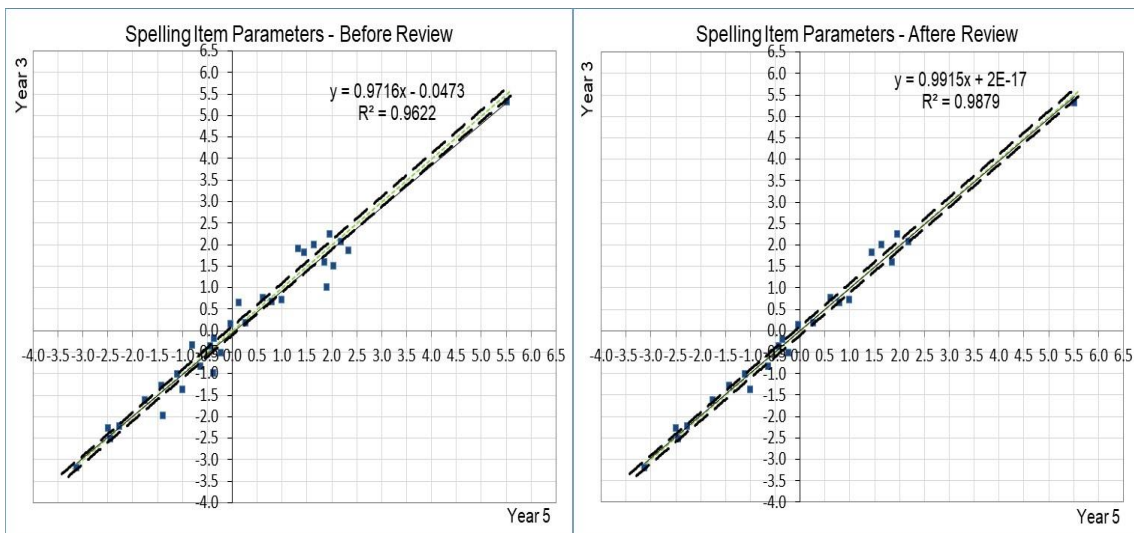


Figure 71. Scatterplot for vertical link item review for spelling between Year 3 and Year 5 online tests

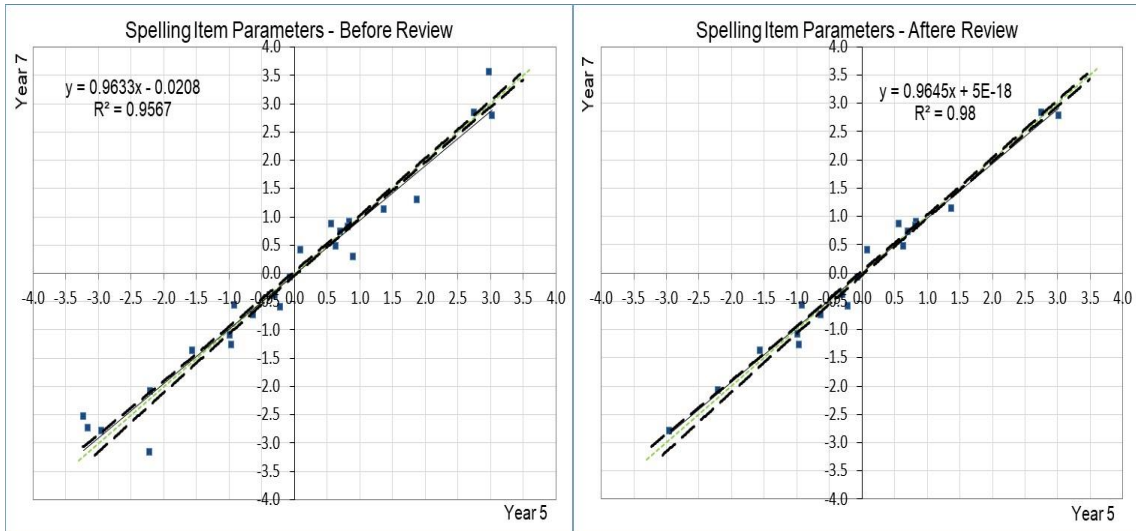


Figure 72. Scatterplot for vertical link item review for spelling between Year 5 and Year 7 online tests

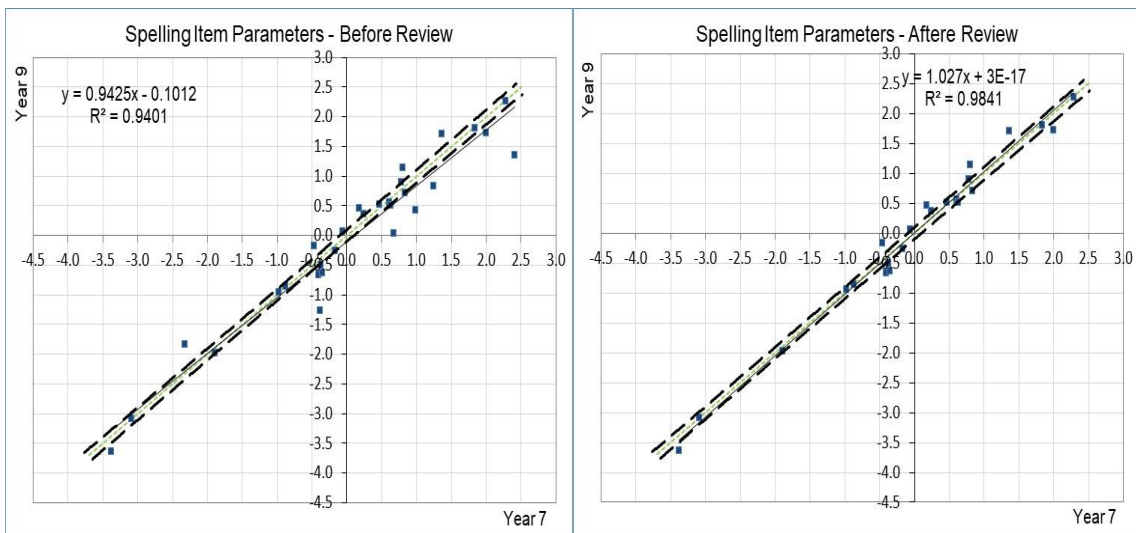


Figure 73. Scatterplot for vertical link item review for spelling between Year 7 and Year 9 online tests

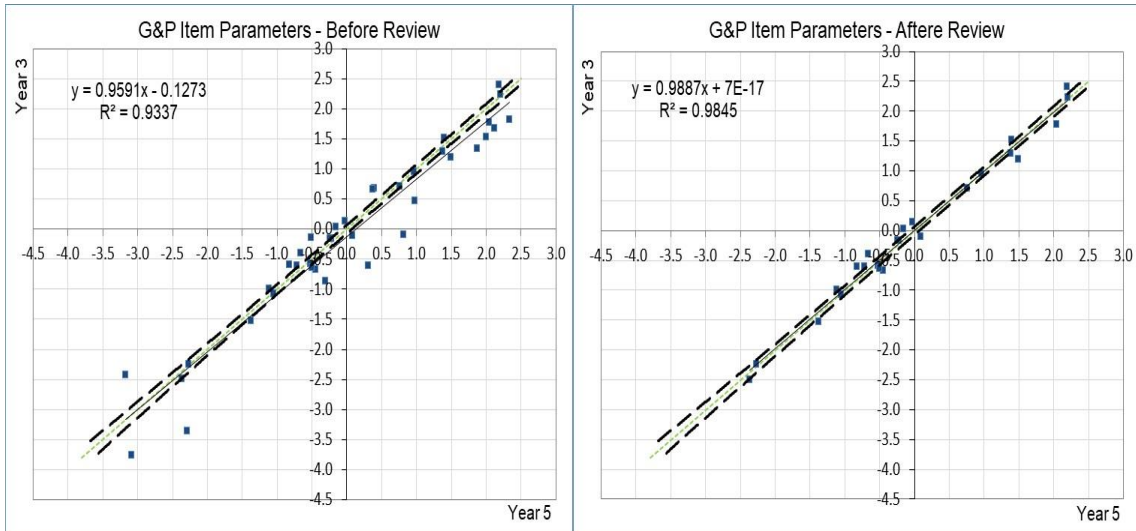


Figure 74. Scatterplot for vertical link item review for grammar and punctuation between Year 3 and Year 5 online tests

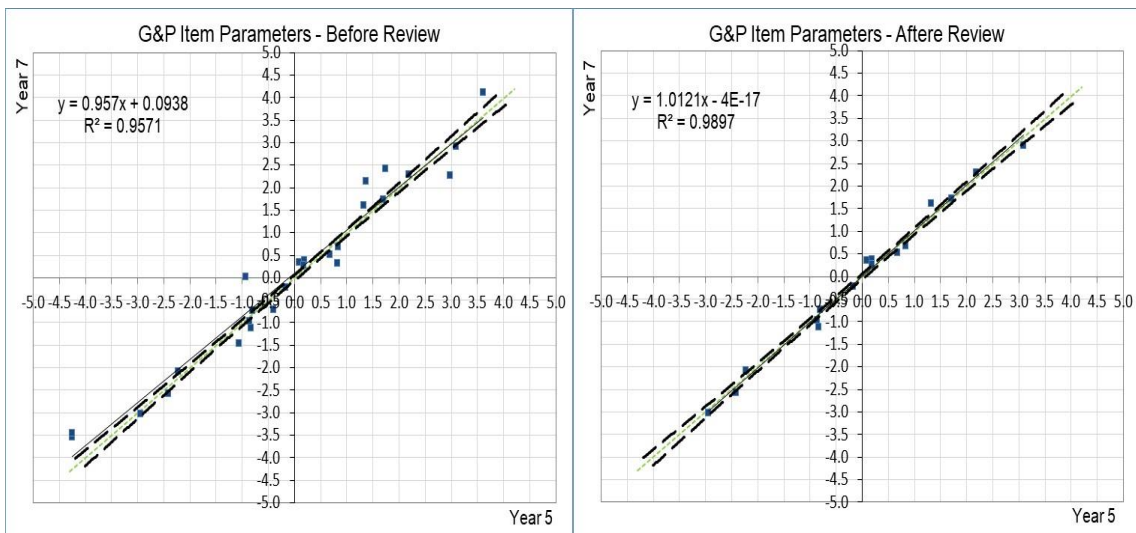


Figure 75. Scatterplot for vertical link item review for grammar and punctuation between Year 5 and Year 7 online tests

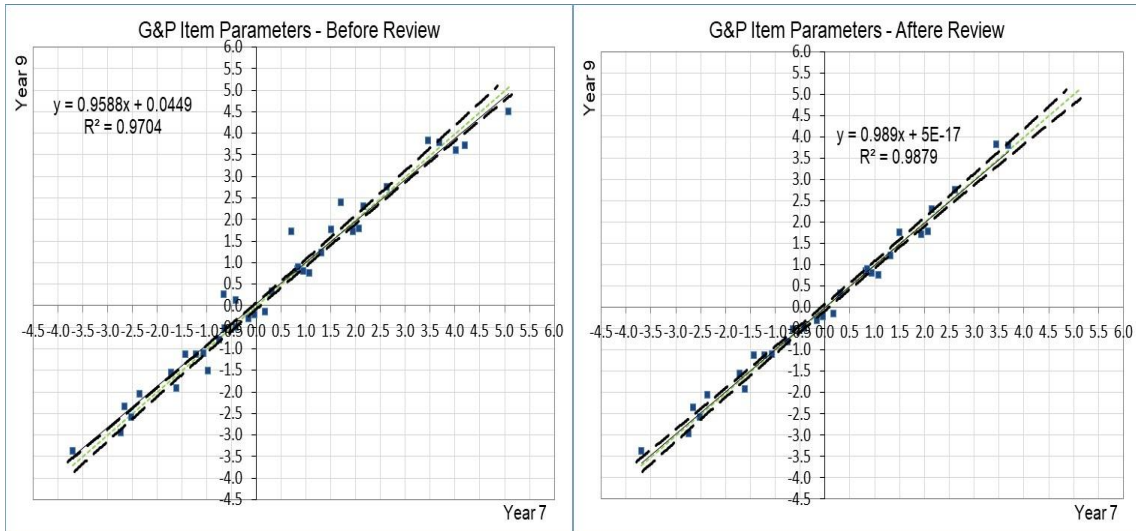


Figure 76. Scatterplot for vertical link item review for grammar and punctuation between Year 7 and Year 9 online tests

Vertical link item review of paper tests

The 12 plots of the vertical equating for the online tests are shown in Figure 77 to Figure 88. As there were not many common items in the paper tests between year levels, it was agreed to maximize the number of links retained, where possible, as in previous years.

