



NAPLAN 2019

Technical Report

June 2020



National Assessment Program – Literacy and Numeracy (NAPLAN) 2019: technical report

Copyright

© Australian Curriculum, Assessment and Reporting Authority (ACARA) 2020, 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) 2020, 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
2020, *National Assessment Program – Literacy and Numeracy*
2019: Technical Report, ACARA, Sydney

Acknowledgements

ACARA worked with several contractors to successfully complete NAPLAN 2019. 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, Martin Murphy and Jorge Fallas from ACER; Stephen Humphry from UWA; and Eveline Gebhardt, Anna Cohen and Rassoul Sadeghi 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.....	20
Test administration	22
Participants.....	22
Marking	23
Psychometric analysis	23
Item selection for the 2019 NAPLAN tests.....	26
Chapter 3: NAPLAN test design.....	27
Paper test design.....	27
Online test design	29
Construction of NAPLAN Online tests	31
Test length.....	32
Difficulty of testlets.....	33
Item types for online tests	33
Curriculum coverage.....	34
Example items in reporting bands	41
Setting branching rules	54
Branching rules for NAPLAN Reading and Numeracy tests	55
Branching rules for spelling.....	58
Pathway utilisation	60
Chapter 4: Sampling	62
Sample frames	62
The calibration samples	62
Exclusions	62
Sample size.....	64
Stratification.....	65
Sample selection	66
Substitution.....	66
Weighting	67
Participation.....	68
Equating samples	70

Exclusions	70
Sample sizes	71
Stratification.....	73
Sample selection	74
Assignment to cognitive domains.....	74
Chapter 5: Data collection and preparation.....	76
Data collection and delivery	76
Data cleaning validation process.....	77
Data preparation.....	78
Distribution of not reached items	80
Not reached items in paper tests.....	80
Not reached items in online tests	81
Final student participation rates.....	84
Chapter 6: Scaling methodology and outcomes	85
Scaling model.....	85
Software used for analyses.....	85
Item calibration	86
Review of test and item characteristics.....	86
Test reliability.....	87
Test targeting and item spread.....	88
Item fit	92
Differential Item Functioning (DIF) Analyses.....	94
Estimation of student ability and generation of PVs	100
Chapter 7: Equating procedures.....	103
Equating of numeracy, reading, spelling, and grammar & punctuation results.....	103
Horizontal equating shifts.....	104
Vertical equating shifts.....	128
Horizontal-vertical regression (HVR) equating shifts	140
Scaling factors.....	144
Equating of writing results	146
Standardisation of scales from logits to reporting scales.....	151
Equipercntile equating.....	151
Summary of equating parameter estimates for NAPLAN 2019.....	153
Estimating equating errors	155
Chapter 8: NAPLAN proficiency bands	159
Illustrations	164
Chapter 9: Reporting of national results	166
Calculation of statistics using plausible values.....	166
Computation of standard errors.....	166
Sampling error.....	167
Measurement error	167
Testing for differences	168

Effect sizes.....	169
Reporting of geographically classified statistics	170
References	171
Appendix A: Percentages and ability distribution by pathway.....	172
Appendix B: Item analysis details.....	173
Appendix C: Item summary tables.....	174
Appendix D: Item characteristic curves.....	175
Appendix E: Item–person maps.....	176
Appendix F: Gender DIF	177
Appendix G: LBOTE DIF.....	178
Appendix H: ATSI DIF	179
Appendix I: DIF summary tables	180
Appendix J: Jurisdictional DIF for writing.....	181
Appendix K: Horizontal link item comparisons	182
Appendix L: Vertical link item comparisons.....	183
Appendix M: Measurement errors in individual achievement scores	184

List of tables

Table 1. Number of items developed for numeracy	17
Table 2: Composition of the trial numeracy item pool	21
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: Writing by task and total responses	22
Table 7: NAPLAN Numeracy paper test number of items and time available.....	27
Table 8: NAPLAN Reading paper test number of items and time available	28
Table 9: NAPLAN Language Conventions paper test number of items and time available	28
Table 10: NAPLAN Writing test criteria and score categories	29
Table 11 NAPLAN Online Numeracy test: number of items and time available	32
Table 12: NAPLAN Online Reading test: number of items and time available	32
Table 13: NAPLAN Online Conventions of Language test: number of items and time available	32
Table 14: NAPLAN Online numeracy and reading: predefined difficulty parameters for each testlet.....	33
Table 15: NAPLAN Online grammar and punctuation: predicted logit range and average for each testlet.....	33
Table 16: NAPLAN Online spelling: predicted logit range and average for each testlet..	33
Table 17: NAPLAN Online Numeracy: item types in the suite by year level.....	34
Table 18: NAPLAN Online Reading: item types in the suite by year level.....	34
Table 19: NAPLAN Online Conventions of Language: item types in the suite by year level.....	34
Table 20: NAPLAN Numeracy Year 3 curriculum coverage by mode and pathway.....	35
Table 21: NAPLAN Numeracy Year 5 curriculum coverage by mode and pathway.....	35
Table 22: NAPLAN Numeracy Year 7 curriculum coverage by mode and pathway.....	35
Table 23: NAPLAN Numeracy Year 9 curriculum coverage by mode and pathway.....	36
Table 24: NAPLAN Reading Year 3 curriculum coverage by mode and pathway	36
Table 25: NAPLAN Reading Year 5 curriculum coverage by mode and pathway	37
Table 26: NAPLAN Reading Year 7 curriculum coverage by mode and pathway	37
Table 27: NAPLAN Reading Year 9 curriculum coverage by mode and pathway	38
Table 28: NAPLAN Conventions of Language Year 3 curriculum coverage by mode and pathway	38
Table 29: NAPLAN Conventions of Language Year 5 curriculum coverage by mode and pathway	39
Table 30: NAPLAN Conventions of Language Year 7 curriculum coverage by mode and pathway	40
Table 31: NAPLAN Conventions of Language Year 9 curriculum coverage by mode and pathway	40
Table 32: Numeracy example items in reporting bands.....	41
Table 33: Reading example items in reporting bands.....	44
Table 34: Grammar and punctuation example items in reporting bands.....	49
Table 35: Spelling items in bands	52

Table 36: Stage 1 cut scores (Testlet A to C B D).....	57
Table 37: Stage 2 cut scores (testlet AB to C E F).....	57
Table 38: Stage 2 cut scores (testlet AD–C E F).....	58
Table 39: Stage 1, Testlet SA–SB SD cut scores.....	59
Table 40: Stage 2, Testlets SA–SB to PB PD cut scores.....	60
Table 41: Stage 2, Testlet SA–SD to PB PD cut scores.....	60
Table 42: Calibration sample exclusions.....	63
Table 43: Number and percentage of primary schools by assessment mode, nationally and for each jurisdiction.....	63
Table 44: Number and percentage of secondary schools by assessment mode, nationally and for each jurisdiction.....	64
Table 45: Numbers of schools and students in the calibration sample.....	64
Table 46: Quintile cut points for primary school NAPLAN performance by jurisdiction...	65
Table 47: Quintile cut points for secondary school NAPLAN performance by jurisdiction.....	65
Table 48: ABS remoteness classification.....	66
Table 49: Percentage of sampled schools (%) included in the calibration sample (2019).....	67
Table 50: Calibration sample: Distribution (%) by gender by year level (2019).....	68
Table 51: Calibration sample: Distribution (%) by language background by year level (2019).....	69
Table 52: Calibration sample: Distribution (%) by Indigenous status by year level (2019).....	69
Table 53: Calibration sample: Distribution (%) by geolocation by year level (2019).....	70
Table 54: Equating samples (paper and online) – exclusions.....	71
Table 55: Target school sample sizes, 2019 equating studies.....	71
Table 56: Achieved number of schools and students in the equating sample (reading) (2019).....	72
Table 57: Achieved number of schools and students in the equating sample (spelling) (2019).....	72
Table 58: Achieved number of schools and students in the equating sample (grammar and punctuation) (2019).....	72
Table 59: Achieved number of schools and students in the equating sample (numeracy) (2019).....	73
Table 60: Quintile cut points for the online primary school equating sample by jurisdiction.....	73
Table 61: Quintile cut points for the paper primary school equating sample by jurisdiction.....	74
Table 62: Quintile cut points for the online secondary school equating sample by jurisdiction.....	74
Table 63: Quintile cut points for the paper secondary school equating sample by jurisdiction.....	74
Table 64: School domain assignment, online equating 2019.....	75
Table 65: School Domain Assignment, paper equating 2019.....	75
Table 66: Rules for data recoding.....	79
Table 67: Pathway assignment rules to incomplete online tests.....	80
Table 68: Student participation rates by year level and domain, nationally and for each	

jurisdiction.....	84
Table 69: Reliability (WLE) for NAPLAN 2019 paper tests	88
Table 70. Summay of item statistics in NAPLAN 2019 paper tests	93
Table 71. Number of items showing gender DIF by domain by year level	95
Table 72. Numer of Items Showing LBOTE DIF by Domain by Year Level.....	96
Table 73. Numer of items showing Indigenous DIF by domain by year level	98
Table 74. Number of items showing state/territory DIF by domain by year level.....	99
Table 75: Equating design for both assessment modes	103
Table 76. Horizontal link review summary	127
Table 77. Horizontal equating shifts between 2019 item locations and 2009 item locations by test mode	127
Table 78. Vertical link review summary	139
Table 79. Vertical shift constants between adjacent year levels.....	140
Table 80. Vertical shift constants from each year level to Year 5.....	140
Table 81: Example of comparing horizontal shifts with vertical shifts (numeracy, online test)	141
Table 82. Regression intercepts and slopes	143
Table 83: Final HVR shifts applied for equating NAPLAN 2019 onto the NAPLAN historic scale.....	144
Table 84: Local means and scaling factors.....	145
Table 85: Domain mean and standard deviation for transforming logits to NAPLAN scale scores.....	151
Table 86: Equipercntile equating parameters.....	153
Table 87: Summary of parameters for transforming the 2019 logit scores to the NAPLAN reporting scales.....	154
Table 88. Standard errors of equating	157
Table 89: Lower bounds of proficiency bands in scale scores and in logits	159
Table 90: Described scale for numeracy	160
Table 91: Described scale for reading.....	161
Table 92: Described scale for writing	162
Table 93: Described scale for conventions of language	163

List of figures

Figure 1: A sample ICC for a poor performing item	24
Figure 2: A sample MC distractor curve for a poor performing item	24
Figure 3: A sample ICC for a good performing item	24
Figure 4: A sample MC distractor curve for a good performing item.....	25
Figure 5: A sample ICC displaying gender DIF in favor of girls	25
Figure 6: A sample ICC displaying gender DIF in favor of boys	26
Figure 7: The multistage tailored test design for numeracy and reading.....	30
Figure 8: Online test design for grammar and punctuation.	30
Figure 9: Multistage tailored test design for spelling.....	31
Figure 10: Test information functions curves for testlet C, B and E	56
Figure 11: Stage 1. Testlet A to C B D cut scores	56
Figure 12: Stage 2. Testlet AB to C E F cut scores	57
Figure 13: Stage 2. Testlet AD to C E F cut scores	58
Figure 14: Stage 1. Testlet SA to SB SD cut scores.....	59
Figure 15: Stage 2. Testlet SA-SB to PB PD cut scores	59
Figure 16: Stage 2. Testlet SA-SD to PB PD cut scores.....	60
Figure 17: Percentage of students assigned to each pathway in Year 3 numeracy.....	61
Figure 18: Ability distribution by pathway for Year 3 numeracy	61
Figure 19: Trailing missing percentage in numeracy, reading, grammar & punctuation, and spelling.....	81
Figure 20: Trailing missing percentage in numeracy	82
Figure 21: Trailing missing percentage in reading.....	82
Figure 22: Trailing missing percentage in grammar & punctuation.....	83
Figure 23: Trailing missing percentage in spelling.....	83
Figure 24: Wright map for Year 3 numeracy paper test (an example)	89
Figure 25: Wright map for paper writing test (a polytomous example).....	90
Figure 26: Thurstonian thresholds for writing test.....	91
Figure 27: Year 5 reading information function for test path A1B1E1	91
Figure 28: Item characteristic curves for an item with $\text{infit} = 1.01$	93
Figure 29: Item characteristic curves for an item with $\text{Infit} = 1.25$	94
Figure 30: Example of item characteristic curves displaying gender DIF [†]	96
Figure 31: Example of item characteristic curves displaying LBOTE DIF [†]	97
Figure 32: Example of item characteristic curves displaying Indigenous DIF [†]	98
Figure 33: Conditioning variables for the multidimensional item response model with latent regression model.....	102
Figure 34: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 3 paper students	105
Figure 35: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 5 paper students)	105
Figure 36: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 7 paper students	106
Figure 37: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 9 paper students	106
Figure 38 Scatterplot of spelling, horizontal equating items between 2019 and 2009 for	

Year 3 paper students	107
Figure 39: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 5 paper students	107
Figure 40: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 7 paper students	108
Figure 41: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 9 paper students	108
Figure 42: Scatterplot of grammar and punctuation, horizontal equating items between 2019 and 2009 for Year 3 paper students	109
Figure 43: Scatterplot of grammar and punctuation, horizontal equating items between 2019 and 2009 for Year 5 paper students	109
Figure 44: Scatterplot of grammar and punctuation, horizontal equating items between 2019 and 2009 for Year 7 paper students	110
Figure 45: Scatterplot of grammar and punctuation, horizontal equating items between 2019 and 2009 for Year 9 paper students	110
Figure 46: Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 3 paper students.....	111
Figure 47: Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 5 paper students.....	111
Figure 48: Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 7 paper students.....	112
Figure 49: Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 9 paper students.....	112
Figure 50: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 3 online students	113
Figure 51: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 5 online students	113
Figure 52: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 7 online students	114
Figure 53: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 9 online students	114
Figure 54: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 3 online students	115
Figure 55: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 5 online students	115
Figure 56: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 7 online students	116
Figure 57: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 9 online students	116
Figure 58: Scatterplot of GP testlet C, horizontal equating items between 2019 and 2009 for Year 3 online students	117
Figure 59: Scatterplot of GP testlet E1&E2, horizontal equating items between 2019 and 2009 for Year 3 online students.....	117
Figure 60. Scatterplot of GP testlet E3, horizontal equating items between 2019 and 2009 for Year 3 online students.....	118
Figure 61. Scatterplot of GP testlet F, horizontal equating items between 2019 and 2009 for Year 3 online students	118
Figure 62. Scatterplot of GP testlet C, horizontal equating items between 2019 and 2009 for Year 5 online students	119

Figure 63. Scatterplot of GP testlet E1&E2, horizontal equating items between 2019 and 2009 for Year 5 online students.....	119
Figure 64. Scatterplot of GP testlet E3, horizontal equating items between 2019 and 2009 for Year 5 online students.....	120
Figure 65 Scatterplot of GP testlet F, horizontal equating items between 2019 and 2009 for Year 5 online students	120
Figure 66. Scatterplot of GP testlet C, horizontal equating items between 2019 and 2009 for Year 7 online students	121
Figure 67. Scatterplot of GP testlet E1&E2, horizontal equating items between 2019 and 2009 for Year 7 online students.....	121
Figure 68. Scatterplot of GP testlet E3, horizontal equating items between 2019 and 2009 for Year 7 online students.....	122
Figure 69. Scatterplot of GP testlet F, horizontal equating items between 2019 and 2009 for Year 7 online students	122
Figure 70. Scatterplot of GP testlet C, horizontal equating items between 2019 and 2009 for Year 9 online students	123
Figure 71. Scatterplot of GP testlet E1&E2, horizontal equating items between 2019 and 2009 for Year 9 online students.....	123
Figure 72. Scatterplot of GP testlet E3, horizontal equating items between 2019 and 2009 for Year 9 online students.....	124
Figure 73. Scatterplot of GP testlet F, horizontal equating items between 2019 and 2009 for Year 9 online students	124
Figure 74. Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 3 online students	125
Figure 75. Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 5 online students	125
Figure 76. Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 7 online students	126
Figure 77. Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 9 online students	126
Figure 78. Scatterplot for vertical link item review for reading between Year 3 and Year 5 paper tests	129
Figure 79. Scatterplot for vertical link item review for reading between Year 5 and Year 7 paper tests	129
Figure 80. Scatterplot for vertical link item review for reading between Year 7 and Year 9 paper tests	130
Figure 81. Scatterplot for vertical link item review for spelling between Year 3 and Year 5 paper tests	130
Figure 82. Scatterplot for vertical link item review for spelling between Year 5 and Year 7 paper tests	131
Figure 83. Scatterplot for vertical link item review for spelling between Year 7 and Year 9 paper tests	131
Figure 84. Scatterplot for vertical link item review for grammar and punctuation between Year 3 and Year 5 paper tests	132
Figure 85. Scatterplot for vertical link item review for grammar and punctuation between Year 5 and Year 7 paper tests	132
Figure 86. Scatterplot for vertical link item review for grammar and punctuation between Year 7 and Year 9 paper tests	133
Figure 87. Scatterplot for vertical link item review for numeracy between Year 3 and Year	

5 paper tests	133
Figure 88. Scatterplot for vertical link item review for numeracy between Year 5 and Year 7 paper tests	134
Figure 89. Scatterplot for vertical link item review for numeracy between Year 7 and Year 9 paper tests	134
Figure 90. Scatterplot for vertical link item review for reading between Year 3 and Year 5 online tests.....	135
Figure 91. Scatterplot for vertical link item review for reading between Year 5 and Year 7 online tests.....	135
Figure 92. Scatterplot for vertical link item review for reading between Year 7 and Year 9 online tests.....	136
Figure 93. Scatterplot for vertical link item review for spelling between Year 3 and Year 5 online tests.....	136
Figure 94. Scatterplot for vertical link item review for spelling between Year 5 and Year 7 online tests.....	137
Figure 95. Scatterplot for vertical link item review for spelling between Year 7 and Year 9 online tests.....	137
Figure 96. Scatterplot for vertical link item review for numeracy between Year 3 and Year 5 online tests.....	138
Figure 97. Scatterplot for vertical link item review for numeracy between Year 5 and Year 7 online tests.....	138
Figure 98. Scatterplot for vertical link item review for numeracy between Year 7 and Year 9 online tests.....	139
Figure 99: Example HVR-shift (numeracy, online test).....	141
Figure 100. Comparisons of horizontal and vertical shifts of the paper tests.....	142
Figure 101. Comparisons of horizontal and vertical shifts of the online tests	143
Figure 102: Scatterplot for writing criteria between 2019 and 2016 paper and online tests	146
Figure 103: Scatterplot of the NAPLAN rubric and pairwise scale locations for 2016 and 2019 paper performances.	148
Figure 104: Scatterplot of the NAPLAN rubric and pairwise scale locations, comparing 2019 paper and online performances.	149
Figure 105: Scatterplot of the NAPLAN rubric and pairwise scale locations, for all 2016 and 2019 with online performances included.....	150
Figure 106: 2019 Online locations based on direct online–online (x-axis) and indirect online – paper (y-axis)	150
Figure 107: Scatterplot of percentiles in 2019 and 2017, Year 3 online numeracy	152
Figure 108. A schematic of the equating errors accumulated across NAPLAN administrations.	156
Figure 109: Schematic picture of proficiency bands by year levels	164
Figure 110: Examples in SPSS and SAS for estimating sampling variance.....	167

Chapter 1: Introduction

The first 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 the same year level 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 parents, schools, government and the non-government school sectors with important 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.

In May 2019, the National Assessment Program – Literacy and Numeracy (NAPLAN) 2019 tests were administered nationally to all students in Years 3, 5, 7 and 9. As in previous cycles of NAPLAN, students at each of these year levels 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 2019 administration.

The central analysis of NAPLAN data essentially involves a first step of 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. The reporting scale information with score-equivalence tables for the tests and proficiency bands provide necessary information for the jurisdictions to report to parents and schools.

Over one million students in Years 3, 5, 7 and 9 in all states and territories of Australia participated in NAPLAN 2019. From 2008 to 2017, NAPLAN delivered only paper-based tests. In 2018 and 2019, 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 2019, 52 per cent of students took the NAPLAN test online and 48 per cent of students took the test on paper. In 2018, the percentage were approximately 15 per cent and 85 per cent, respectively.

There were approximately 5,540 schools from all eight jurisdictions that participated in the online tailored tests.

Reporting of preliminary student performance and final national reporting combined results of online and paper participants. The delivery of the online tailored tests alongside the paper-

based NAPLAN 2019 presented new challenges in data analyses, including the equating of the online tests to the NAPLAN historic scales.

Five outcome reports were produced for NAPLAN 2019. 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 2019 National Report that was based on the second cut of the census data. This report for 2019 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 2019. Chapter 2 of this report describes the NAPLAN 2019 item trial. Chapter 3 describes the test design. Chapter 4 gives a summary of the methodology used in drawing samples for equating and calibration, and participation rates. Chapter 5 describes the data preparation process. Chapter 6 describes scaling methodology and outcomes. Chapter 7 describes the test equating processes to place the NAPLAN 2019 tests on the NAPLAN historic scales. Chapter 8 describes the proficiency bands on the NAPLAN scales. Chapter 9 describes the methodology used for reporting of NAPLAN 2019 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 NAPLAN 2019 item trial. There are two main components in the NAPLAN item trial: 1) item development and 2) psychometric analysis. The first part of this chapter describes the test development process, while the second part focuses on the psychometric analysis.

Item development

Item development required contractors to conform to the following documents:

- 2019 item development guidelines
- 2018 NAPLAN style guide
- ACARA accessibility guidelines
- ADS user guide
- Web Content Accessibility Guidelines (WGAG2.0 AA).

Items were delivered from both contractors in four year-level batches across the project period, from November until June. Items in each batch were reviewed by the National Testing Working Group (NTWG), feedback was synthesised by ACARA and the items were then returned to the contractor for revisions before the final ACARA checks and delivery in June 2019.

With each batch, compliance tables were submitted showing the spread of items across the curriculum, as well as item types and proficiencies across each year level. Source files of all graphics were also supplied.

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

Items that contained table shading were copied and then added as alternative items for students who required items in black and white, or fully shaded items (lilac, blue, yellow and green).

Audio was recorded for all items prior to trialling. This required scripting of all items, including alternative items, recording, editing, attaching audio to each item (and its accessibility alternative, where applicable) and checking of all recordings in each item.

Numeracy item development

Items for the NAPLAN 2019 Numeracy tests were procured from two separate contractors. The main contractor, the Australian Council for Educational Research (ACER), provided ACARA with items from both the Measurement and Geometry, and Statistics and Probability strands. Approximately 16 per cent additional accessibility substitute items were prepared by ACER for students with disability.