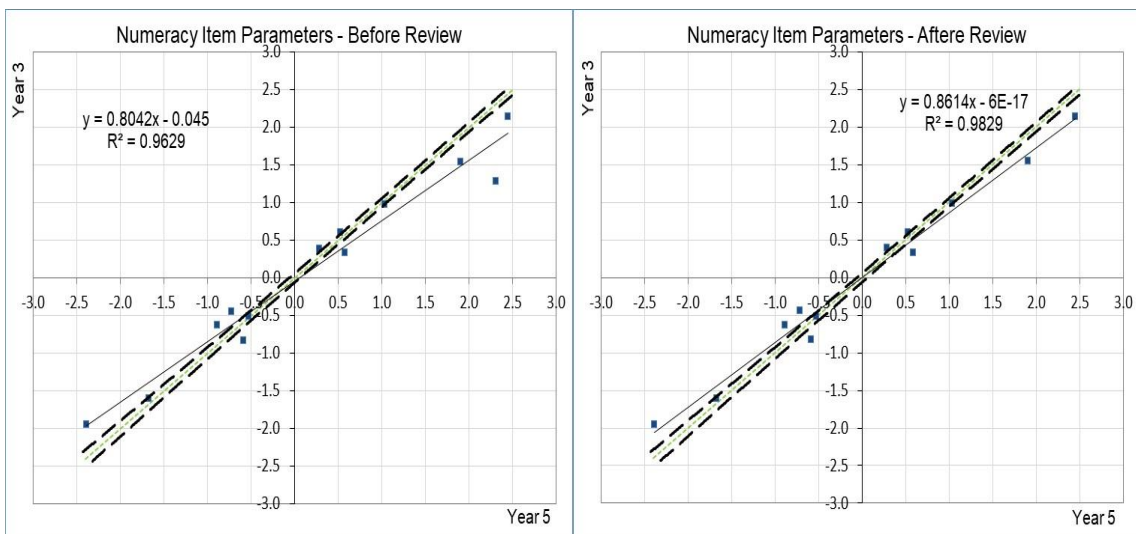


Figure 77. Scatterplot for vertical link item review for numeracy between Year 3 and Year 5 paper tests

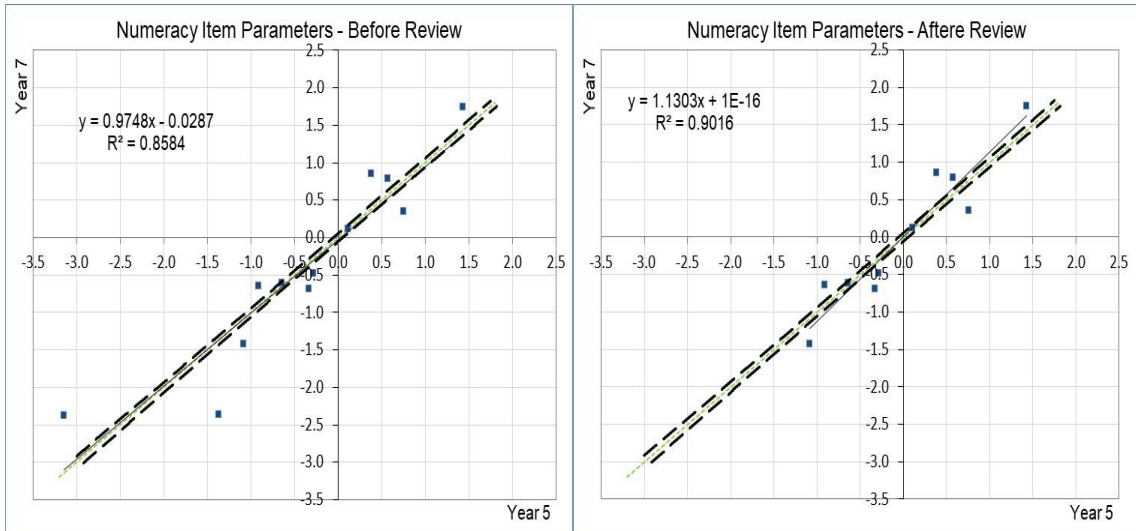


Figure 78. Scatterplot for vertical link item review for numeracy between Year 5 and Year 7 paper tests

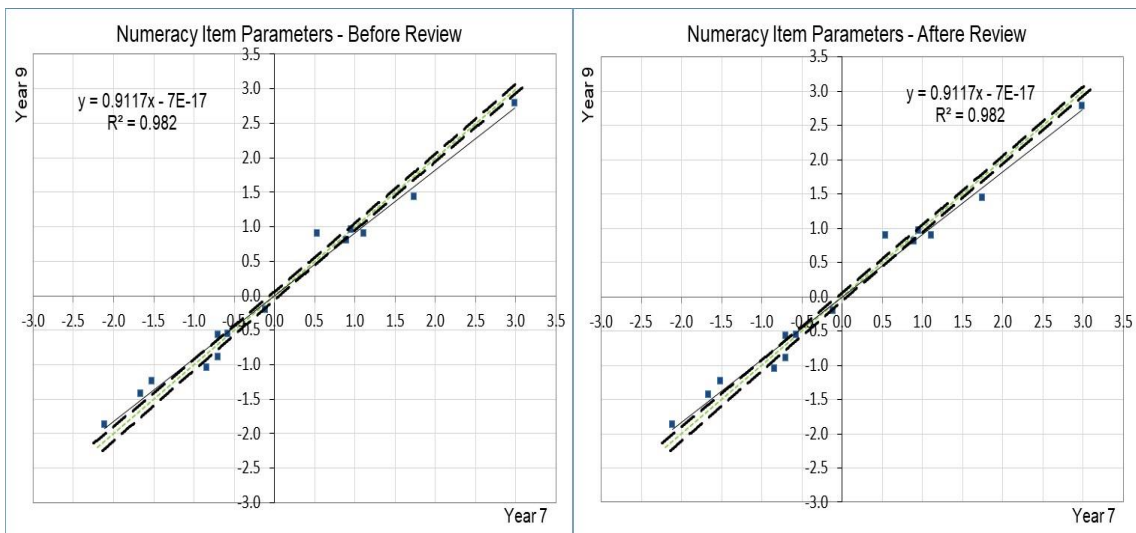


Figure 79. Scatterplot for vertical link item review for numeracy between Year 7 and Year 9 paper tests

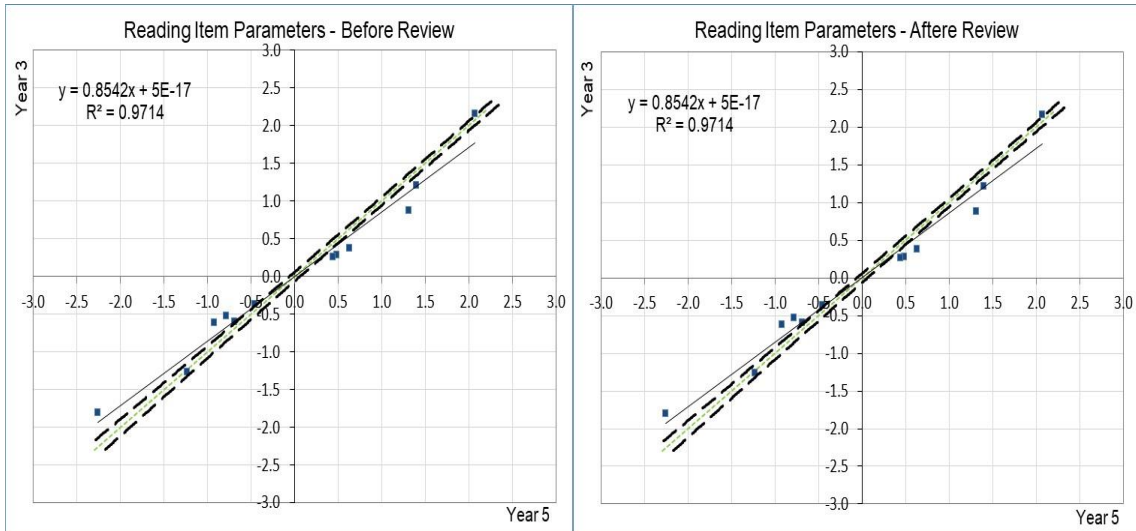


Figure 80. Scatterplot for vertical link item review for reading between Year 3 and Year 5 paper tests

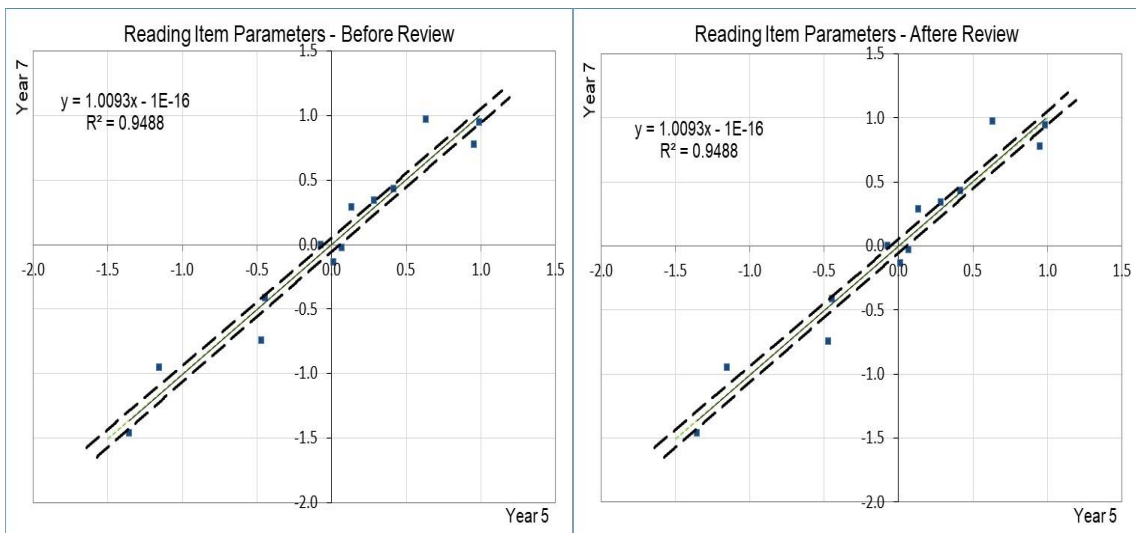


Figure 81. Scatterplot for vertical link item review for reading between Year 5 and Year 7 paper tests

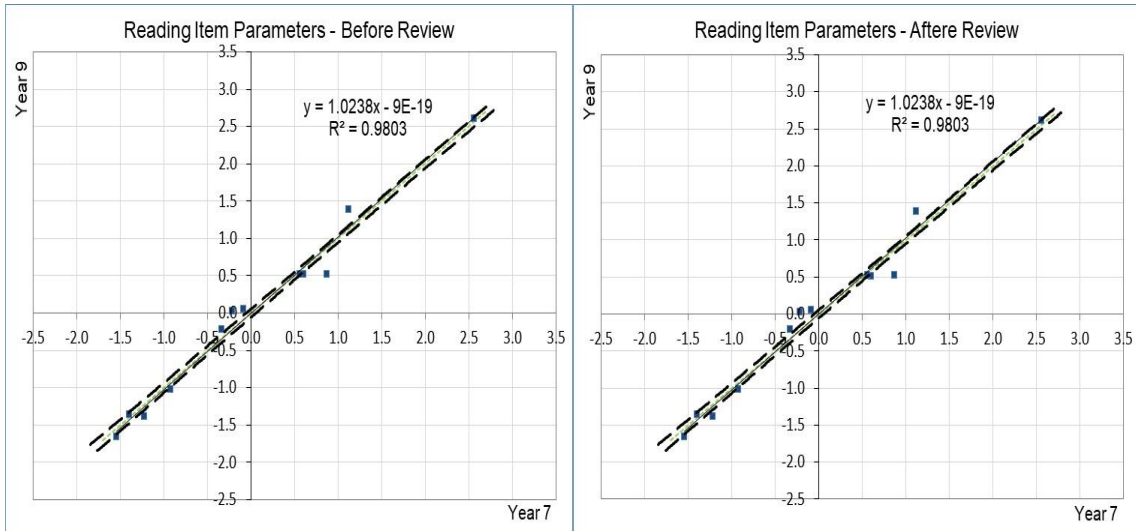


Figure 82. Scatterplot for vertical link item review for reading between Year 7 and Year 9 paper tests

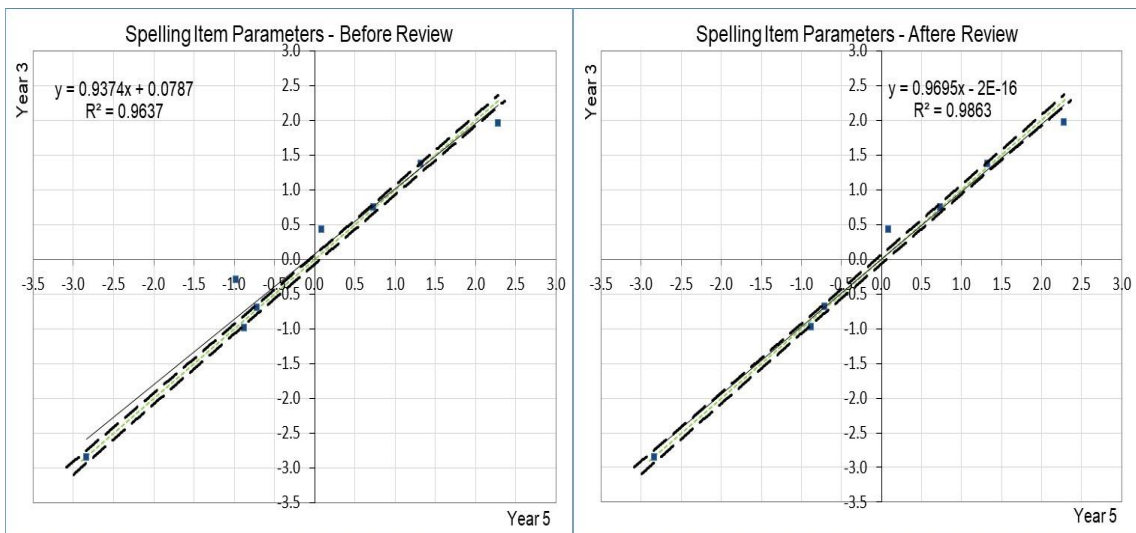


Figure 83. Scatterplot for vertical link item review for spelling between Year 3 and Year 5 paper tests