The second contractor, the National Foundation for Educational Research (NFER), provided items from the Number and Algebra strand.

Approximately 5 per cent additional accessibility substitute items were prepared by the NFER for students with disability.

The number of items developed for each strand are included in Table 1. Items were developed across the full range of item difficulties needed for the main study test design. The main study test is built from testlets of varying difficulty, it utilises a branched design (see Chapter 3).

Table 1. Number of items developed for numeracy

	Measurement and Geometry	Statistics and Probability	Number and Algebra
Year 3	30	15	52
Year 5	36	17	62
Year 7	42	19	70
Year 9	40	18	70
Total	148	69	254

Items were allocated to one of the four proficiencies to cover a range of cognitive demands – fluency, understanding, problem-solving and reasoning, with different percentage targets at each year level.

Items were supplied to cover three broad items types: 40 per cent multiple choice(s), 15 per cent text entry and 45 per cent technology-enhanced items.

Reading item development

ACARA contracted the University of New South Wales Global through the business group University of New South Wales Global Assessments (UNSWG) to produce a full suite of 24 reading testlets as part of the online NAPLAN Reading assessment for Years 3, 5, 7 and 9 for trial in 2018 and use in 2019. The UNSWG's final delivery included 71 stimulus texts and 478 items.

An additional package, procured from the NFER in May 2018, consisted of 30 items to supplement pre-existing reading units, most of which had been trialled but not used in a main study.

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 and targeted testlet, and source. This iterative matrix was submitted and revised throughout the item development cycle.

The difficulty of items, to a large extent, depended on the complexity of the stimulus texts. A common concern for NAPLAN reading item writing was the appropriate targeting of early childhood and, for all years, entry level texts. 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.

Entry level texts targeted students working at a skill level 1–3 years below their school year level but with subject matter that was still engaging and age appropriate for these students.

The UNSWG focused on producing very basic picture-book-like texts for the Year 3 easiest testlets. All easy testlets contained three or at the most four stimulus texts. Easy testlets for Year 3 contained four very short texts. It was expected that these short texts would allow even very low-ability readers to demonstrate the skills that they do have, providing useful information for their teachers.

Suitable paired texts were commissioned for use in the most difficult testlets to give greater scope for developing items, targeting very able readers functioning well above the literacy levels of their cohort. Special attention was paid to texts in testlets that would be linked during trialling to ensure that they were appropriate in content and style for each of the year levels in which they would appear.

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

- online readability, particularly in font selection and layout choices aimed at reducing the need for scrolling
- accessibility for visually impaired students, taking into account ACARA's guidelines on colour, contrast and font selection and the layouts shared with UNSWGA at an early feedback meeting
- resource file size. ACARA requested that the resource file size be kept at a maximum of 150 kb per text
- HTML texts. The requirement to provide texts in the HTML format created some challenges, but by the end of the project UNSWG exceeded the requirement of 25 per cent of texts designed in HTML, providing over half of the stimulus in this format.

Seventy-three stimuli made it through the development cycle to be accepted for item development. These stimuli were reviewed in three batches by panels of assessment and curriculum experts convened in each jurisdiction. Following the review and modification stages, 71 stimuli were accepted for item development.

Stage 2 of the cycle involved the development of over 500 items by the UNSW Global and 30 by the NFER.

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, item development managers and editors. For the UNSWG, an experienced Indigenous reviewer examined stimulus texts and was available for consultation throughout the project. ACARA also requested follow-up cultural reviews for some texts and these were provided. The NFER's stimuli had already been reviewed nationally and for cultural inclusion. A fact check was carried out for all information texts by a team member other than the text writer and then 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 also from 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 (CoL) tests consisted 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 were most likely to misspell as well as the likely error patterns. The words were put into 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 then placed in trial 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 the UNSW Global. The UNSWGA supplied [to ACARA] four separate batches of items, totalling approximately 351 grammar and 94 punctuation items for six testlets each for Years 3, 5, 7, 9. ACARA facilitated five reviews of the grammar and punctuation items over a six-month period. Feedback on accessibility alternative items was sought from the NTWG and ACARA's student diversity specialist. All modifications to items were made by ACARA.

Writing task development

Prompts for the 2019 NAPLAN Writing test were developed and trialled according to the following process:

1. Education experts from all over Australia developed a large pool of writing tasks to engage students in Years 3 and 5, and Years 7 and 9. Each jurisdiction (state or territory) created panels of experts with significant experience in writing assessment and educators that represented key groups that have special needs.
2. Expert panels undertook four stages of review of all of the 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. Once a shortlist of eight topics was chosen and refined, over 5,000 students responded to two different 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 show which tasks were the best in terms of fairness and measurement reliability. Psychometric analysis of the tasks ensured 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 was consulted and 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 that will be used to guide construction of the final NAPLAN tests and build each domain's item bank. Trialling also allowed other quantitative and qualitative feedback on the tests to be gathered, including time on task, engagement with test content and identification of online display issues. Single items and suites of test items (based on common stimuli) were authored in the test delivery system to be 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. Item locations obtained from trial data were also used to guide appropriate targeting in the construction of final tests.

The Educational Measurement Solution (EMS) was engaged to analyse items that were included in tests according to the trial design developed by ACARA for each of the test domains.

Item trial test design

Trial tests were designed so that a large number of items could be placed on a common scale. To achieve this, all testlets had to be linked, either directly or indirectly:

- Direct linkage: two testlets appear in the same pathway.
- Indirect linkage: two testlets appear in different pathways, but those pathways have other testlets in common.

In past paper trials, items were placed on the population scale by drawing a representative stratified sample of schools and students. Greater confidence in item location can be obtained by incorporating previously administered items, with known difficulty, into the trial tests in order to calibrate the new items. This has been a feature of online trials, while still drawing as strong a stratified sample as is possible within the constraints of technology transition. To achieve this linkage, two testlets of previously trialed items (equating testlets) were included in each test. These items approximated a NAPLAN test.

It is possible that items presented at the end of a test will perform differently from those presented at the beginning, due to accumulated cognitive load or time pressure. To counteract this potential position effect, the trial tests were designed so that testlets were presented at two or more positions within the tests. To illustrate, reading had the following rotational design for each year level:

- Sixteen testlets plus one testlet of stand-alone items¹
- Three nodes: node 1 had one testlet with approximately 10 stand-alone items. Students do just one item in this testlet. Nodes 2 and 3 had eight testlets each.
- Students started by doing a single stand-alone item then *either* did one testlet in node 1 then one testlet in node 2 *or* one testlet in node 2 then one testlet in node 1. This means every testlet was trialed in two different positions, i.e. every testlet was seen by half of the students first and by half of the students last.
- The equating testlets were placed in the trial design to approximate their position in the main study. Testlet A units were placed towards the start of the testlet, and testlet E units, towards the end of the testlet. The order of items was preserved.
- There were 128 parallel test pathways at each year level.

After assigning the newly developed and some historical items to the test designs and choosing items that were to be administered in two year levels, a total of 2,664 unique items were trialed, as well as a short survey to collect additional information: gender, device, device usage, where computer skills were learnt and whether students were used to typing stories or essays school.

The total item pool of unique items for numeracy was 626; for reading, 1,140; for grammar and punctuation, 413; and for spelling, 485.

¹ An item set consisting of a very short stimulus text and usually just one, occasionally two items. These are used to target specific reading skills and/or locations on the scale. They are also a bridge between the rigidity of an online test delivery system that needs constancy of total items and the flexibility of a test based on reading units.

Table 2: Composition of the trial numeracy item pool

	MC	CR	Total
Year 3	70	74	144
Year 5	79	89	168
Year 7	117	123	240
Year 9	127	113	240
Total	393	399	792

Table 3: Composition of the trial reading item pool

	MC	CR	Total
Year 3	332	96	428
Year 5	317	121	438
Year 7	327	150	477
Year 9	326	150	476
Total	1,302	517	1,819

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

	MC	CR	Total
Year 3	68	62	130
Year 5	60	70	130
Year 7	45	84	129
Year 9	37	93	130
Total	210	309	519

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

Eight writing tasks were trialled each at 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. That is, 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 each completed two tasks online. Approximately 250 Year 3 students completed one task online and one task on paper.

Table 6: Writing by task and total responses

Prompt	Year 3	Year 5	Year 7	Year 9	Total
Task 1	264	308	315	298	1,185
Task 2	352	375	311	277	1,315
Task 3	292	339	314	309	1,254
Task 4	277	307	364	347	1,295
Task 5	237	285	370	367	1,259
Task 6	304	333	327	318	1,282
Task 7	310	460	323	276	1,369
Task 8	348	389	320	308	1,365
Task 1 paper	137				137
Task 7 paper	110				110
Total	2,631	2,796	2,644	2,500	10,571

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 2019 NAPLAN program – reading, writing, language conventions (spelling, and grammar and punctuation), and numeracy. 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 various aspects of the trial administration. This feedback, in conjunction with a range of other sources of feedback, informed the selection and refinement of items for the final pool of assessment items and for the final student survey.

Participants

A convenience 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–CC field trial or main study
- distance education schools
- Montessori, Steiner, Waldorf schools
- special schools
- schools without NAPLAN performance data.

In total, 19,561 students from the trial schools across all states and territories participated in the trial. Each student completed two tests. The target number of responses for each item was set at 250 to achieve stable item parameters.

Marking

Development of marking materials and management of the marking operation were part of the trial administration contract awarded to Pearson for the NAPLAN 2019 item trial. A team of experienced NAPLAN markers was engaged by Pearson for marking the writing scripts. ACARA's writing test development manager supported Pearson's training of the markers, and they also remained on-site to oversee the marking process. On completion of marking each prompt, a debriefing session was held with the test developers, 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

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

The following steps have been taken to analyse NAPLAN 2019 trial data:

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. As a result, the first round of analyses treated '9' was treated as incorrect. Output files were sent to Test Development 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 second round of analyses (with acceptable items) was to calibrate items, therefore '9' was treated as missing rather than incorrect.

The Rasch measurement model (Rasch, 1960), using ACER Conquest (Adams, Wu, Cloney & Wilson; 2020) and RUMM software, was used in item calibration and analyses of trial data. 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 6.

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 curved 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.

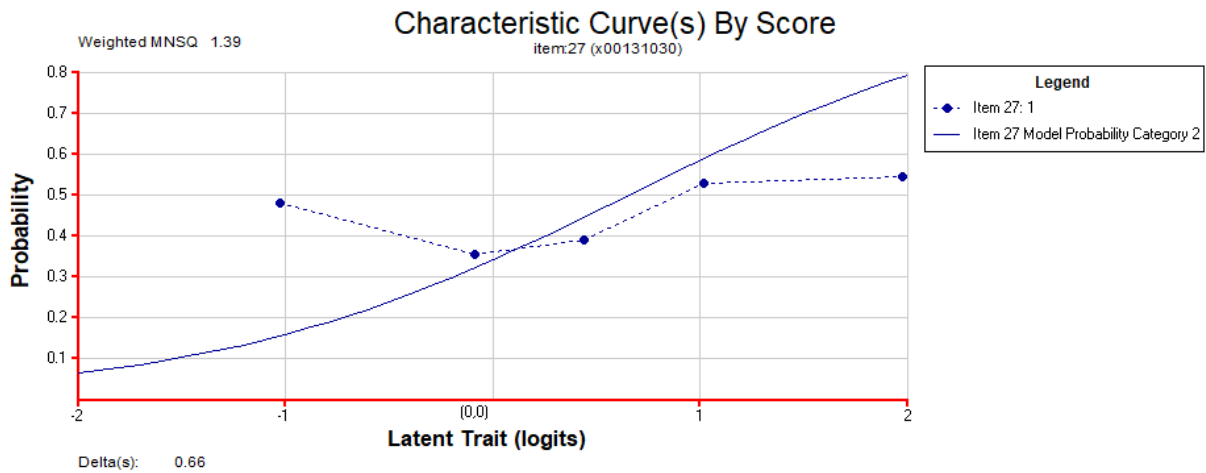


Figure 1: A sample ICC for a poorly performing item

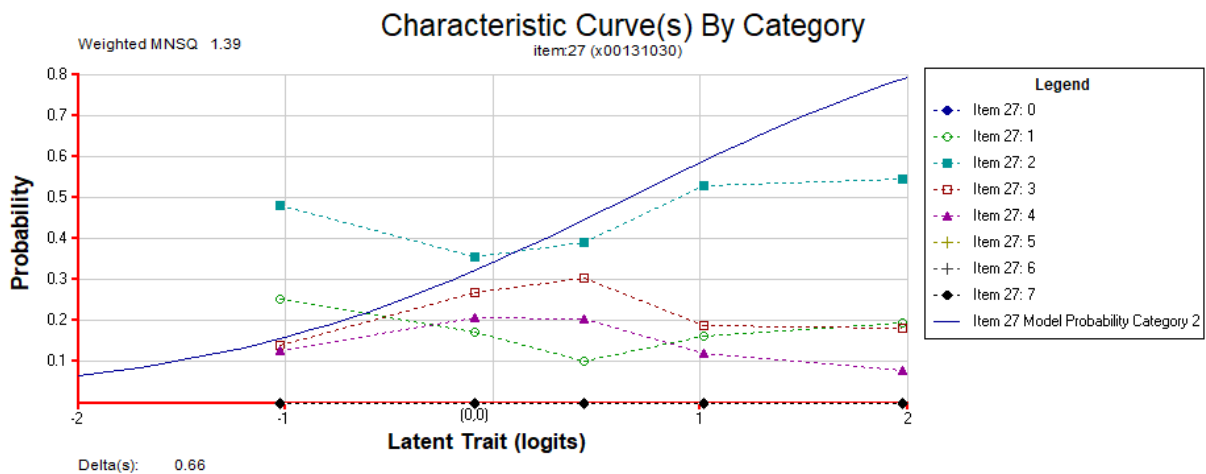


Figure 2: A sample MC distractor curve for a poorly performing item

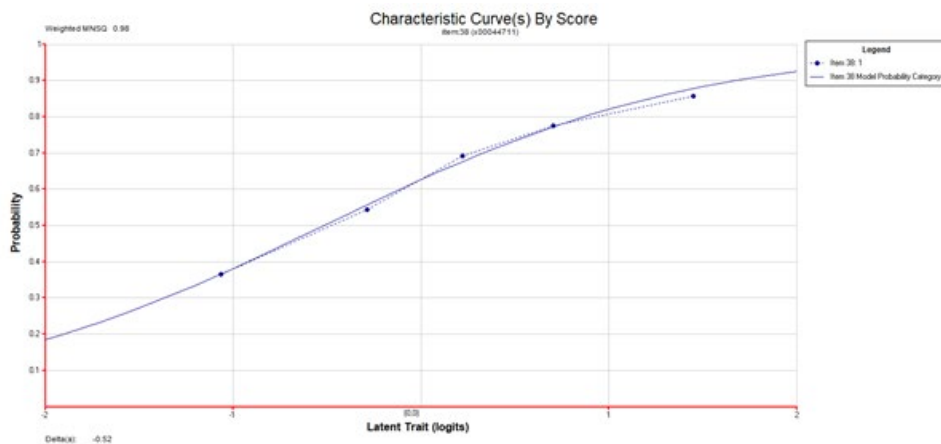


Figure 3: A sample ICC for a well-performing item

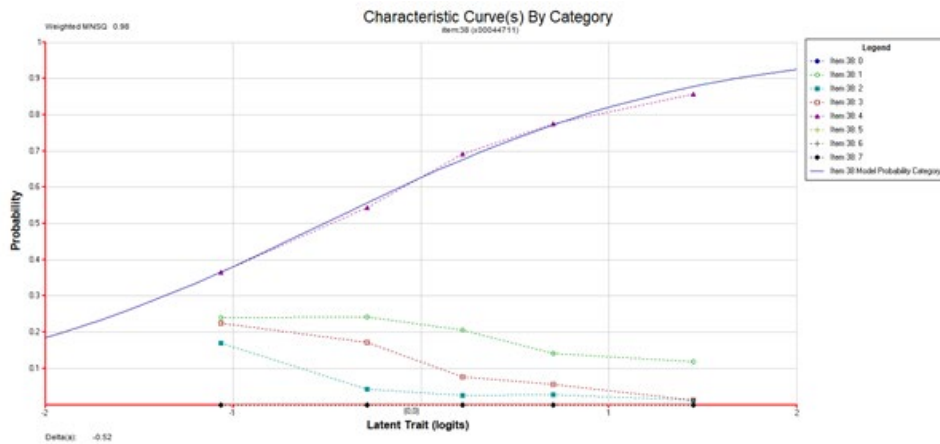


Figure 4: A sample MC distractor curve for a well- performing item

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.

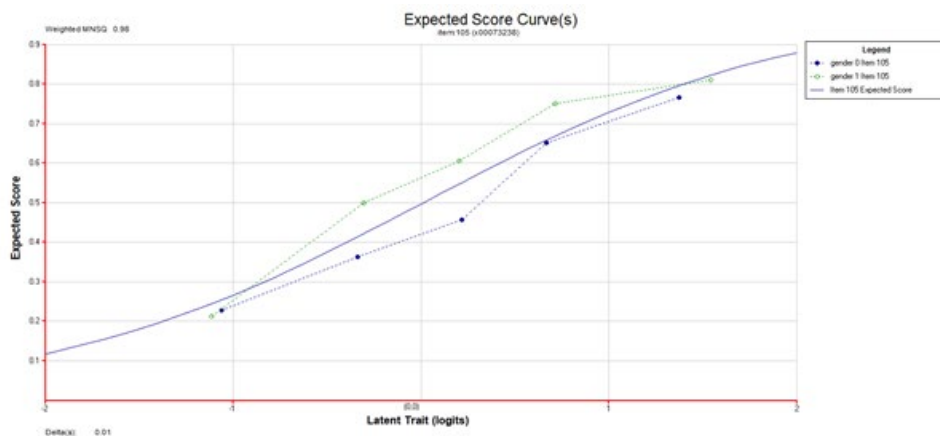


Figure 5: A sample ICC displaying gender DIF in favor of girls

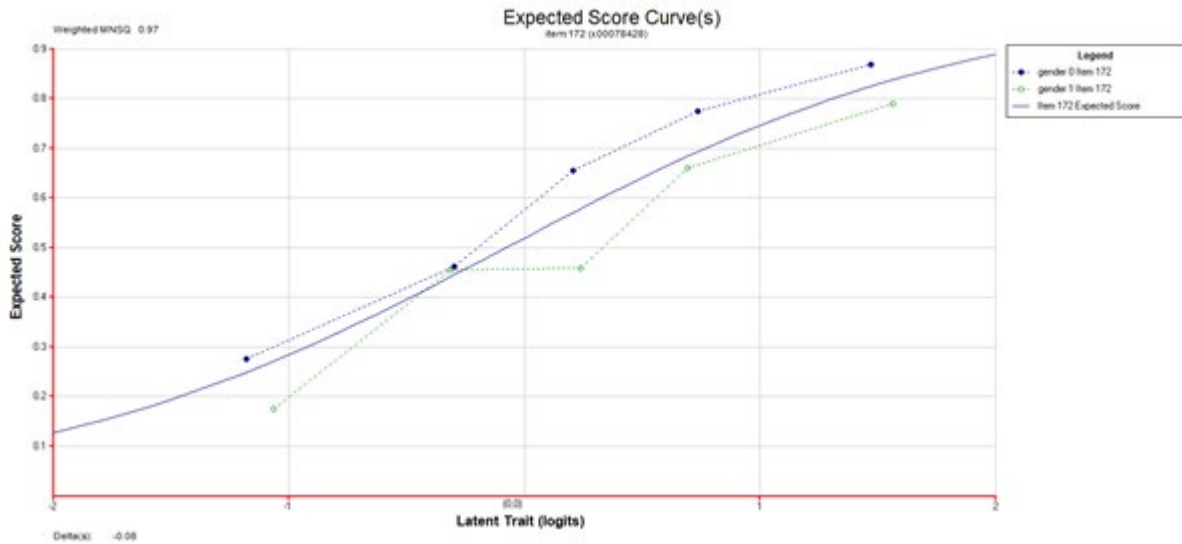


Figure 6: A sample ICC displaying gender DIF in favor of boys

Domain level analyses: Domain level analyses were IRT analyses where all acceptable items across the year levels for a domain were combined in one data file and scaled together. In all domain level analyses, '9' was treated as not administered (missing). Domain level analyses, using vertically linked items, enable placing all items from a domain on a common scale. A numeracy scale, for example, contains all numeracy items in Years 3, 5, 7 and 9 calibrated simultaneously.

Item selection for the 2019 NAPLAN tests

The results emerged from psychometric analysis provided a pool of psychometrically sound items to test managers to select items for inclusion in the final NAPLAN 2019 tests. Furthermore, results obtained from DIF analysis enabled test managers to exclude those items that displayed bias against a particular sub-group.

Chapter 3: NAPLAN test design

The aim of this chapter is to describe the NAPLAN 2019 test design. The first part of this chapter describes the test design for both paper and online tests. The branching method implemented in the NAPLAN multistage tailored test design is discussed in the second part.

Paper test design

Four paper-based linear 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. The test results were reported on five scales: numeracy, reading, spelling, grammar and punctuation, and writing. The paper-based tests were linear, or fixed, tests in that all students within a year level, who sat the tests on paper, completed the same 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 for 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 within each quartile of the scale. Items were ordered 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.

Year 3 and Year 5 numeracy tests consisted of 36 and 42 items, respectively. The use of calculators was not permitted in the numeracy tests in Years 3 and 5. Each of the Year 7 and Year 9 numeracy tests consisted of 48 items where in both year levels, the use of calculators was permitted in 40 of the items but not in eight of the items. The calculator items preceded the non-calculator items in the paper test.

Table 7 to Table 9 outline the total number of items in each test at each year level and the time available to students to complete the tests.

Table 7: 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 8: 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 9: 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 items from the online tests. Tables outlining other test specifications, encompassing average difficulty (logits), alignment to the Australian Curriculum and item types, are included in the next section about the online test design. Comparison of test specifications by online test pathways are also included in these tables so the paper and online test modes can be compared.

For the writing task, all students were required to write a narrative text in 2019. Students from Years 3 and 5 responded to one writing prompt, while students in Years 7 and 9 responded to a separate prompt. The scripts were rated based on the same 10 criteria (criteria 1–10) across four year levels. Each of these 10 criteria was rated polytomously. The ratings on the 10 criteria were treated as scores on 10 different items. The 10 criteria with the associated number of score categories are listed in Table 10.