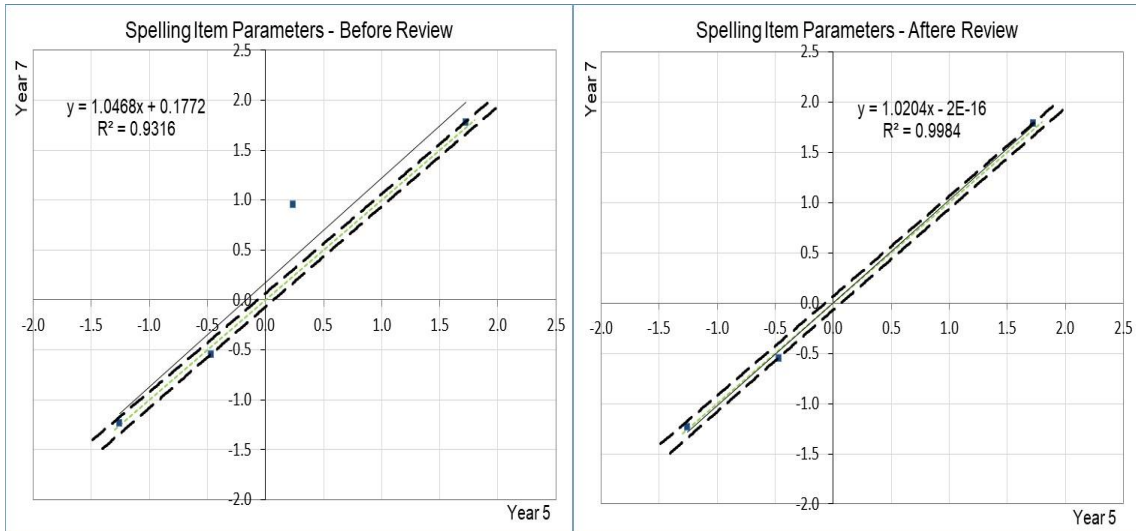


Figure 84. Scatterplot for vertical link item review for spelling between Year 5 and Year 7 paper tests

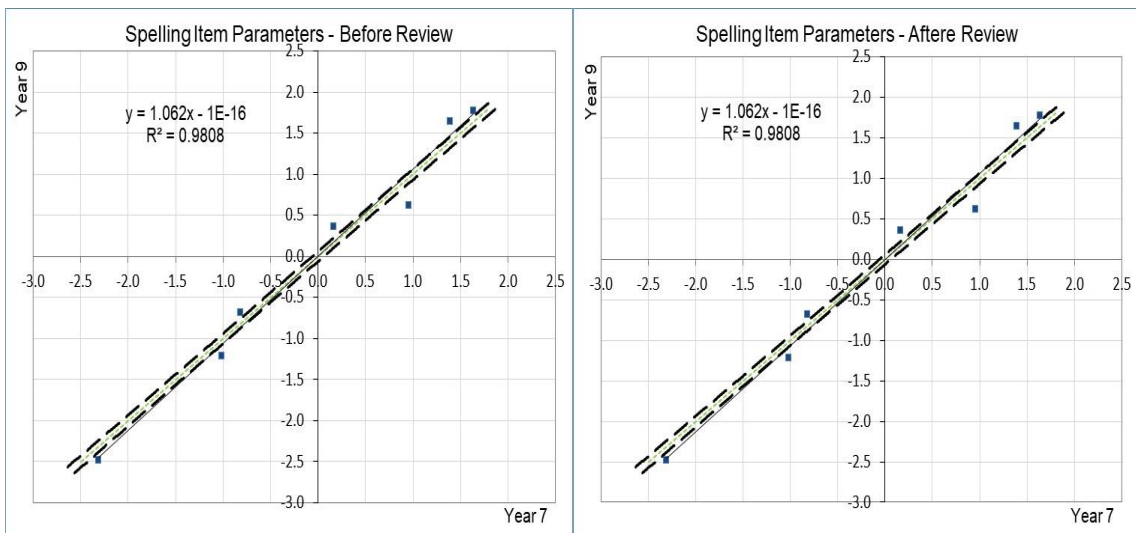


Figure 85. Scatterplot for vertical link item review for spelling between Year 7 and Year 9 paper tests

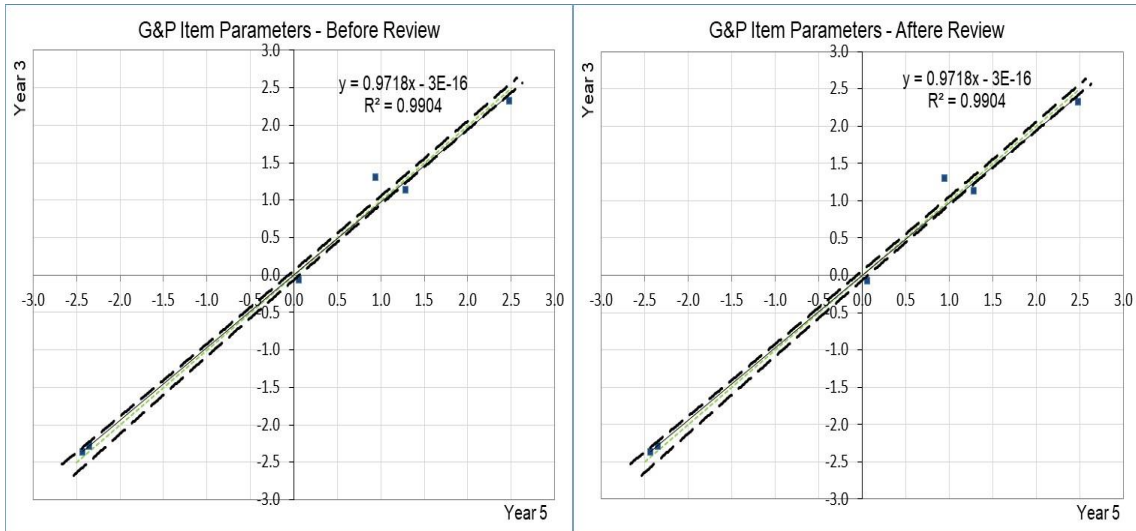


Figure 86. Scatterplot for vertical link item review for grammar and punctuation between Year 3 and Year 5 paper tests

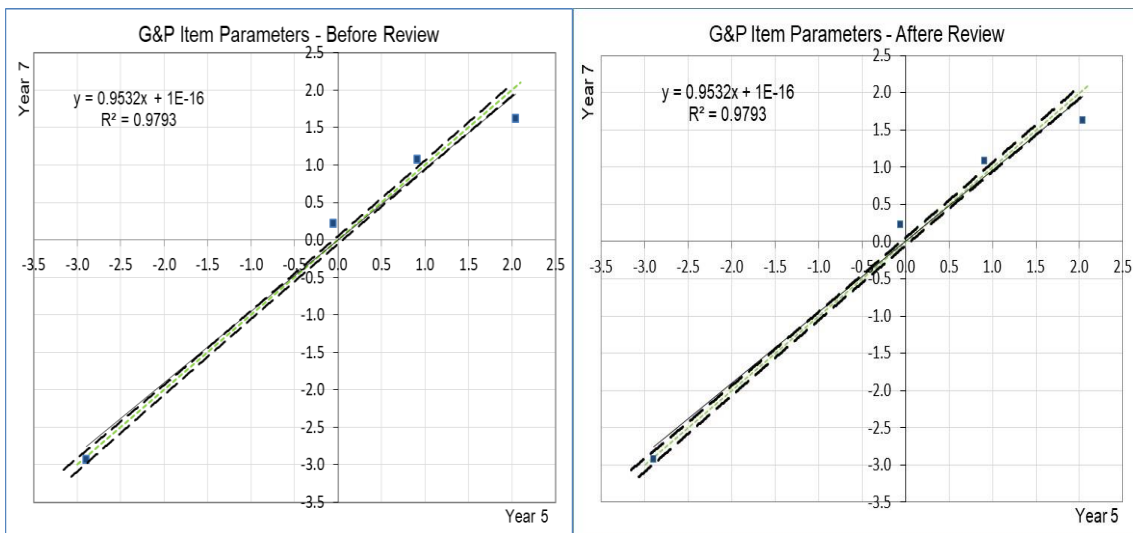


Figure 87. Scatterplot for vertical link item review for grammar and punctuation between Year 5 and Year 7 paper tests

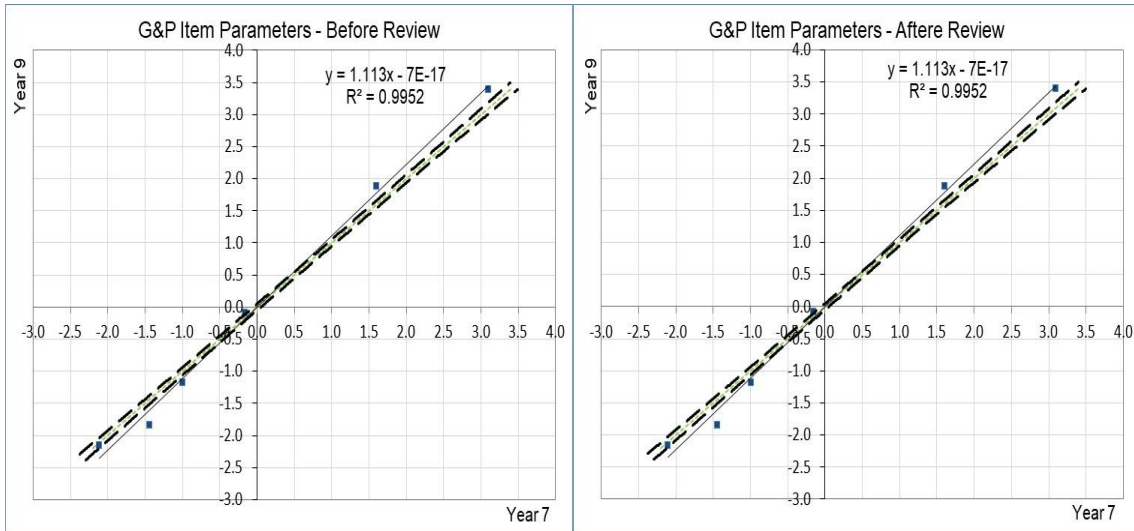


Figure 88. Scatterplot for vertical link item review for grammar and punctuation between Year 7 and Year 9 paper tests

The numbers of vertical links used and retained for each adjacent pair of year levels are shown in Table 71 by test mode. Appendix L presents the 2021 vertical link item locations (Rasch difficulty parameters), standard errors, and differences in the item locations by domain for each adjacent pair of year levels.

Table 71. Vertical link review summary

Test mode	Link	Numeracy	Reading	Spelling	Grammar and punctuation
Online	Years 3 to 5	31/45	51/64	23/32	24/38
	Years 5 to 7	24/36	39/45	19/25	16/26
	Years 7 to 9	46/63	54/61	23/29	29/37
Paper	Years 3 to 5	12/13	12/12	7/8	6/6
	Years 5 to 7	10/12	13/13	3/4	4/4
	Years 7 to 9	14/14	12/12	7/7	6/6

The mean shifts between two adjacent year levels for each of the four domains are shown in Table 72 and mean shifts between each year level and Year 5 are shown in Table 73.

Table 72. Vertical shift constants between adjacent year levels

Test mode	Shift	Numeracy	Reading	Spelling	Grammar and punctuation
Online	Years 3 to 5	-1.207	-0.949	-2.098	-0.871
	Years 7 to 5	0.849	0.689	0.987	0.675
	Years 9 to 7	0.693	0.262	1.138	0.477
Paper	Years 3 to 5	-1.220	-1.247	-1.711	-1.202
	Years 7 to 5	0.591	0.997	1.358	0.470
	Years 9 to 7	0.615	0.082	1.274	-0.099

Table 73. Vertical shift constants from each year level to Year 5

Test mode	Shift	Numeracy	Reading	Spelling	Grammar and punctuation
Online	Years 3 to 5	-1.207	-0.949	-2.098	-0.871
	Years 5 to 5	0.000	0.000	0.000	0.000
	Years 7 to 5	0.849	0.689	0.987	0.675
	Years 9 to 5	1.542	0.951	2.125	1.152
Paper	Years 3 to 5	-1.220	-1.247	-1.711	-1.202
	Years 5 to 5	0.000	0.000	0.000	0.000
	Years 7 to 5	0.591	0.997	1.358	0.470
	Years 9 to 5	1.207	1.079	2.632	0.371

The final equating parameters to place the 2021 paper tests on each of the historical NAPLAN domain scales were determined by taking both the horizontal equating shifts and the vertical equating shifts into consideration. The procedure and results are described in the following section.

Horizontal–vertical regression (HVR) equating shifts (paper tests)

The NAPLAN historical scale spanning Years 3, 5, 7 and 9 was established in 2008 through vertical equating of the year level tests. The horizontal equating tests for each year level provided one basis for placing the NAPLAN 2021 tests on the historical scale for each domain. The horizontal equating tests were first used in 2009 and reused every subsequent year.

Table 66 depicts the horizontal and vertical equating design schematically. In principle, each year level test can be equated directly onto the NAPLAN scale through the horizontal equating shifts without the vertical equating shifts. The vertical equating shifts, however, serve as a quality assurance check and as a tool to fine tune the horizontal shifts using the predicted values from a regression analysis of the horizontal shifts onto the vertical

First, vertical shifts are calculated from each year level to the Year 5 scale. The shifts in the second column of Table 74 are equal to the shifts presented in Table 70. These shifts are transformed in column three by subtracting the Year 5 horizontal shift from each of the year level horizontal shifts. If both horizontal and vertical equating shifts were error free,

columns one and three should be identical. In this example, there are some noticeable differences.

Table 74. Example of comparing horizontal shifts with vertical shifts (numeracy, paper test)

	2021 vertical shift to Year 5	Horizontal shift 2021 to NAPLAN	Adjusted horizontal shift	Predicted horizontal shift
Year 3	-1.220	-0.722	-1.237	-0.782
Year 5	0.000	0.516	0.000	0.590
Year 7	0.591	1.162	0.646	1.255
Year 9	1.207	2.053	1.538	1.947

Therefore, the horizontal shifts in column two (Y) were regressed onto the vertical shifts in column one (X). A scatterplot of these shifts as applied to the paper tests are presented by domain in Figure 89. The broken line represents the regression line. The Y-coordinates of the dots are the observed horizontal shifts. The predicted values of these shifts lie on the regression line. The predicted values were the HVR equating shifts used to place the NAPLAN 2021 paper test results onto the historical scale.

Figure 89 shows the plots of the positions of the four 2021 tests (Years 3, 5, 7 and 9), based on the horizontal equating (vertical axes), against their relative positions centred at Year 5, based on the common-item vertical equating (horizontal axes), for paper tests. The regression equation and *R*-square are shown at the top of each plot. There is one plot for each of reading, spelling and numeracy by test mode, and one plot for grammar and punctuation paper tests.

Ideally, each regression line would have a slope of 1.0 and pass through all four points, showing perfect correspondence of the two methods. It can be seen from the plots that this is not always the case. For the paper tests, the best fit lines for reading, spelling and numeracy show that the horizontal equating and vertical equating align well, the correlation between the vertical and horizontal equating shifts were close to one, although Year 5 spelling showed a slight deviation away from the line. For grammar and punctuation paper tests, although the correlation between the vertical shifts and horizontal shifts is lower than that for the other three domains, there was no particular year level that stood out as an outlier. These regression shifts were used for final equating of the 2021 paper tests to the NAPLAN historical scale.

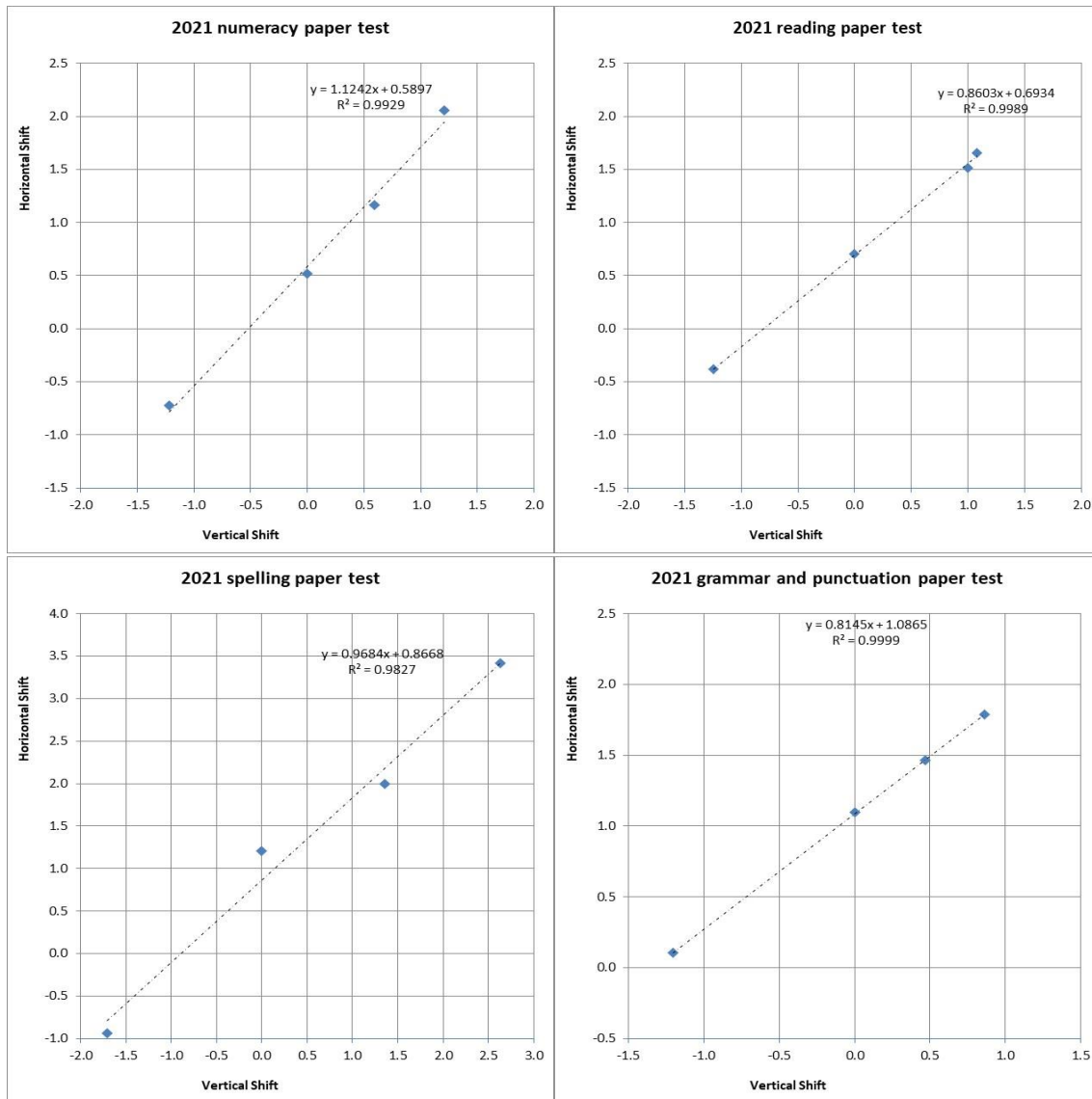


Figure 89. Comparisons of horizontal and vertical shifts of the paper tests

Table 75 displays the intercepts and slopes for the regression-based combination of the vertical and horizontal equating shifts for paper tests.

Table 75. Regression intercepts and slopes of paper tests

Regression coefficient	Numeracy	Reading	Spelling	Grammar and punctuation
Intercept (a)	0.590	0.693	0.867	1.087
Slope (b)	1.124	0.860	0.968	0.814

As in previous years, the final equating shifts were calculated using the regression lines of best fit:

$$\hat{Y} = a + bX \tag{4}$$

where \hat{Y} is the HVR shift from 2021 onto the historical NAPLAN scale; X is the Year 5 centred shifts based on vertical equating; b is the regression slope; and a is the regression intercept.

In other words, the final equating shift that places the 2021 results for each year level onto the historical scale is equal to the *estimated* horizontal shift from a regression of the *observed* (computed) horizontal shifts onto the *observed* (computed) vertical shifts.

Final shifts

To place the 2021 online tests onto the NAPLAN historical scales, the horizontal shift between 2021 and 2019 tests was first applied and then all equating parameters (ACARA, 2020) were applied that placed the NAPLAN 2019 tests on to the NAPLAN historical scale. The NAPLAN 2021 paper tests were placed on to the NAPLAN historical scale by applying the final regression-based shifts that were calculated using equation $\hat{Y} = a + bX$ (4) for each domain. The final online horizontal shifts and paper HVR shifts are shown in Table 76 by year level.

Table 76. Final shifts applied for equating NAPLAN 2021

Test mode	Year level	Numeracy	Reading	Spelling	Grammar and punctuation
Online (horizontal shifts to 2019)	3	0.08946	-0.12757	0.50267	1.10486
	5	0.02940	-0.02776	0.60279	0.85955
	7	-0.30456	-0.09311	0.25374	0.66904
	9	-0.32354	-0.29216	0.23621	0.61441
Paper (HVR shifts to 2008)	3	-0.78200	-0.37957	-0.79029	0.10715
	5	0.58969	0.69340	0.86677	1.08651
	7	1.25462	1.55132	2.18209	1.46936
	9	1.94652	1.62156	3.41619	1.78709

Scaling factors

Applying a scaling factor is sometimes necessary due to the potential impact that differences in test reliability can have on the spread of student scores. As the NAPLAN tests measure the same construct within a domain, it is expected to result in the same latent distribution for the same group of students. In this case, the scale factor would be very close to 1. However, due to differences in test reliabilities of NAPLAN tests, either between the current year test and previous year test, between the tests across year levels, or between the equating test and the NAPLAN test, the spreads of scores between samples of two equated tests can be quite different for some year levels and domains. The scaling factor was derived as the standard deviation (square root of the latent variance) ratio between the 2021 NAPLAN test and the test to be equated to. A scale factor that was greater than 1.0 indicated that the test to be equated spread the students out more than the 2021 test did for that domain at the particular year level. Conversely, a scale factor that was less than 1.0 indicated that the NAPLAN 2021 test spread the students out more than the equating test for that domain at that particular year level.

In 2021, the scale factors of online tests were estimated using the standard deviation ratio of link items between the NAPLAN 2019 test and the NAPLAN 2021 test. No scaling factor was applied if the ratio was within the range 0.94 to 1.05. Only three online tests, numeracy year 7, spelling year 5 and grammar and punctuation year 3, required a scale factor.

The scale factors of paper tests were estimated between the NAPLAN 2019 test and the NAPLAN 2021 test.

For each domain at each year level, a linear transformation was applied to scores on the delta-centred logit scale to correct for the spread in the scores and to apply the appropriate equating constant to put the scores onto the NAPLAN historical scale. The linear transformation formula applied for each domain at each year level by test mode is given by:

$$\text{TransformedLogitScore} = SF \cdot (\text{LogitScore} - \text{LocalMean}) + \text{LocalMean} + \text{EqShift} \quad (5)$$

where

LocalMean = the mean of the latent distribution estimated using the 2021 calibration sample based on the delta-centred item parameters. As all students have a weight equal to one, no student weights were applied. In other words, by subtracting the local mean, the average of the scale becomes zero. Applying the scaling factor now results in a change in variance only while the mean stays zero. Adding the local mean back recovers the original mean of the scale.

SF = the scale factor is the factor used for correcting the spread of the scores.

EqShift = the equating constant pertinent for the domain at the particular year level (provided in Table 76).

The values for *LocalMean* and *SF* are presented in Table 77 for each year level by domain. Following these transformations, the 2019 equating parameters were then applied to the NAPLAN 2021 online tests.

Table 77. Local means and scaling factors

Domain and year	Online		Paper	
	Local mean	Scale factor	Local mean	Scale factor
N3	0.12854	1.00000	0.04375	1.10615
N5	0.26065	1.00000	0.30951	0.92438
N7	0.23353	0.95819	0.53365	1.00000
N9	0.06790	1.00000	0.41191	0.81621
R3	0.08570	1.00000	0.71441	1.12442
R5	0.25824	1.00000	0.76304	1.02541
R7	0.07500	1.00000	0.38195	1.00000
R9	0.23665	1.00000	0.76346	1.00000
S3	-0.02228	1.00000	-0.05370	1.04731
S5	0.07738	0.97071	0.17112	0.90960
S7	0.29852	1.00000	0.29636	1.00000
S9	-0.08522	1.00000	-0.06482	1.00000
G3	-0.00411	1.15198	0.34426	1.22924
G5	0.24575	1.00000	0.35440	1.00000
G7	0.02195	1.00000	0.31983	1.07673
G9	-0.02288	1.00000	0.62801	1.00000

The same transformation steps were applied to the WLE ability logit scores in the score equivalence tables, the item parameters and the plausible values.

Equating of writing results

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

To review the stability of item difficulty steps, the 2021 data were freely calibrated and compared to the item difficulties of 2019 by test mode. The year 2019 was chosen because the writing genre was narrative in 2021 and in 2019 while the genre was persuasive in 2018 and in 2017. The scatter plot between the two calendar years are shown by test mode in Figure 90. They indicate that the consistency of relative difficulties supported using the anchoring method in 2021.