Table 10: NAPLAN Writing test 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
Raw score range		0–47

Students' writing was marked by assessors who received intensive training in the application of this set of 10 writing criteria. Test administration authorities in each state and territory were responsible for the marking of the writing tests within their jurisdictions. All markers across Australia used the same marking rubric, received the same training and are subject to the same quality assurance measures.

Online test design

The NAPLAN Online numeracy, reading and spelling assessments incorporated a multistage tailored test design. A multistage tailored test is a type of Computerised Adaptive Test (CAT) but adaptivity takes place at testlet level as opposed to item level. Multistage tailored tests are considered a balanced compromise between linear paper-and-pencil and item-level adaptive tests (Hendrickson, 2007). Tailored testing allows students to demonstrate what they know and encourages students to stay engaged with the test.

The benefits of tailored testing include:

- Tailored tests provide a more precise measurement of student performance. This allows for greater differentiation of students by using a wider range of question 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 current paper-based format of NAPLAN too challenging.

The multistage tailored test design for numeracy and reading is illustrated in Figure 7. This figure shows a design with six testlets A, B, C, D, E and F, and students follow one of seven possible pathways (ABC, ABE, ABF, ADC, ADE, ADF and ACB) through the testlets. Each student completes three testlets. This multistage design will be discussed in more detail in the 'Setting branching rules' section.

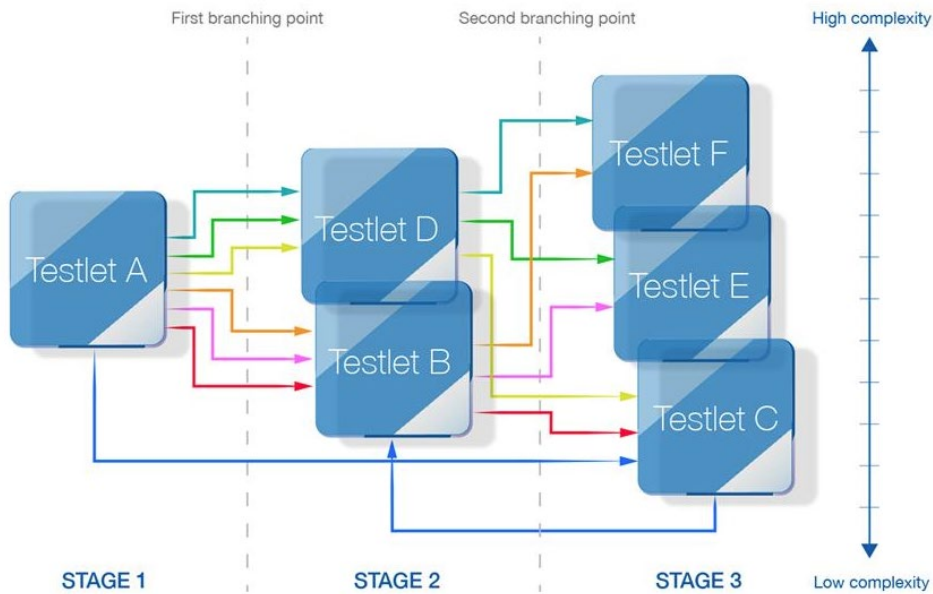


Figure 7: The multistage tailored test design for numeracy and reading

For NAPLAN 2019, the online reading and numeracy tests included three or four versions of each testlet, taking the total number of testlets for each year level to 18. The testlet versions were designed with comparable item difficulties, curriculum coverage and skills assessed. This resulted in 126 possible 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 first version of each testlet was populated with items from the paper tests and new online items. The other testlets were populated with horizontal links items from NAPLAN 2018 as well as items new and unique to the online tests. Vertical link items were included in testlets A, B, D and E.

The grammar and punctuation tailored test design consisted of three testlets: high complexity items (F), medium complexity items (E) and low complexity items (C). Students were directed to the appropriately difficult testlets based on the outcome of their reading tests. The graphical representation of the grammar and punctuation test designs is illustrated in figure 8.

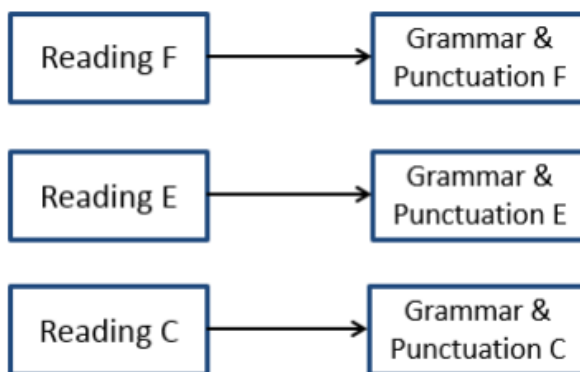


Figure 8: Online test design for grammar and punctuation

As Figure 8 shows, students who finished reading tests on the most challenging set of items, testlet F, were given testlet F in the grammar and punctuation tests; students who finished reading tests with testlet E received the grammar and punctuation testlet E. Finally, students who finished reading tests with testlet C or B received the grammar and punctuation testlet C. There were some common items between testlet C and E, and some common items between testlet E and F.

The spelling tests had a similar design to reading and numeracy tests but with only two testlets in the third stage. Figure 9 illustrates the multistage test design for spelling tests.

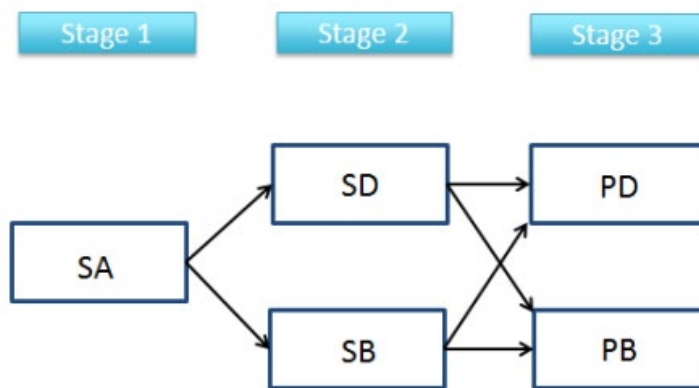


Figure 9: Multistage tailored test design for spelling

As in the reading and numeracy multistage tests, all students at each year level started with the same set of questions (testlet SA). Depending on a student's test performance in testlet SA, the second testlet included questions with overlapping content that was easier (SB) or more difficult (SD).

At the end of the second testlet, the student was directed to the third testlet, depending on their test performance in the first and second testlets. The final testlet also included overlapping content of varying difficulty: hard (PD) and easy (PB). As figure 9 shows, the first two stages of the test are focused on an audio component while the third stage is used to test a proofreading component. Spelling multistage design will be discussed in more detail in the 'Setting branching rules' section.

Construction of NAPLAN Online tests

The trial data largely determined the placement of items within testlets. Skills, curriculum strands and proficiencies were balanced across pathways. 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 MNSQ must stay between 0.9 and 1.1.
- Items should not display gender DIF at either year level.
- 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.
- The items need to be representative of the balance of strands in the tests.

Online tests were constructed around the link items with a balance across the strands and proficiencies. The items were selected to comply with the testlet logit ranges and averages.

Test length

Table 11 to Table 13 outline the test lengths for each domain. The grammar and punctuation and spelling sections of the conventions of language tests were not delineated by year level as there were no differences in the specifications for each.

Table 11 NAPLAN Online Numeracy test: number of items and time available

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

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

Table 12: 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 13: NAPLAN Online Conventions of Language test: number of items and time available

Conventions of language	Items per testlet	Item per section	Total CoL test items	Time available
Grammar and punctuation	25	25	50	45 minutes
Spelling	6 items x 1 testlet (audio dictation) 9 items x 1 testlet (audio dictation) 10 items x 1 testlet (proofreading)	25		

² NC – non-calculator

³ CA – calculator-allowed

Difficulty of testlets

NAPLAN assessments need to align to the breadth and depth of the Australian Curriculum and target the full range of students' abilities, so testlets need to vary in complexity and difficulty.

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

Table 14 to Table 16 outline the predefined difficulty ranges in logits and average difficulty for the testlets in each test.

Table 14: NAPLAN Online numeracy and reading: predefined difficulty parameters for each testlet

Numeracy & reading	Lower bound	Upper bound	Average
A	-3	1	-1
B	-2	0.5	-0.75
C	(low) -3.5	-0.5	-2
D	-0.5	2.0	0.75
E	-1.5	1.5	0
F	0.5	3.5 (high)	2

Table 15: NAPLAN Online grammar and punctuation: predicted logit range and average for each testlet

Grammar & punctuation	Lower bound	Upper bound	Average
C	-4	1	-1.5
E	-1	2	0.5
F	1	4	2.5

Table 16: NAPLAN Online spelling: predicted logit range and average for each testlet

Spelling	Lower bound	Upper bound	Average
SA	-2.5	0.5	-1
SB	-3.0	3.0	-0.75
SD	0.25	2.25	0.75
PB	-2.0	1.0	-0.5
PD	0.5	2.5	1

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. Table 17 to Table 19 show the final distribution of item types in the suite of items at each year level.

Table 17: NAPLAN Online Numeracy: item types in the suite by year level

Numeracy	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in pool
Year 3	103	28	47	178
Year 5	119	33	69	221
Year 7	133	26	69	228
Year 9	138	25	60	223

Table 18: NAPLAN Online Reading: item types in the suite by year level

Reading	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in pool
Year 3	170	-	38	208
Year 5	176	-	32	208
Year 7	222	-	34	256
Year 9	217	-	39	256

Table 19: NAPLAN Online Conventions of Language: item types in the suite by year level

Conventions of language	Number of MC/MCs items	Number of CR items	Number of technology-enhanced items	Total in pool
Spelling Year 3	0	118	0	118
Spelling Year 5	0	100	0	100
Spelling Year 7	0	100	0	100
Spelling Year 9	0	100	0	100
Gr & Pn Year 3	76	0	99	175
Gr & Pn Year 5	65	0	110	175
Gr & Pn Year 7	74	0	101	175
Gr & Pn Year 9	74	0	101	175

Curriculum coverage

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

Curriculum coverage is summarised in Table 20 to Table 31. 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 only is for 40 per cent Algebra, and 15 per cent Number at Year 9.