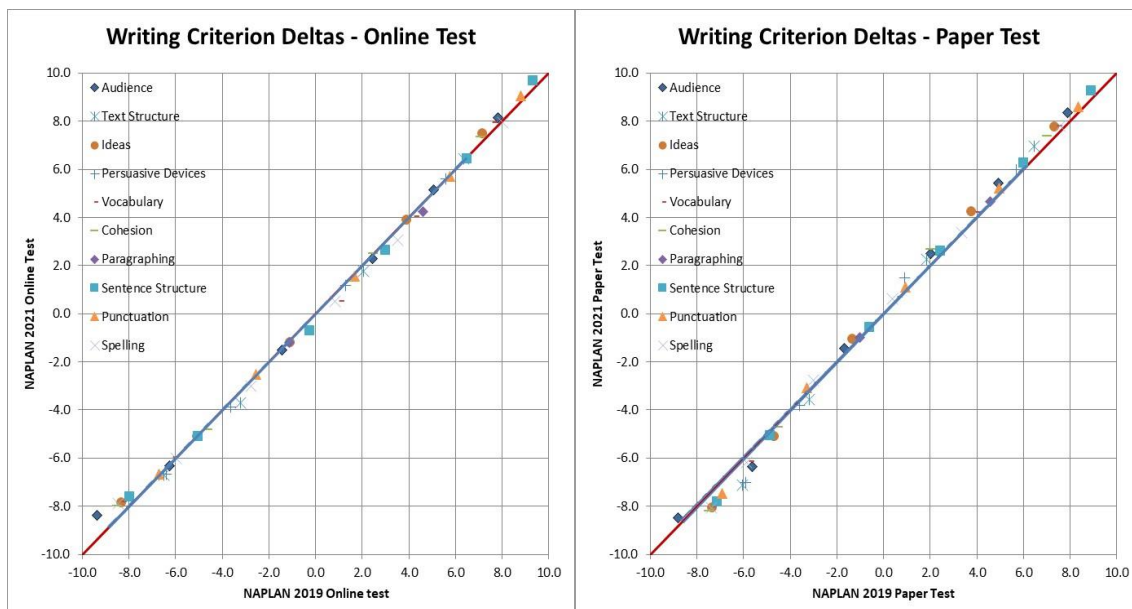


Figure 90. Scatterplot for writing criteria between 2021 and 2019 online and paper tests

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. More information about the pairwise comparison methodology can be found in Humphry & McGrane (2014).

The pairwise study for Writing in the NAPLAN 2021 assessment was designed to equate 2021 performances directly onto the 2016 pairwise scale. The design was similar to that used in 2019, although it did not include an additional pairwise comparison component cross-referencing paper and online performances.

The focus of the 2021 design was to directly align the 2021 performances to the 2016 scale. The 2021 pairwise study showed that 2021 and 2016 paper-based performances scaled together well to form a single scale.

The equating verification involved a set of comparisons between 2021 performances and 2016 performances. 257 paper-based performances from 2016 were compared against 154 paper scripts from 2021 and 360 online performances from 2021. The 2021 set of performances included the following numbers of performances per task: 103 OnT1 scripts, 145 OnT2 scripts, 112 OnT3 scripts and 154 PT1.

Scripts were selected across an approximately uniform score distribution for each task. 37 judges made a total of 25 573 comparisons of 2021 performances against 2016 performances. The statistical fit index Outfit mean square was used to test whether or not each judge agreed (on aggregate) with the consensus of judges. All judges had Outfit values of less than 1.31, apart from one, indicating good consistency of judgements across the set of judges.

A joint 2015/2016/2019/2021 pairwise scale was formed by adding comparisons from previous NAPLAN writing pairwise projects to the judgments from the 2021 pairwise project. 70 713 comparisons in total were analysed using the Bradley-Terry-Luce model to form this scale.

The purpose of the pairwise study is to cross-check other sources of information in the equating of writing. The purpose is to ascertain whether there were differences in rubric marks that are inconsistent with the results of direct comparisons of scripts. The design allows evaluation of whether, for a given scale location based on pairwise comparisons, a similar rubric score was predicted for 2016 and 2021 scripts, and whether a similar rubric score was predicted for 2021 paper and 2021 online scripts. Thus, the purpose of the pairwise study is to obtain a common frame of reference by which to compare marking in 2016 with marking in 2021 (paper and online) as well as to compare 2021 paper marking with 2021 online marking. In particular, the objective is to examine whether there is evidence for differences in marker harshness that might affect the comparability of results.

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

Pairwise Study Results

To evaluate fit to the Bradley-Terry-Luce model used to analyse data, judge outfit indices were calculated after removing extreme observations (comparisons for which the standardized residuals were greater than 7). For the 2021 pairwise study, all but 1 judge had good outfit indices (less than 1.31). The highest judge outfit was 1.506.

Figure 91 shows the plot of pairwise scale locations (x -axis) against locations based on the NAPLAN rubrics (y -axis) for 2016 and 2021 paper scripts separately. The correlation overall is approximately $r = 0.971$ for 2016 paper scripts and $r = 0.937$ for 2021 paper scripts. As can be seen, the fitted curves are somewhat curvilinear as in previous years of the programme. Note that for 2021, the paper prompt was administered to years 3 and 5 students only.

The pairwise scale locations show the ordering of the paper scripts based on direct comparisons whereas the NAPLAN scale locations are based on rubric marking. In the plot, 2016 paper and 2021 paper are highlighted separately. Regression lines are also shown separately for each of these years.

It can be seen in Figure 91 that there is a similar correspondence between the pairwise and NAPLAN rubric scale locations for 2016 and 2021 paper scripts. Rubric locations for

2021 performances are based on the same correspondence table, between raw scores and logits, as the rubric locations for 2016.

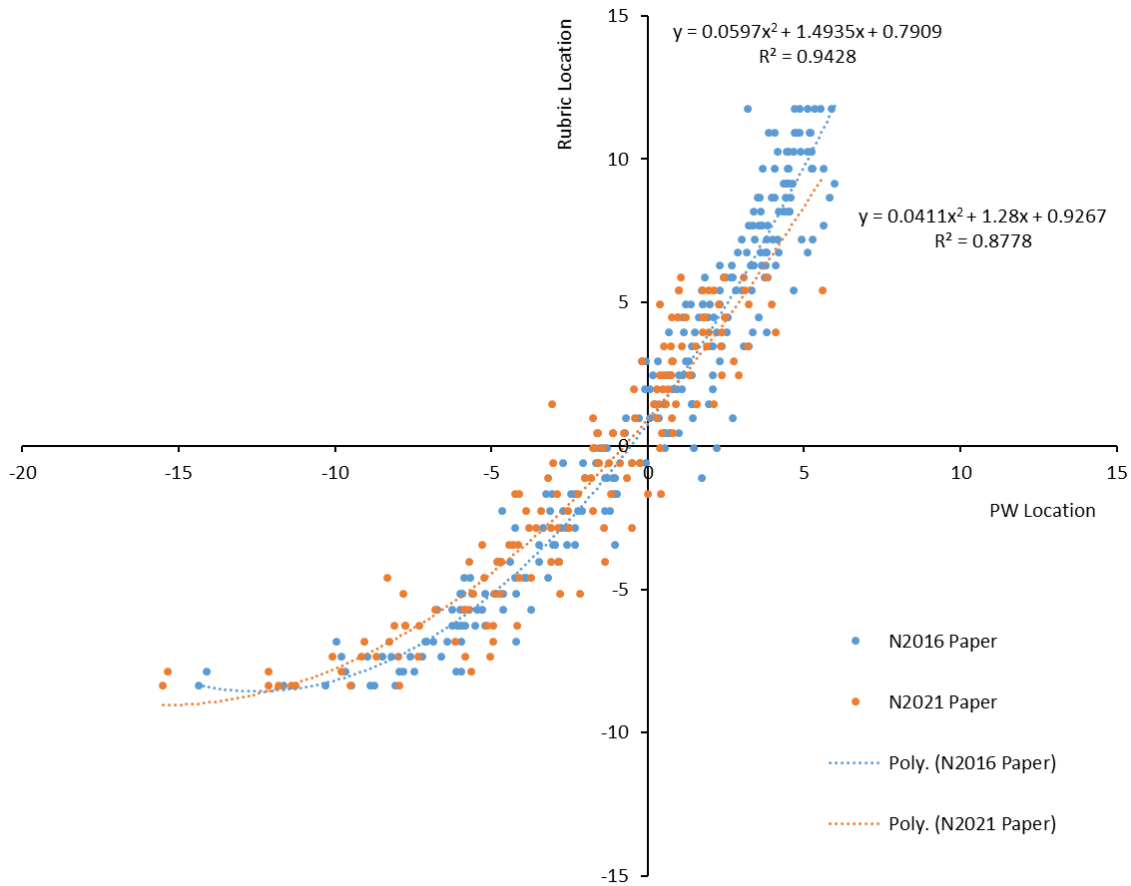


Figure 91. Scatterplot of the NAPLAN rubric and pairwise scale locations for 2016 and 2019 paper performances.

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

Similarly, as shown in Figure 92, 2019 online performances are marked consistently with 2021 online performances. For a given pairwise location, the range of rubric locations for 2019 performances is similar to the range of rubric locations for 2021 performances.

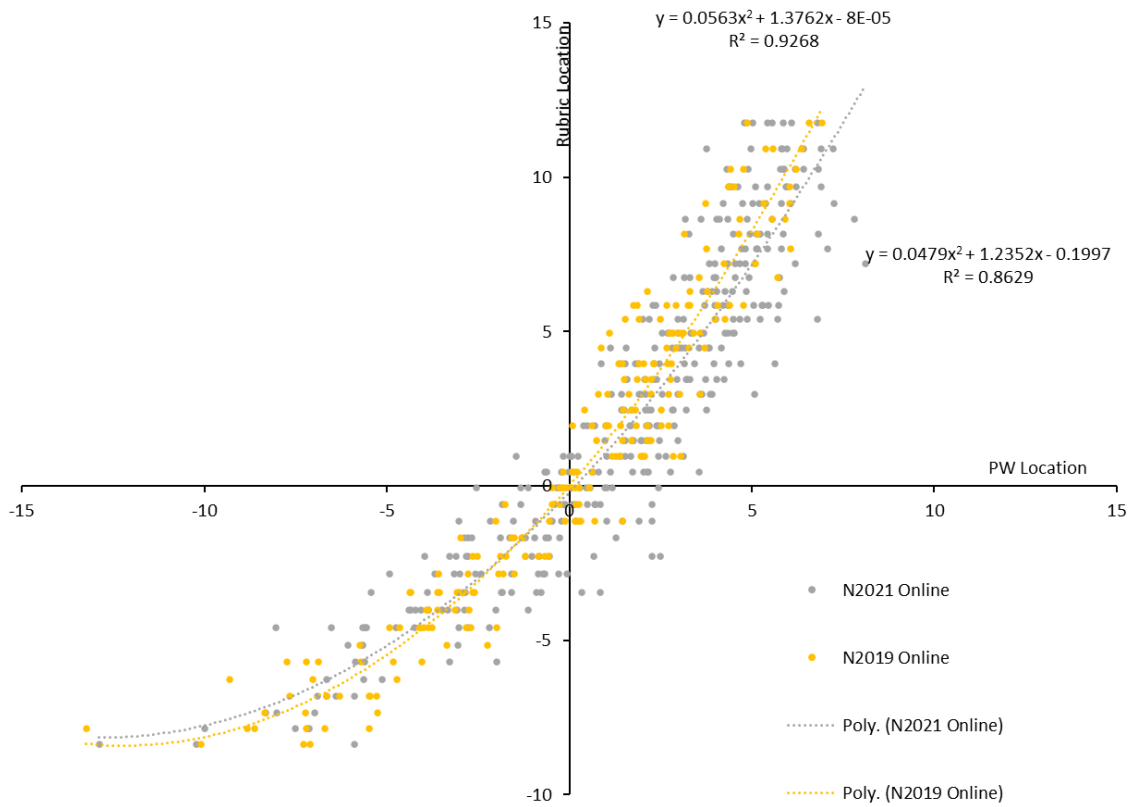


Figure 92. Scatterplot of the NAPLAN rubric and pairwise scale locations for 2021 online performances and 2019 online performances.

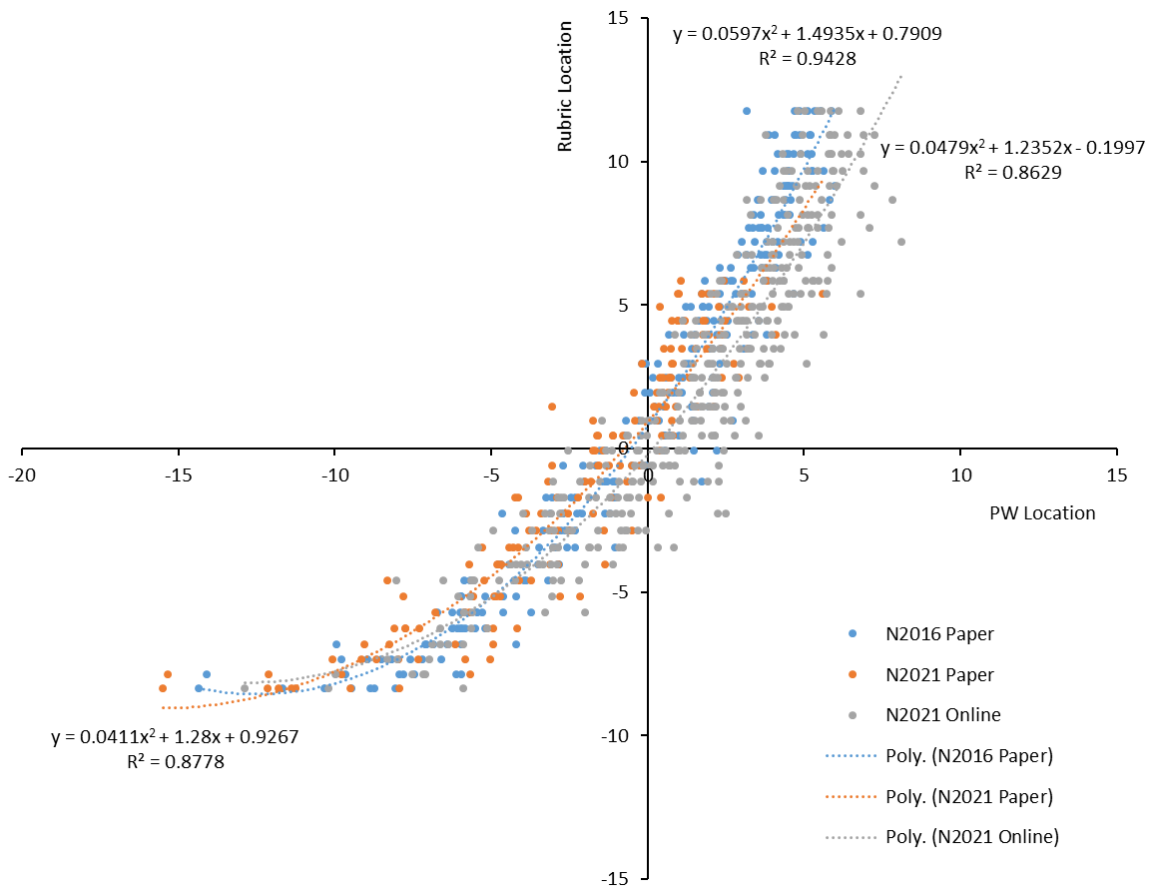


Figure 93. Scatterplot of the NAPLAN rubric and pairwise scale locations, for all 2016 and 2021 performances used in the pairwise equating.

Figure 93 shows the association between pairwise and rubric locations for all 2016 and 2021 performances. Despite the comparability of rubric scores for online and paper performances over time, considered separately, there are differences between the marking of performances administered online compared to marking of performances administered in paper form.

Taking the data on face value, the 2021 online performances appear to have been marked more harshly, particularly at the top end of the range. The results suggest this interpretation because for a given pairwise location on the x-axis, the rubric locations for 2021 online performances tend to be lower than the rubric locations for 2021 paper performances.

However, this interpretation of the results is not consistent with system-level data. Considered in conjunction with historical pairwise information and system-performance data, it is most likely that the nature of sampling of paper performances and associated scores differs in some fashion from the sampling of online performances and associated scores.

In summary, the results of the exercise indicate consistency of marking over time for paper and for online, considered separately. The results show a difference between marking of paper and online, which appears to be an artefact of sampling, possibly combined with other factors.

Standardisation of scales from logits to reporting scales

For each domain, estimates in logits were transformed to the NAPLAN reporting scale scores established in NAPLAN 2008 as follows:

$$NAPLANScaleScore = 100 \cdot (Score_{logit} - DomainMean_{08}) / (DomainStdDeviation_{08}) + 500 \quad (6)$$

where $DomainMean_{08}$ and $DomainStdDeviation_{08}$ were the estimated overall domain mean and domain standard deviation calculated using the 2008 scientific sample. These are presented in Table 78.

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:

$$SE_{NAPLANScale} = 100 \cdot \frac{SE_{logit}}{DomainStdDeviation} \quad (7)$$

Table 78. Domain mean and standard deviation for transforming logits to NAPLAN scale scores

Domain	Domain mean Year 5	Domain SD overall
Numeracy	0.8102	1.6652
Reading	1.1629	1.4867
Writing	1.1160	3.3679
Spelling	0.9406	2.6241
Grammar and punctuation	1.2529	1.3605

Summary of equating parameter estimates for NAPLAN 2021

In 2021, the equating procedures for the NAPLAN results were applied separately for the online and the paper tests. The 2-step formula used for the equating procedures to place the 2021 online results onto the NAPLAN historical scale is:

$$\theta_{21on19}^* = SF_{21}(\theta_{21} - LM_{21}) + LM_{21} + Shift \quad (8)$$

$$\theta_{21}^* = 100 * (SF_{19}(\theta_{21on19}^* - LM_{19}) + LM_{19} + HVR_{19} - MN_{\theta_{Y5,08}}) / SD_{\theta_{All,08}} + 500 \quad (9)$$

Where θ_{21on19}^* is the equated 2021 achievement score onto the 2019 scale, θ_{21}^* is the equated 2021 achievement score onto the NAPLAN historical scale, θ_{21} the original achievement score in logits, SF_{21} and SF_{19} the scaling factor of online test in 2021 or 2019, LM_{21} and LM_{19} the local mean of 2021 and 2019, respectively, $Shift$ 2021 horizontal shift, HVR_{19} 2019 shift for the online tests, $MN_{\theta_{Y5,08}}$ the mean achievement in logit of Year 5 students in 2008, and $SD_{\theta_{All,08}}$ the standard deviation in logits of all year levels in 2008.

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

$$\theta_{21}^{**} = a + b * (\theta_{21}^*)^2 + c * \theta_{21}^* \quad (10)$$

The combined formula for the equating procedures to place the 2021 paper results onto the historical scale, as described in this chapter, is:

$$\theta_{21}^* = 100 * (SF_{21}(\theta_{21} - LM_{21}) + LM_{21} + Shift - MN_{\theta_{Y5_08}}) / SD_{\theta_{All_08}} + 500 \quad (11)$$

Where θ_{21}^* is the equated 2021 achievement score, θ_{21} the original achievement score in logits, SF_{21} the scaling factor of paper test, LM_{21} the local mean, *Shift* HVR shift for the paper tests, $MN_{\theta_{Y5_08}}$ the mean achievement in logit of Year 5 students in 2008, and $SD_{\theta_{All_08}}$ the standard deviation in logits of all year levels in 2008.

Table 79. Summary of parameters for transforming the 2021 logit scores to the NAPLAN reporting scales

Mode	Domain & year	LM ₂₁	SF ₂₁	Shift	LM ₁₉	SF ₁₉	HVR	MN08	SD08	a	b	c
Online	N3	0.12854	1.00000	0.08946	0.2832	1.0293	-1.0910	0.8102	1.6652	100.49140	0.00048	0.56178
	N5	0.26065	1.00000	0.02940	0.2744	0.8408	0.3039	0.8102	1.6652	145.33553	0.00058	0.42416
	N7	0.23353	0.95819	-0.30456	-0.0116	0.9673	1.6987	0.8102	1.6652			
	N9	0.06790	1.00000	-0.32354	-0.2495	0.9782	2.4713	0.8102	1.6652	253.30670	0.00035	0.36937
	R3	0.08570	1.00000	-0.12757	-0.1172	1.1951	0.1399	1.1629	1.4867	134.01356	0.00059	0.43734
	R5	0.25824	1.00000	-0.02776	0.1707	0.9642	1.0718	1.1629	1.4867			
	R7	0.07500	1.00000	-0.09311	0.0485	0.9742	1.7694	1.1629	1.4867			
	R9	0.23665	1.00000	-0.29216	-0.0411	1.1291	2.3102	1.1629	1.4867	35.61401	-0.00015	1.03585
	S3	-0.02228	1.00000	0.50267	0.3575	0.9877	-1.6518	0.9406	2.6241			
	S5	0.07738	0.97071	0.60279	0.4816	1.0624	0.2715	0.9406	2.6241	114.13480	0.00018	0.69006
	S7	0.29852	1.00000	0.25374	0.5037	0.9068	1.5922	0.9406	2.6241			
	S9	-0.08522	1.00000	0.23621	0.2447	0.9209	2.8255	0.9406	2.6241			
	G3	-0.00411	1.15198	1.10486	0.0000	1.0000	-0.7518	1.2529	1.3605	95.63552	0.00041	0.58622
	G5	0.24575	1.00000	0.85955	0.0000	1.0000	0.2612	1.2529	1.3605	125.42115	0.00053	0.46176
	G7	0.02195	1.00000	0.66904	0.0000	1.0000	0.9034	1.2529	1.3605	121.30717	0.00043	0.55426
	G9	-0.02288	1.00000	0.61441	0.0000	1.0000	1.7331	1.2529	1.3605	-25.55819	-0.00012	1.10480
	W3	0.00000	1.00000	0.00000	0.0000	1.0000	0.0000	1.1160	3.3679			
	W5	0.00000	1.00000	0.00000	0.0000	1.0000	0.0000	1.1160	3.3679			
W7	0.00000	1.00000	0.00000	0.0000	1.0000	0.0000	1.1160	3.3679				
W9	0.00000	1.00000	0.00000	0.0000	1.0000	0.0000	1.1160	3.3679				
Paper	N3	0.04375	1.10615	-0.78200				0.8102	1.6652			
	N5	0.30951	0.92438	0.58969				0.8102	1.6652			
	N7	0.53365	1.00000	1.25462				0.8102	1.6652			
	N9	0.41191	0.81621	1.94652				0.8102	1.6652			
	R3	0.71441	1.12442	-0.37957				1.1629	1.4867			
	R5	0.76304	1.02541	0.69340				1.1629	1.4867			
	R7	0.38195	1.00000	1.55132				1.1629	1.4867			
	R9	0.76346	1.00000	1.62156				1.1629	1.4867			
	S3	-0.05370	1.04731	-0.79029				0.9406	2.6241			
	S5	0.17112	0.90960	0.86677				0.9406	2.6241			
	S7	0.29636	1.00000	2.18209				0.9406	2.6241			
	S9	-0.06482	1.00000	3.41619				0.9406	2.6241			
	G3	0.34426	1.22924	0.10715				1.2529	1.3605			
G5	0.35440	1.00000	1.08651				1.2529	1.3605				

Mode	Domain & year	LM ₂₁	SF ₂₁	Shift	LM ₁₉	SF ₁₉	HVR	MN08	SD08	a	b	c
	G7	0.31983	1.07673	1.46936				1.2529	1.3605			
	G9	0.62801	1.00000	1.78709				1.2529	1.3605			
	W3	0.00000	1.00000	0.00000				1.1160	3.3679			
	W5	0.00000	1.00000	0.00000				1.1160	3.3679			
	W7	0.00000	1.00000	0.00000				1.1160	3.3679			
	W9	0.00000	1.00000	0.00000				1.1160	3.3679			

Estimating equating errors

As with all statistics, equating shifts have an associated level of uncertainty. Had a different set of items been chosen for the equating test or had a different group of students been selected for the equating sample, 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 which 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 2021, equating errors were first estimated separately for the online test and for the paper tests. The final equating errors for comparing student achievement between 2021 and 2019 were then combined from the equating error of the online tests and the equating error of the paper tests according to the proportion of students that sat the online test (70%) and the proportion of students that sat the paper test (30%).

The equating errors were determined for comparing student achievement between 2021 and the base year or between 2021 and 2019. Multiple steps were involved in the equating of numeracy, reading, spelling, and grammar and punctuation. An equating error was estimated for each step by test mode. The equating errors were combined on the assumption that the errors from the steps are independent.

The errors considered in the equating processes over the course of the program are shown in Figure 94.

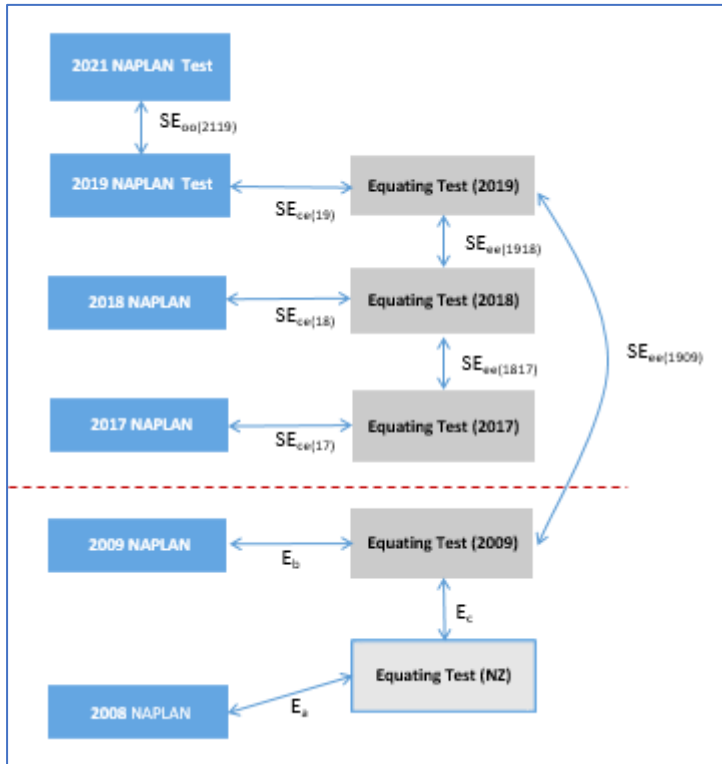


Figure 94. A schematic of the equating errors accumulated across NAPLAN administrations

For each domain and year level except writing:

- E_a is the standard error associated with equating the offshore equating test and the 2008 NAPLAN test;
- E_c is the standard error associated with equating the offshore and onshore equating tests; E_a , E_b and E_c were determined during 2009 equating process.
- $SE_{ce(xx)}$ is the standard error associated with equating the NAPLAN 20xx test with the equating test (calibration to equating), xx stands for 19 or 21;
- $SE_{ee(2119)}$ is the standard error associated with equating the 2021 and 2019 administrations of the equating test (equating to equating);
- $SE_{oo(2119)}$ is the standard error associated with equating the NAPLAN 2021 online test and the NAPLAN 2019 online test (equating to equating); and so forth.