Table 20: 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%	55%	56%	58%	56%		56%	56%
<i>Measurement and Geometry</i>	30%	28%	29%	27%	29%		29%	28%
<i>Statistics and Probability</i>	15%	17%	16%	15%	15%		16%	16%
Proficiencies								
<i>Fluency</i>	20%	22%	31%	40%	36%		32%	27%
<i>Understanding</i>	30%	28%	32%	34%	33%		32%	30%
<i>Problem-solving</i>	30%	33%	21%	13%	18%		17%	22%
<i>Reasoning</i>	20%	17%	16%	14%	13%		19%	21%
Item types								
<i>MC/MCS</i>	60%	75%	59%	59%	56%		63%	63%
<i>Text entry</i>	15%	25%	15%	18%	16%		12%	14%
<i>Interactive</i>	25%	-	26%	23%	28%		25%	24%

Table 21: NAPLAN Numeracy Year 5 curriculum coverage by mode and pathway

Year 5	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	55%	55%	51%	52%	56%	57%
<i>Measurement and Geometry</i>	30%	26%	30%	32%	32%	29%	28%
<i>Statistics and Probability</i>	15%	19%	15%	17%	17%	15%	15%
Proficiencies							
<i>Fluency</i>	20%	19%	25%	30%	25%	22%	19%
<i>Understanding</i>	30%	29%	30%	33%	33%	33%	29%
<i>Problem-solving</i>	30%	31%	31%	23%	26%	30%	39%
<i>Reasoning</i>	20%	21%	14%	13%	17%	14%	13%
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 22: NAPLAN Numeracy Year 7 curriculum coverage by mode and pathway

Year 7	Specified	Paper	Online	ABC	ABE	ADE	ADF
Australian Curriculum strands							
<i>Number and Algebra</i>	55%	54%	53%	54%	54%	56%	55%
<i>Measurement and Geometry</i>	30%	29%	31%	29%	32%	29%	27%
<i>Statistics and Probability</i>	15%	17%	16%	16%	14%	15%	19%

Year 7	Specified	Paper	Online	ABC	ABE	ADE	ADF
Proficiencies							
<i>Fluency</i>	20%	17%	22%	29%	24%	25%	23%
<i>Understanding</i>	30%	33%	32%	31%	31%	31%	26%
<i>Problem-solving</i>	30%	27%	27%	24%	25%	25%	31%
<i>Reasoning</i>	20%	23%	19%	16%	19%	19%	20%
Item types							
<i>MC/MCS</i>	60%	71%	58%	57%	65%	63%	50%
<i>Text entry</i>	15%	29%	10%	8%	7%	8%	13%
<i>Interactive</i>	25%	-	32%	35%	28%	30%	37%

Table 23: 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%	58%	55%	56%	58%	59%	59%
<i>Measurement and Geometry</i>	30%	27%	29%	28%	26%	26%	27%
<i>Statistics and Probability</i>	15%	15%	16%	16%	16%	15%	15%
Proficiencies							
<i>Fluency</i>	20%	23%	22%	23%	24%	26%	22%
<i>Understanding</i>	30%	23%	23%	29%	24%	31%	20%
<i>Problem-solving</i>	30%	29%	38%	33%	35%	35%	41%
<i>Reasoning</i>	20%	25%	18%	15%	17%	17%	17%
Item types							
<i>MC/MCS</i>	60%	75%	63%	69%	68%	57%	53%
<i>Text entry</i>	15%	25%	12%	11%	11%	13%	13%
<i>Interactive</i>	25%	-	26%	21%	21%	31%	32%

Table 24: 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>	30-50%	43%	39%	46%	44%	36%	26%
<i>Literature</i>	30-50%	41%	47%	44%	57%	50%	55%
<i>Literacy</i>	10-20%	16%	13%	10%	9%	15%	19%
Cognitive processes							
<i>Locating & identifying</i>	30-50%	43%	39%	46%	44%	36%	26%
<i>Integrating & interpreting</i>	30-50%	41%	47%	44%	57%	50%	55%
<i>Analysing & evaluating</i>	10-20%	16%	13%	10%	9%	15%	19%

Year 3	Specified	Paper	Online	ABC	ABE	ADE	ADF
Stimulus texts							
<i>Number of texts</i>		6	-	7	6	6	7
<i>Average word count</i>		231	185	128	197	219	219
Item types							
<i>MC</i>	90-100%	89%	76%	77%	85%	79%	71%
<i>MCs</i>	0-10%	5%	5%	3%	4%	6%	8%
<i>Other</i>	0-10%	5%	18%	20%	11%	15%	21%

Table 25: 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%	15%	22%	20%	21%	25%	25%
<i>Literature</i>	5-15%	15%	12%	7%	10%	15%	13%
<i>Literacy</i>	70-90%	69%	66%	73%	68%	61%	62%
Cognitive processes							
<i>Locating & identifying</i>	30-50%	28%	31%	49%	38%	25%	22%
<i>Integrating & interpreting</i>	30-50%	46%	49%	40%	50%	54%	47%
<i>Analysing & evaluating</i>	10-20%	26%	21%	11%	13%	21%	31%
Stimulus texts							
<i>Number of texts</i>		6	-	6	6	6	7
<i>Average word count</i>		298	285	227	273	291	304
Item types							
<i>MC</i>	90-100%	82%	78%	86%	88%	83%	73%
<i>MCs</i>	0-10%	8%	6%	4%	4%	8%	12%
<i>Other</i>	0-10%	10%	15%	11%	8%	9%	15%

Table 26: 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%	14%	22%	22%	21%	21%	24%
<i>Literature</i>	10-20%	16%	12%	13%	17%	13%	12%
<i>Literacy</i>	50-70%	70%	66%	66%	63%	67%	64%
Cognitive processes							
<i>Locating & identifying</i>	20-40%	24%	28%	33%	28%	26%	25%
<i>Integrating & interpreting</i>	40-60%	48%	47%	53%	51%	49%	45%
<i>Analysing & evaluating</i>	20-40%	28%	25%	14%	20%	26%	30%

Year 7	Specified	Paper	Online	ABC	ABE	ADE	ADF
Stimulus texts							
<i>Number of texts</i>		8	-	9	9	9	9
<i>Average word count</i>		356	315	261	314	344	351
Item types							
<i>MC</i>	90-100%	84%	79%	91%	88%	79%	71%
<i>MCs</i>	0-10%	6%	8%	2%	4%	11%	13%
<i>Other</i>	0-10%	10%	13%	8%	8%	10%	16%

Table 27: 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%	24%	18%	20%	17%	17%	20%
<i>Literature</i>	10-20%	14%	16%	18%	15%	16%	16%
<i>Literacy</i>	50-70%	62%	66%	62%	68%	67%	64%
Cognitive processes							
<i>Locating & identifying</i>	20-40%	20%	20%	22%	22%	22%	21%
<i>Integrating & interpreting</i>	40-60%	46%	48%	58%	57%	44%	37%
<i>Analysing & evaluating</i>	20-40%	34%	33%	20%	22%	34%	42%
Stimulus texts							
<i>Number of texts</i>		8	-	9	9	9	9
<i>Average word count</i>		380	335	305	352	349	340
Item types							
<i>MC</i>	90-100%	86%	76%	87%	83%	75%	65%
<i>MCs</i>	0-10%	6%	9%	3%	6%	9%	13%
<i>Other</i>	0-10%	8%	15%	10%	10%	16%	23%

Table 28: NAPLAN Conventions of Language Year 3 curriculum coverage by mode and pathway

Year 3	Spec.	Paper	Online	G&P C	G&P E	G&P F	SASB PB	SASB PD	SASD PB	SASD PD
Australian Curriculum threads and test item formats										
<i>G&P grammar</i>	50%	72%	69%	76%	69%	62%	-	-	-	-
<i>G&P punctuation</i>	15%	28%	31%	24%	31%	38%	68%	68%	68%	68%
<i>Sp audio-dictation</i>	60%	0%	60%	-	-	-	18%	11%	18%	11%
<i>Sp mistake identified</i>	20%	48%	18%	-	-	-	15%	21%	15%	21%
<i>Sp mistake not identified</i>	20%	52%	22%	-	-	-	-	-	-	-
Australian Curriculum alignment to sub-domains										
<i>Editing</i>	-	-	2%	4%	3%	2%	-	-	-	-
<i>Punctuation</i>	-	14%	15%	24%	31%	38%	-	-	-	-

Year 3	Spec.	Paper	Online	G&P C	G&P E	G&P F	SASB PB	SASB PD	SASD PB	SASD PD
<i>Sentence-level grammar</i>	-	10%	13%	24%	32%	26%	-	-	-	-
<i>Text cohesion</i>	-	10%	6%	14%	13%	14%	-	-	-	-
<i>Vocabulary</i>	-	-	2%	8%	4%	-	-	-	-	-
<i>Word-level grammar</i>	-	14%	10%	26%	17%	20%	-	-	-	-
<i>Spelling</i>	-	52%	53%	-	-	-	100%	100%	100%	100%
Item types										
<i>MC/MCs</i>	-	50%	28%	46%	44%	40%	-	-	-	-
<i>Text entry</i>	-	50%	36%	-	-	-	100%	100%	100%	100%
<i>Interactive</i>	-	-	36%	54%	56%	60%	-	-	-	-

Table 29: NAPLAN Conventions of Language Year 5 curriculum coverage by mode and pathway

Year 5	Spec.	Paper	Online	G&P C	G&P E	G&P F	SASB PB	SASB PD	SASD PB	SASD PD
Australian Curriculum threads and test item formats										
<i>G&P grammar</i>	-	68%	71%	76%	67%	74%	-	-	-	-
<i>G&P punctuation</i>	-	32%	29%	24%	33%	26%	-	-	-	-
<i>Sp audio-dictation</i>	-	-	60%	-	-	-	-	68%	71%	-
<i>Sp mistake identified</i>	-	48%	13%	-	-	-	-	32%	29%	-
<i>Sp mistake not identified</i>	-	52%	27%	-	-	-	-	-	60%	-
Australian Curriculum alignment to sub-domains										
<i>Editing</i>	-	-	1%	-	-	-	-	4%	0%	0%
<i>Punctuation</i>	-	16%	14%	-	-	-	-	24%	33%	26%
<i>Sentence-level grammar</i>	-	14%	13%	-	-	-	-	18%	32%	36%
<i>Text cohesion</i>	-	8%	7%	-	-	-	-	16%	19%	10%
<i>Vocabulary</i>	-	-	2%	-	-	-	-	6%	3%	6%
<i>Word-level grammar</i>	-	12%	10%	-	-	-	-	32%	13%	22%
<i>Spelling</i>	-	50%	55%	100%	100%	100%	100%	-	-	-
Item types										
<i>MC/MCs</i>	-	50%	24%	-	-	-	-	34%	48%	24%
<i>Text entry</i>	-	50%	36%	100%	100%	100%	100%	66%	52%	76%
<i>Interactive</i>	-	-	40%	-	-	-	-	-	-	-

Table 30: NAPLAN Conventions of Language Year 7 curriculum coverage by mode and pathway

Year 7	Spec.	Paper	Online	G&P C	G&P E	G&P F	SASB PB	SASB PD	SASD PB	SASD PD
Australian Curriculum threads and test item formats										
<i>G&P grammar</i>	-	68%	59%	82%	76%	72%	-	-	-	-
<i>G&P punctuation</i>	-	32%	23%	18%	24%	28%	-	-	-	-
<i>Sp audio-dictation</i>	-	-	63%	-	-	-	68%	68%	68%	68%
<i>Sp mistake identified</i>	-	48%	16%	-	-	-	15%	15%	15%	15%
<i>Sp mistake not identified</i>	-	52%	22%	-	-	-	18%	18%	18%	18%
Australian Curriculum alignment to sub-domains										
<i>Editing</i>	-	-	3%	-	7%	6%	-	-	-	-
<i>Punctuation</i>	-	18%	13%	18%	24%	28%	-	-	-	-
<i>Sentence-level grammar</i>	-	16%	14%	26%	33%	30%	-	-	-	-
<i>Text cohesion</i>	-	2%	5%	22%	9%	4%	-	-	-	-
<i>Vocabulary</i>	-	-	1%	4%	4%	-	-	-	-	-
<i>Word-level grammar</i>	-	14%	12%	30%	23%	32%	-	-	-	-
<i>Spelling</i>	-	50%	52%	-	-	-	100%	100%	100%	100%
Item types										
<i>MC/MCs</i>	-	25%	28%	38%	55%	32%	-	-	-	-
<i>Text entry</i>	-	25%	36%	-	-	-	100%	100%	100%	100%
<i>Interactive</i>	-	50%	36%	62%	45%	68%	-	-	-	-