For reporting results of NAPLAN 2021, the equating errors for equating the 2021 scale to the 2019 and 2008 scales were estimated by combining the relevant standard errors as follows by test mode:

Online test:

$$SE_{2021to2019}^{Online} = SE_{oo(2119)}$$

$$SE_{2021tobase}^{online} = \sqrt{(SE_{2019tobase})^2 + (SE_{2021to2019}^{online})^2}$$

Paper test:

$$SE_{2021tobase}^{Paper} = \sqrt{E_a^2 + E_c^2 + (SE_{ce(21)}^{paper})^2 + (SE_{ee(2109)}^{paper})^2}$$

Final equating error for 2021:

$$SE_{2021to2019} = \sqrt{(SE_{2021to2019}^{Online} * SF_{2021}^{online} * 0.7)^2 + (SE_{2021to2019}^{Paper} * SF_{2021}^{paper} * 0.3)^2}$$

$$SE_{2021tobase} = SE_{2021tobase}^{Online} * SF_{21}^{online}$$

The online equating error between 2021 and 2019 were estimated with taking the clustering of items in units into account. The following approach has been used to estimate the online equating error $SE_{2021to2019}^{oo}$ between 2021 and 2019. Suppose we have a total of L score points in the link items in K modules. Use i to index items in a unit and j to index units so that $\hat{\delta}_{ij}^y$ is the estimated difficulty of item i in unit j for year y , and let:

$$c_{ij} = \hat{\delta}_{ij}^{2021} - \hat{\delta}_{ij}^{2019}$$

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

$$\sum_{j=1}^K m_j = L \text{ and } \bar{m} = \frac{1}{K} \sum_{j=1}^K m_j$$

Further let:

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

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

$$SE_{2021to2019}^{oo} = \sqrt{\frac{\sum_{j=1}^k m_j^2 (c_{\cdot j} - \bar{c})^2}{K(K-1)\bar{m}^2}} = \sqrt{\frac{\sum_{j=1}^k m_j^2 (c_{\cdot j} - \bar{c})^2}{L^2} \frac{K}{K-1}}$$

Table 80 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 scaling factor, the 2008 standard deviation and 100.

Table 80. Standard errors of equating

Domain	Year	Logit		Scale score	
		2021 to base*	2021 to 2019	2021 to base*	2021 to 2019
Numeracy	3	0.0741	0.0307	4.4501	1.8428
	5	0.0780	0.0267	4.6862	1.6059
	7	0.0552	0.0217	3.3145	1.3026
	9	0.0590	0.0221	3.5447	1.3269
Reading	3	0.0705	0.0310	4.7429	2.0849
	5	0.0641	0.0265	4.3121	1.7791
	7	0.0609	0.0263	4.0993	1.7713

Domain	Year	Logit		Scale score	
		2021 to base*	2021 to 2019	2021 to base*	2021 to 2019
	9	0.0641	0.0237	4.3127	1.5958
Spelling	3	0.1030	0.0443	3.9267	1.6867
	5	0.1078	0.0419	4.1072	1.5951
	7	0.1114	0.0428	4.2465	1.6297
	9	0.1118	0.0495	4.2621	1.8879
Grammar and punctuation	3	0.1176	0.0504	8.6409	3.7024
	5	0.1123	0.0430	8.2579	3.1605
	7	0.1010	0.0385	7.4209	2.8274
	9	0.0927	0.0325	6.8143	2.3910
Writing**	3579	0.1510	0.1300	4.4835	3.8600

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

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

The equating errors were taken into account, together with sampling and measurement errors, in estimating the standard errors used to determine statistical significance in the comparisons between mean scores across years in NAPLAN reports. The equating errors are not included when estimating standard errors of estimates used to determine statistical significance in the comparisons between mean scores of different subgroups within NAPLAN 2021. This is further explained in Chapter 8.

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

Further details regarding the application of standard errors to testing the statistical significance of performance differences are given in Chapter 8.

Chapter 7: NAPLAN proficiency bands

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

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

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

In 2008, a response probability of 0.62 (RP62) was chosen, which needs to be combined with a band width of 52 NAPLAN scale scores to satisfy the condition that all students in a band are expected to respond correctly to at least 50 per cent of the items in the same band. It was decided to use the same cut scores between bands across all domains. Hence, the width of the bands in logits varies across domains. Table 81 shows the cut points between bands (lower bound) in scale scores and in logits.

Table 81. Lower bounds of proficiency bands in scale scores and in logits

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

Once the proficiency bands were defined, the skills that students in each band mastered were described by reviewing the items with an RP62 difficulty located within each band. The descriptions of the bands are included in Table 82 to Table 85 for each domain.

Table 82. Described scale for numeracy

Proficiency band	Numeracy skills and knowledge
Band 10	Uses mathematical understanding to solve complex problems including those involving irrational numbers. Interprets and uses index notation. Evaluates algebraic expressions and solves equations and inequalities using a range of algebraic strategies. Solves surface area and volume problems using geometric reasoning or formulas. Calculates and compares numerical probabilities. Applies knowledge of line and angle properties to spatial problems.
Band 9	Solves complex reasoning problems. Uses square roots and powers. Evaluates algebraic expressions and solves equations and inequalities using substitution. Interprets simple linear graphs. Interrogates data and finds measures of centre. Calculates elapsed time across time zones. Determines angle size, area and volume of polygons and diameter and circumference of circles. Recognises congruence and uses similarity in regular shapes.
Band 8	Solves non-routine problems and compares common fractions, decimals and percentages. Continues linear patterns and identifies non-linear rules. Solves perimeter and area problems. Determines probabilities of outcomes of experiments. Classifies triangles and uses their properties. Identifies transformations of shapes and visualises changes to 3D objects. Determines direction using compass points and angles of turn.
Band 7	Solves multi-step problems involving relational reasoning. Calculates missing values in equations. Interprets rules and patterns and completes simple inequalities. Finds perimeters and areas of composite shapes. Calculates elapsed times across midday and midnight. Expresses probability as a fraction. Compares and classifies angles and solves problems involving nets. Uses scale to determine distance on maps.
Band 6	Applies appropriate strategies to solve multi-step problems, simple multiplication and division and patterning. Converts between familiar units of measure. Calculates durations of events. Interprets and uses data from a variety of displays. Recognises nets of familiar 3D objects and symmetry in irregular shapes. Uses simple legends and coordinate systems to interpret maps and grids.
Band 5	Solves routine problems using a range of strategies. Demonstrates knowledge of simple fractions and decimals. Continues number and spatial patterns. Uses familiar measures to estimate, calculate and compare area or volume. Reads graduated scales. Compares likelihood of outcomes in chance events. Recognises the effect of transformations on 2D shapes. Uses major compass points and follows directions to locate positions.
Band 4	Solves problems involving unit fractions, combinations of addition and subtraction of two-digit numbers and number facts to 10×10 . Identifies repeating parts of patterns. Interprets timetables and calendars and reads time on clocks to the quarter hour. Locates information in tables and graphs. Recognises familiar 2D shapes after a transformation and identifies a line of symmetry. Visualises 3D objects from different viewpoints.
Band 3	Solves single-step problems involving addition, subtraction or simple multiplication. Recognises representations of unit fractions and completes simple number sentences. Compares length and mass using familiar units of measure. Describes outcomes of simple chance events. Uses common features and properties to classify families of shapes and objects, and recognises symmetrical grid references.
Band 2	Compares and orders different representations of three-digit numbers. Applies addition and subtraction facts up to 20 to solve problems. Identifies equal groups of collections. Uses language of time and chance in familiar contexts. Visually compares area and locates information in simple tables. Recognises common features of positions on simple maps and plans by following directions.
Band 1	Uses counting strategies to solve problems and demonstrates knowledge of place value of three-digit numbers. Identifies the next term in a simple pattern. Interprets tally marks. Recognises and compares length and mass of familiar objects. Names common 2D shapes and familiar 3D objects and shows some understanding of spatial positioning.

Table 83. Described scale for reading

Proficiency band	Reading skills and knowledge
Band 10	Analyses and critically evaluates aspects of complex texts to recognise an author's purpose and stance, and to identify an underlying message, subtle character traits, tone and point of view.
Band 9	Evaluates and processes implicit ideas in a range of complex narrative and informative texts and interprets complex vocabulary. Analyses and evaluates key evidence in persuasive texts. Identifies language and text features to infer an author's intended purpose and audience.
Band 8	Interprets ideas and processes information in a range of complex texts. Analyses how characters' traits and behaviours are used to develop stereotypes. Analyses and interprets persuasive texts to identify bias and to infer a specific purpose and audience. Interprets vocabulary, including technical words, specific to an informative text or topic.
Band 7	Applies knowledge and understanding of different text types and features to enhance meaning and infer themes and purpose. Identifies details that connect implied ideas across and within texts to process information and form conclusions. Interprets character motivation in narrative texts, the writer's values in persuasive texts and the main ideas in informative texts.
Band 6	Makes meaning from a range of text types of increasing difficulty and understands different text structures. Recognises the purpose of general text features such as titles and subheadings. Makes inferences by connecting ideas across different parts of texts. Draws conclusions about the feelings and motivations of characters, and sequences events and information.
Band 5	Applies knowledge, makes inferences and processes information to infer the main idea in texts. Draws conclusions about a character in narrative texts. Connects and sequences ideas in informative texts and identifies opinions in persuasive texts.
Band 4	Makes inferences from clearly stated information in short informative texts and stories. Identifies the meaning of some unfamiliar words from their context. Finds specific information in longer stories and informative texts including those with tables and diagrams.
Band 3	Makes meaning from simple texts with familiar content and themes and finds directly stated information. Makes some connections between ideas that are not clearly stated and identifies simple cause and effect. Makes some inferences and draws conclusions, such as identifying the main idea of a text.
Band 2	Makes some meaning from short texts, such as simple reports and stories, that have some visual support. Makes connections between pieces of clearly stated information.
Band 1	Makes some meaning from simple texts with familiar content. Texts have short sentences, common words and pictures to support the reader. Finds clearly stated information.

Table 84. Described scale for writing

Proficiency band	Writing skills and knowledge
Band 10	Writes a cohesive, engaging text that explores universal issues and influences the reader. Creates a complete, well-structured and well-sequenced text that effectively presents the writer's point of view. Effectively controls a variety of correct sentence structures. Uses punctuation correctly, including complex punctuation. Spells all words correctly, including many difficult and challenging words.
Band 9	Incorporates elaborated ideas that reflect a worldwide view of the topic. Makes consistently precise word choices that engage or persuade the reader and enhance the writer's point of view. Punctuates sentence beginnings and endings correctly and uses other complex punctuation correctly most of the time. Shows control and variety in paragraph construction to pace and direct the reader's attention.
Band 8	Writes a cohesive text that begins to engage or persuade the reader. Makes deliberate and appropriate word choices to create a rational or emotional response. Attempts to reveal attitudes and values and to develop a relationship with the reader. Constructs most complex sentences correctly. Spells most words, including many difficult words, correctly.
Band 7	Develops ideas through language choices and effective textual features. Joins and orders ideas using connecting words and maintains clear meaning throughout the text. Correctly spells most common words and some difficult words, including words with less common spelling patterns and silent letters.
Band 6	Organises a text using paragraphs with related ideas. Uses some effective text features and accurate words or groups of words when developing ideas. Punctuates nearly all sentences correctly with capitals, full stops, exclamation marks and question marks. Correctly uses more complex punctuation markers some of the time.
Band 5	Structures a text with a beginning, complication and resolution, or with an introduction, body and conclusion. Includes enough supporting detail for the text to be easily understood by the reader, although the conclusion or resolution may be weak or simple. Correctly structures most simple and compound sentences and some complex sentences.
Band 4	Writes a text in which characters or setting are briefly described, or in which ideas on topics are briefly elaborated. Correctly punctuates some sentences with both capital letters and full stops. May demonstrate correct use of capitals for names and some other punctuation. Correctly spells most common words.
Band 3	Attempts to write a text containing a few related events or ideas on topics, although these are usually not elaborated. Correctly orders the words in most simple sentences. May experiment with using compound and complex sentences but with little success. Orders and joins ideas using a few connecting words but the links are not always clear or correct.
Band 2	Shows audience awareness by using common text elements, for example, begins writing with Once upon a time; or I think ... because ... Uses some capital letters and full stops correctly. Correctly spells most simple words used in the writing. Some other one- and two-syllable words may also be correct.
Band 1	Writes a small amount of simple content that can be read. May name characters or a setting; or write a few content words on a topic. May write some simple sentences with correct word order but full stops and capital letters are usually missing or incorrect. Correctly spells a few simple words used in the writing.

Table 85. Described scale for conventions of language

Proficiency band	Conventions of language skills and knowledge
Band 10	Identifies errors and correctly spells difficult words and challenging words (<i>interrupt, camouflaged, instantaneous</i>). Demonstrates knowledge of the correct use of a wide range of grammar and punctuation conventions in complex texts.
Band 9	Identifies errors and correctly spells words with difficult spelling patterns (<i>rehearsals, deliberately, consistently</i>). Demonstrates knowledge of grammar and punctuation conventions in more complex texts, such as the correct use of possessive pronouns (<i>its</i>) and rhetorical questions.
Band 8	Identifies errors and correctly spells most words with difficult spelling patterns (<i>angrily, substantial, performance</i>). Demonstrates knowledge of grammar and punctuation conventions in more complex texts, such as the correct use of adverbs, pairs of conjunctions (<i>neither, nor</i>), cause and effect structures, quotation marks for effect and for speech and apostrophes for plural possession (<i>parents'</i>).
Band 7	Identifies errors and correctly spells words with common spelling patterns and some words with difficult spelling patterns (<i>applauded, received, achievement</i>). Demonstrates knowledge of grammar and punctuation conventions in more complex texts, such as appropriate and consistent sentence structure and the correct use of italics, apostrophes and commas to separate phrases.
Band 6	Identifies errors and correctly spells most words with common spelling patterns (<i>gloves, collect, hungry, comfortable</i>). Demonstrates knowledge of grammar and punctuation conventions in longer sentences and speech, such as the correct use of commas to separate phrases and apostrophes for contractions (<i>we'll</i>).
Band 5	Identifies errors and correctly spells one- and two-syllable words with common spelling patterns (<i>spill, locked, pleasing, benches</i>). Recognises grammar and punctuation conventions in standard sentences and speech, such as the correct use of adjectives, compound verbs (<i>could have</i>), capital letters for compound proper nouns and commas in lists.
Band 4	Identifies errors and correctly spells most one- and two-syllable words with common spelling patterns (<i>clear, mail, brick, won</i>). Recognises grammar and punctuation conventions in short sentences and speech, such as the correct use of groups of adjectives, referring pronouns (<i>those</i>) and capital letters for simple proper nouns.
Band 3	Identifies errors and correctly spells one-syllable words with simple spelling patterns (<i>out, feet, rain, hose, would</i>). Recognises grammar and punctuation conventions in short sentences, such as the correct use of linking and coordinating words (<i>that, but</i>), describing words, capital letters to begin a sentence, full stops and question marks.
Band 2	Identifies errors and correctly spells some words with simple spelling patterns. Recognises grammar and punctuation conventions in short sentences, such as the correct use of pronouns (<i>herself</i>).
Band 1	Identifies errors and correctly spells a few words with simple spelling patterns. Recognises a small range of grammar and punctuation conventions in short sentences, such as the correct use of simple conjunctions (<i>because</i>) and common verbs (<i>will go</i>).

Out of the 10 bands, only six bands were reported for each year level. Bands 1 to 6 were used for Year 3; bands 3 to 8, for Year 5; bands 4 to 9, for Year 7; and bands 5 to 10, for Year 9. Students in the two lowest band for each level were regarded as achieving below

the National Minimum Standard (NMS), students in the second lowest band were regarded as at the NMS.

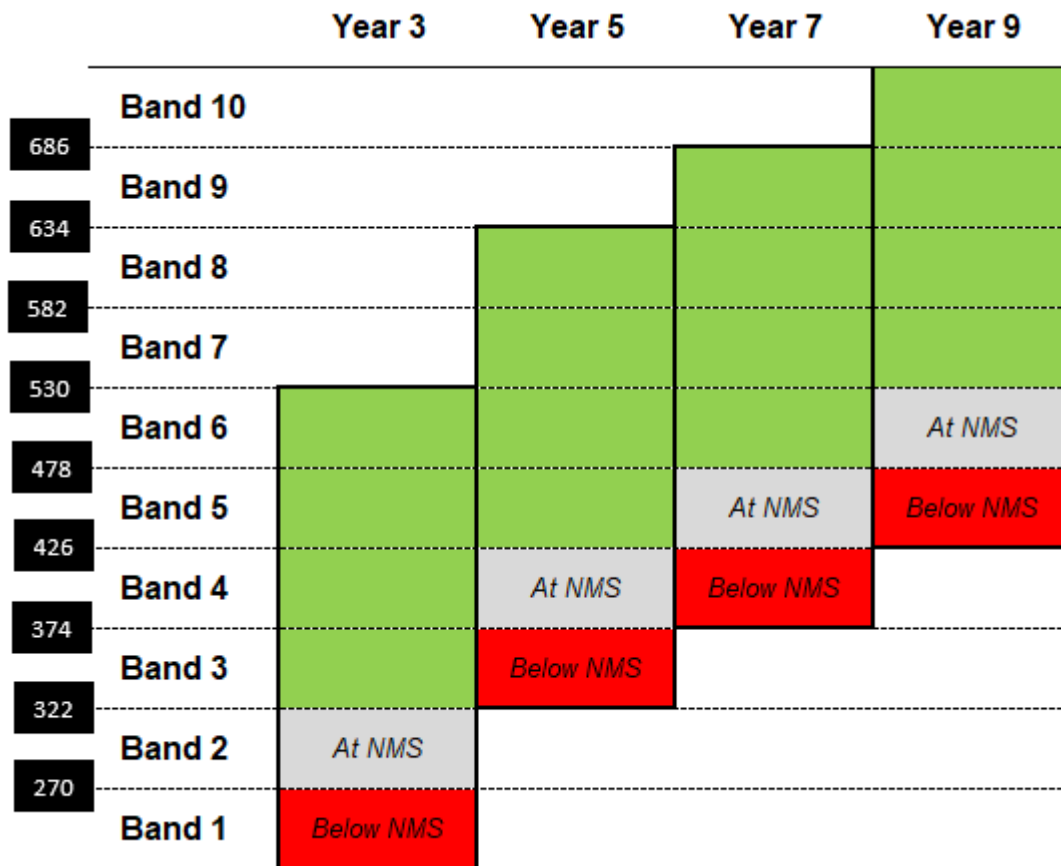


Figure 95. Schematic picture of proficiency bands by year levels

Illustrations

One Year 5 student received a NAPLAN score of 480 for numeracy. A score of 480 is near the lower bound of Band 6. This student is expected to respond correctly to 50 per cent of the items that have an RP62 difficulty between 478 and 530, and therefore, is regarded as mastering the skills that are described for Band 6 (see Table 82). This student is ready to be introduced to some of the skills and concepts described for Band 7.

Another Year 5 student received a NAPLAN score of 530 for numeracy. This student achieves at the very top of Band 6 and is expected to respond correctly to about 70 per cent of the items in this band. The student, therefore, has mastered most skills within Band 6 (see Table 82) and is ready to learn the skills and concepts described for Band 7.

Chapter 8: Reporting of national results

NAPLAN produces several reports for a variety of audiences each year. The student and school summary report (SSSR)¹¹ is a preliminary report with student and school level results for school staff. The individual student report (ISR)¹² is a report for parents about their child's NAPLAN achievement. The summary report is a national report with a selection of preliminary results. The national report replaces the summary report and includes final national statistics to inform policy makers and researchers. Additional reporting is also provided on the website *My School*¹³, with results for individual schools and accessible to the general public. This chapter describes analysis for the national report.

Calculation of statistics using plausible values

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

$$\theta = \frac{1}{5} \sum_{i=1}^5 \theta_i \quad (12)$$

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

Computation of standard errors

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

Three types of errors were estimated and different 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 two types of errors are explained in this chapter.

¹¹ www.nap.edu.au/docs/default-source/default-document-library/how-to-interpret-the-sssr.pdf?sfvrsn=10

¹² www.nap.edu.au/results-and-reports/student-reports

¹³ www.myschool.edu.au/

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 2021. In addition, even in census assessments, there is a certain amount of non-response that must be taken into account. Sampling error was considered at both the student and the school level. At the student level, there is a random element from one assessment year to another with respect to different age cohorts at each year level. At the school level, it needs to be considered that schools may be closed from one year to another or new schools may be opened.

The Taylor Series Linearization method (Wolter, 1995; 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 & Lemeshow, 1999).

Taylor series variance estimation can be done using commercially available statistical software. For NAPLAN 2021, the complex sample module implemented in the SPSS software package and the procedure *Proc Surveymeans* in the SAS software package were used in parallel processing for checking. Example of these procedures are included in Figure 96. 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.csplan' /PLANVARS ANALYSISWEIGHT=WGT /SRSESTIMATOR TYPE=WOR /PRINT PLAN /DESIGN CLUSTER=school_id /ESTIMATOR TYPE=WR.	<pre>proc surveymeans data=temp; cluster schID ; by grade <subgroups>; var PV1-PV5; ods output statistics=PVout; run;</pre>

Figure 96. 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 five statistics from the plausible values would be identical. Since no test is perfectly reliable, the five sets of statistics would not be identical. The measurement variance is estimated as:

$$B_M = \frac{1}{M-1} \sum_{i=1}^M (\theta_i - \theta)^2 \quad (13)$$

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

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

$$V = U + \left(1 + \frac{1}{M}\right) B_M \quad (14)$$

$$SE = \sqrt{V} \quad (15)$$

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 2021 performance in the reports.

Testing for differences

Two types of differences were computed and tested for significance. The first type of comparison was between subgroups within the NAPLAN 2021 data; for example, between male and female students or between jurisdictions. The second type of comparison was between 2021 results and results from earlier assessment years. Differences of the first type were tested for significance using the standard errors estimated from the sampling variance and the measurement variance. For testing the second type of differences, the equating errors needed to be taken into account as well.

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

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

$$SE_{DIFF} = \sqrt{SE_A^2 + SE_B^2} \quad (16)$$

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 two means by the standard error of the difference. The probability level of 0.05 was used for all statistical tests, with corresponding critical values of ± 1.96 . This illustrative example can be taken further by setting θ_A and θ_B to 500 and 515, respectively, and setting SE_A and SE_B to 3 and 4, respectively. Then, θ_B minus θ_A equals 15 and the standard error for this difference is equal to the square root of the

sum of 16 and 9, 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 2021 and 2019, we need to add the equating error between these two years, $SE_{2021to2019}$, to the calculation in the following way:

$$SE_{DIFF} = \sqrt{SE_{A19}^2 + SE_{A21}^2 + SE_{2012to2019}} \quad (17)$$

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 2021 NAPLAN with 2019 and the base year for each domain and year level are included in Chapter 6. No NAPLAN tests were administered in 2020 due to the pandemic, hence 2020 was skipped from reporting of the NAPLAN long term trend.

Only when differences between subgroups are compared between calendar years – for example, the gap between Indigenous and non-Indigenous students over time – the equating error does 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 2021 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 *Hedge's g*, whose formula is:

$$g = \frac{m_2 - m_1}{s_p} \quad (18)$$

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

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (19)$$

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. This formula is known to yield a biased estimate for the population value and is corrected using the following formula:

$$g_{unbiased} = g_{biased} \left[1 - \frac{3}{4(n_1 + n_2 - 2)} \right] \quad (20)$$

The effect size for differences in percentages is given by *Cox's d*, whose formula is:

$$OR = \frac{p_E q_C}{q_E p_C} \quad (21)$$

$$d_{Cox} = \frac{L(OR)}{1.65} \quad (22)$$

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

Three effect sizes were reported in the NAPLAN performance as follows:

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

Reporting of geographically classified statistics

Revisions to the Australian Statistical Geography Standard (ASGS) were undertaken by the Australian Bureau of Statistics in 2016 in an attempt to improve comparability in reporting geolocation structures and subgroups. This standard aims to provide a coherent set of comparable and geospatially integrated regions for implementation in the production and interpretation of geographically classified statistics.

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. & Lazendic, G. (2013). *Observations on the Feasibility of a Multistage Test Design for NAPLAN*. Unpublished technical report.
- Australian Assessment, Curriculum and Reporting Authority (ACARA). (2017). The Australian National Assessment Program Literacy and Numeracy (NAPLAN) assessment framework: NAPLAN Online 2017. ACARA: Sydney.
- Australian Assessment, Curriculum and Reporting Authority (ACARA). (2020). The Australian National Assessment Program Literacy and Numeracy (NAPLAN): 2019 Technical Report. ACARA: Sydney.
- Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs, I. The method of paired comparisons. *Biometrika*, 39, 324–45.
- Hendrickson, A. (2007). An NCME Instructional Module on Multistage Testing, *Educational Measurement: Issues and Practice*, 26, 2.
- Humphry, S. M. & McGrane, J. A. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *The Australian Educational Researcher*, 42, 443–60.
- Lord, F. M. and Novick, M. R. (1968) *Statistical Theories of Mental Test Scores*. Addison-Wesley: Menlo Park.
- Luce, R. D. (1959). *Individual Choice Behaviours: A theoretical analysis*. New York: J. Wiley.
- Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, 19(3), 189–202.
- Mislevy, R.J. & Sheehan, K.M. (1987), Marginal estimation procedures, in Beaton, A.E., Editor, 1987. *The NAEP 1983–84 technical report, National Assessment of Educational Progress*. Educational Testing Service, Princeton, pp. 293–360.
- National Assessment Program Literacy and Numeracy (NAPLAN) assessment framework:*
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmark Paedagpiske Institut.
- Rubin, D. (1991). EM and beyond. *Psychometrika*, 39, 111–21.
- Warm, T. A. (1989), Weighted Likelihood Estimation of Ability in Item Response Theory, *Psychometrika*, 54 (3), pp. 427–50.

Appendix A - Percentages and ability distribution by pathway

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

Appendix B - Item analysis details

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

Appendix C - Item summary tables

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

Appendix D - Item characteristic curves

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

Appendix E - Item-person maps

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

Appendix F - Gender DIF analysis

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

Appendix G - LBOTE DIF analysis

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

Appendix H - ATSI Status DIF analysis

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

Appendix I - DIF summary tables

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

Appendix J - Jurisdictional DIF

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

Appendix K - Horizontal link item comparison

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

Appendix L - Vertical link item comparisons

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

Appendix M - Exception report

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