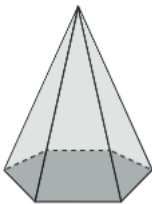
Table 31: NAPLAN Conventions of Language Year 9 curriculum coverage by mode and pathway

Year 9	Spec.	Paper	Online	G&P C	G&P E	G&P F	SASB PB	SASB PD	SASD PB	SASD PD
Australian Curriculum threads and test item formats										
<i>G&P grammar</i>	-	68%	71%	68%	69%	76%	-	-	-	-
<i>G&P punctuation</i>	-	32%	29%	32%	31%	24%	-	-	-	-
<i>Sp audio-dictation</i>	-	-	39%	-	-	-	40%	40%	40%	40%
<i>Sp mistake identified</i>	-	48%	50%	-	-	-	50%	52%	50%	52%
<i>Sp mistake not identified</i>	-	52%	11%	-	-	-	10%	8%	10%	8%
Australian Curriculum alignment to subdomains										
<i>Editing</i>	-	4%	4%	8%	8%	4%	-	-	-	-
<i>Punctuation</i>	-	18%	15%	32%	31%	24%	-	-	-	-
<i>Sentence-level grammar</i>	-	8%	12%	16%	23%	32%	-	-	-	-
<i>Text cohesion</i>	-	4%	6%	16%	7%	12%	-	-	-	-
<i>Vocabulary</i>	-	4%	3%	4%	8%	4%	-	-	-	-
<i>Word-level grammar</i>	-	12%	13%	24%	24%	24%	-	-	-	-
<i>Spelling</i>	-	50%	48%				100%	100%	100%	100%

Year 9	Spec.	Paper	Online	G&P C	G&P E	G&P F	SASB PB	SASB PD	SASD PB	SASD PD
Item types										
MC/MCs	-	25%	23%	22%	8%	16%	-	-	-	-
Text entry	-	25%	48%	-	-	-	100%	100%	100%	100%
Interactive	-	50%	29%	78%	92%	84%	-	-	-	-

Example items in reporting bands

Table 32: 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"/> $30\,000 + 200$ <input type="radio"/> $30\,000 + 100$ </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. The graph below shows the results.</p> <table border="1"> <caption>Data from the stacked bar chart</caption> <thead> <tr> <th>Age (years)</th> <th>No (%)</th> <th>Sometimes (%)</th> <th>Yes (%)</th> </tr> </thead> <tbody> <tr> <td>12</td> <td>50</td> <td>14</td> <td>36</td> </tr> <tr> <td>13</td> <td>45</td> <td>17</td> <td>38</td> </tr> <tr> <td>14</td> <td>52</td> <td>10</td> <td>38</td> </tr> <tr> <td>15</td> <td>30</td> <td>24</td> <td>46</td> </tr> <tr> <td>16</td> <td>30</td> <td>15</td> <td>55</td> </tr> <tr> <td>17</td> <td>24</td> <td>12</td> <td>64</td> </tr> <tr> <td>18</td> <td>17</td> <td>18</td> <td>65</td> </tr> </tbody> </table> <p>There were 150 12-year-old students at the high school. How many 12-year-old students responded 'No'?</p> <p>21 50 54 75 100</p> <p><input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/></p>	Age (years)	No (%)	Sometimes (%)	Yes (%)	12	50	14	36	13	45	17	38	14	52	10	38	15	30	24	46	16	30	15	55	17	24	12	64	18	17	18	65	D
Age (years)	No (%)	Sometimes (%)	Yes (%)																																
12	50	14	36																																
13	45	17	38																																
14	52	10	38																																
15	30	24	46																																
16	30	15	55																																
17	24	12	64																																
18	17	18	65																																
9	686	<p>33 At the entrance to a harbour there are two lights. A red light flashes every 5 seconds. A green light flashes every 7 seconds. The red light and the green light both flash together at 7:00 am. How many more times will the lights both flash at the same time in the next 3 minutes?</p> <p><input type="text"/></p>	5																																
10	738	<p>38 Suki makes a regular hexagon from six identical triangular tiles. Each tile has an area of 3.9 cm^2.</p> <p>Suki then adds more tiles to make a hexagon with double the side length of this hexagon. What will be the area of this larger hexagon?</p> <p>7.8 cm^2 23.4 cm^2 46.8 cm^2 93.6 cm^2</p> <p><input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/></p>	D																																

Table 33: 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 x00073144	<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 x00074156	<p>Why didn't Abbey and Michael <i>see what was coming?</i> (second last paragraph)</p> <ul style="list-style-type: none"><input type="radio"/> The sun was shining brightly in their eyes.<input type="radio"/> Dingle's head was blocking their view.<input checked="" type="radio"/> They were not looking where they were going.<input type="radio"/> Dingle made them dizzy.

A great southern secret—*two views*

View 1

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

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

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

View 2

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

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

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



Band	NAPLAN scale score	Stimulus text	Item
7	545	A great southern secret – two views x00074162	<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 x00074170	<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 x00074168	<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	<p>In View 1, what is the main point of contrast between the river and the sea?</p> <p><input checked="" type="radio"/> sound</p> <p><input type="radio"/> depth</p> <p><input type="radio"/> colour</p> <p><input type="radio"/> beauty</p>


Table 34: Grammar and punctuation example items in reporting bands

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

Band	NAPLAN scale score	Item	Key / key string
4	420	<p>Place the correct word in the box to complete this sentence.</p> <div style="border: 1px solid #ccc; padding: 5px; margin: 10px 0;"> hard hardly hardest </div> <p>It is harder to ride a horse than a bike.</p>	
5	458	<p>Place the correct word in the box to complete this sentence.</p> <div style="border: 1px solid #ccc; padding: 5px; margin: 10px 0;"> each much every </div> <p>The teacher asked how many parents would come to the concert.</p>	
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. 	
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. 	

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="481 427 1305 728"> <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="488 846 1315 920" style="border: 1px solid #ccc; padding: 5px; text-align: center;"> : ; </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>																										

Table 35: 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 style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">free</div>	
2	325.7	<p>The spelling mistake in this sentence is underlined.</p> <p>The toy began to <u>spinn</u> around.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">spin</div>	
3	362.6	<p>The spelling mistake in this sentence is underlined.</p> <p>He <u>kickd</u> the football through the goals.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">kicked</div>	

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 <u>bigga</u> than a mouse.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div data-bbox="488 591 1082 698" style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;"> <p>bigger</p> </div>	
5	430.0	<p>The spelling mistake in this sentence is underlined.</p> <p>One <u>rool</u> in our class is to raise your hand to ask for help.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div data-bbox="488 990 1008 1088" style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;"> <p>rule</p> </div>	
6	516.6	<p>The spelling mistake in this sentence is underlined.</p> <p>The children saved the day and were <u>heros</u>.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div data-bbox="488 1411 1088 1518" style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;"> <p>heroes</p> </div>	
7	534.3	<p>The spelling mistake in this sentence is underlined.</p> <p>A rock band often has a <u>gitar</u> player.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div data-bbox="488 1848 1098 1955" style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;"> <p>guitar</p> </div>	

Band	NAPLAN scale score	Item	Key / key string
8	611.2	<p>There is one spelling mistake in this sentence.</p> <p>The students had a very efficient method for completing their homework.</p> <p>Type the correct spelling of the word in the box.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">efficient</div>	
9	654.6	<p>There is one spelling mistake in this sentence.</p> <p>The performance was given spontaneous applause.</p> <p>Type the correct spelling of the word in the box.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">spontaneous</div>	
10	716.3	<p>The spelling mistake in this sentence is underlined.</p> <p>The mouse was a <u>nusence</u> when it chewed through the electricity cord.</p> <p>Type the correct spelling of the underlined word in the box.</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;">nuisance</div>	

Setting branching rules

Students are branched to more or less difficult 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 point 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 rout 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 reading and numeracy 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 NAPLAN Reading and Numeracy tests

Figure 7 illustrates a three-stage tailored test design (1–2–3) with one testlet in Stage 1; two testlets in Stage 2; and three testlets in Stage 3. These six testlets form seven pathways (ABC, ABE, ABF, ADC, ADE, ADF and ACB), which are shown using different colouring arrows.

All students at each year level and domain started with testlet A (Stage 1). Once testlet A was completed, a decision was made to branch a student to either an easier testlet B or a harder testlet D, which was the *first branching point*. Assuming that a student was sent to testlet D and completed this testlet, then another decision was made to branch this student to testlet C (low complexity items), testlet E (items with average complexity) or testlet F (high complexity items), which was the *second branching point*. If a student was branched to testlet E, pathway ADE (highlighted in green in Figure 7) was completed. As discussed earlier, students with very low performance on testlet A were first assigned the easiest testlet C as a second testlet before finally being assigned testlet B as the third testlet (pathway ACB, highlighted in blue). 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 C, B and D are illustrated in figure 10. As this figure shows, the peak of the information function for testlets B and D was about -1 and 1 logits, respectively. This means that the items were allocated to testlets B and D so that D was more suited to 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.

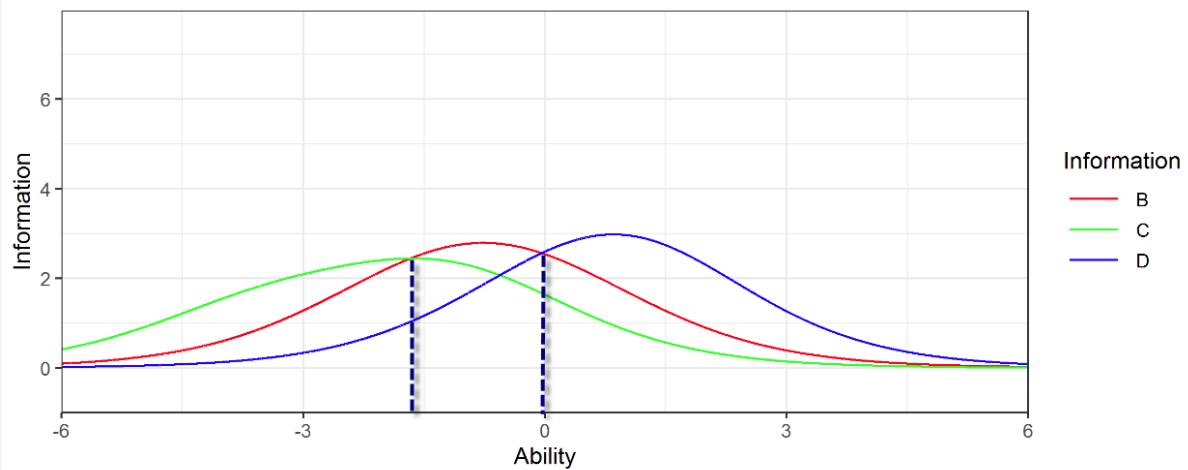


Figure 10: Test information functions: curves for testlets C, B and E

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 11 illustrates how the test characteristic curve for one testlet (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.

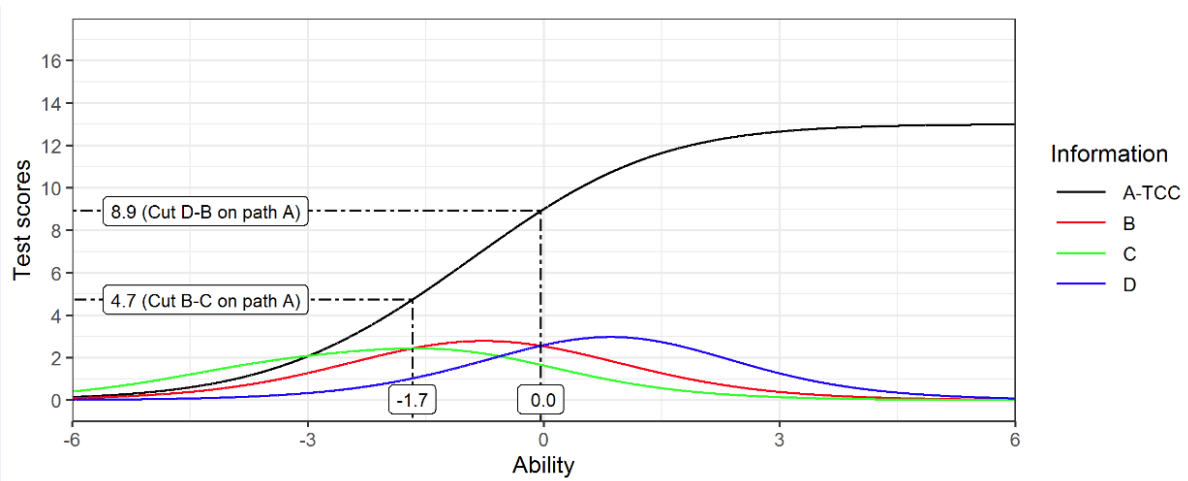


Figure 11: Stage 1. Testlet A-C|B|D cut scores

The branching rules for the first branching point discussed above are presented in Table 36 .

Table 36: 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 12 and Table 37).

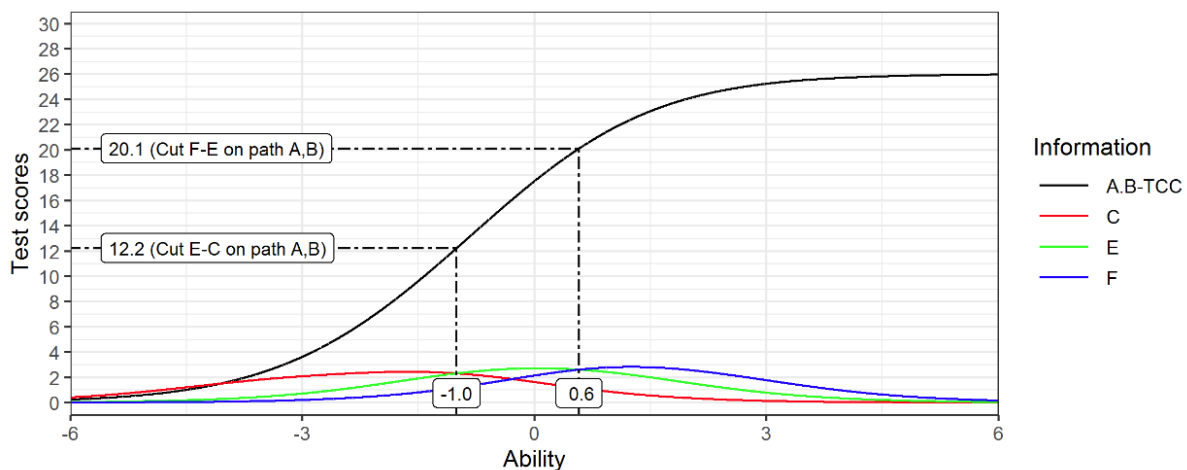


Figure 12: Stage 2. Testlet AB-C|E|F cut scores

In Figure 12, 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 37.

Table 37: 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 13, 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 38.

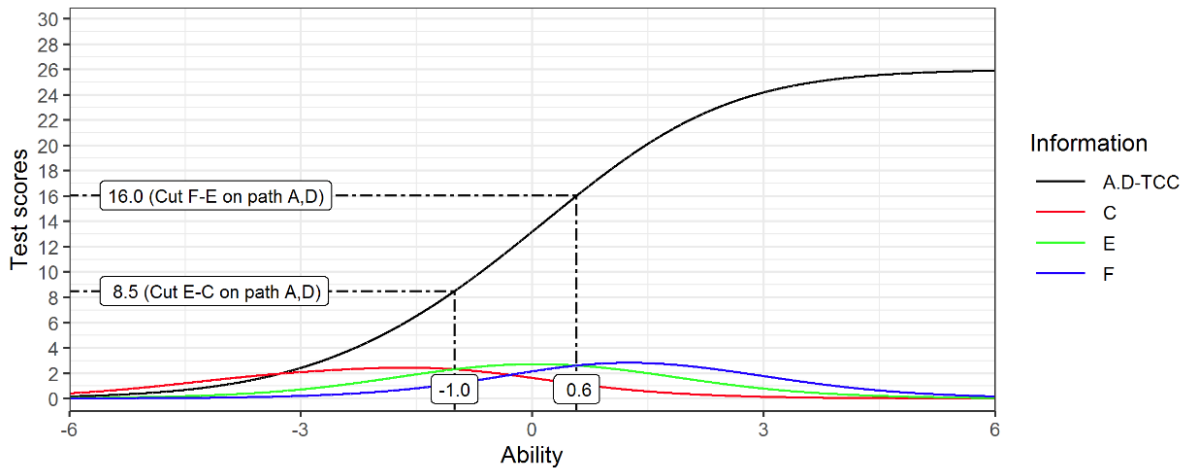


Figure 13: Stage 2. Testlet AD-C|E|F cut scores

Table 38: 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

Figure 9 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 reading and numeracy 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 14 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 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 14 and Figure 15.

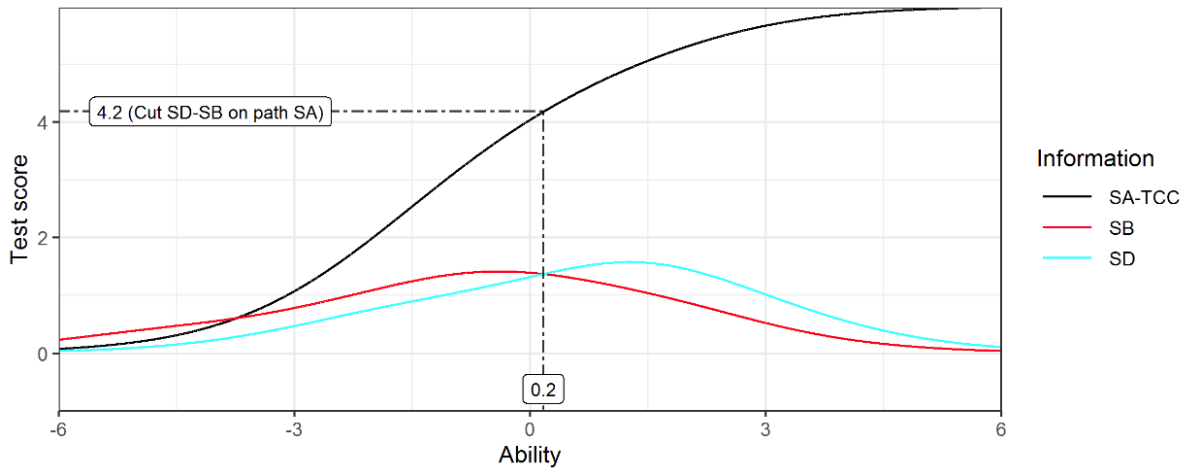


Figure 14: Stage 1. Testlet SA–SB|SD cut scores

In Figure 14, 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 39.

Table 39: 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

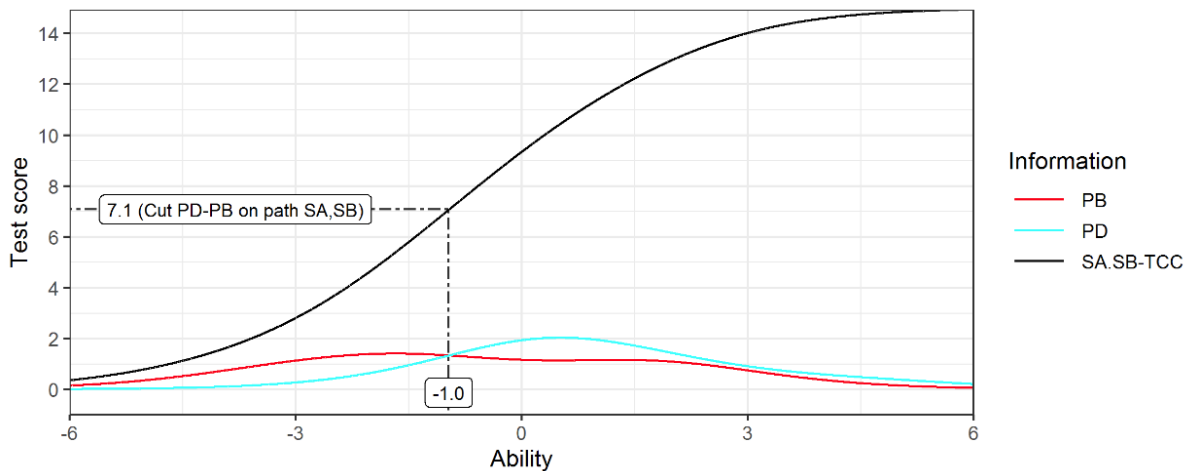


Figure 15: Stage 2. Testlet SA–SB to PB|PD cut scores

In Figure 16, 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 testlet SA and SB, then their ability estimate was in a region for which testlet PB provided most precision; whereas if a student had a raw score greater than 7 on testlet SA, 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 40.

Table 40: 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

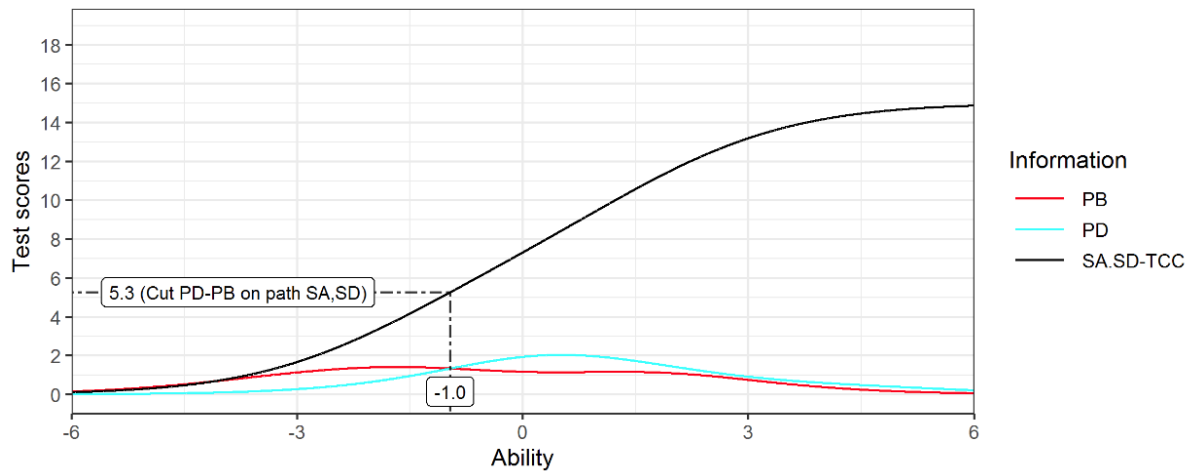


Figure 16: Stage 2. Testlets SA–SD to PB|PD cut scores

Figure 16 where 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 raw score greater than 5 on testlet SA, 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 41.

Table 41: 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 2019 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 17 and Figure 18.

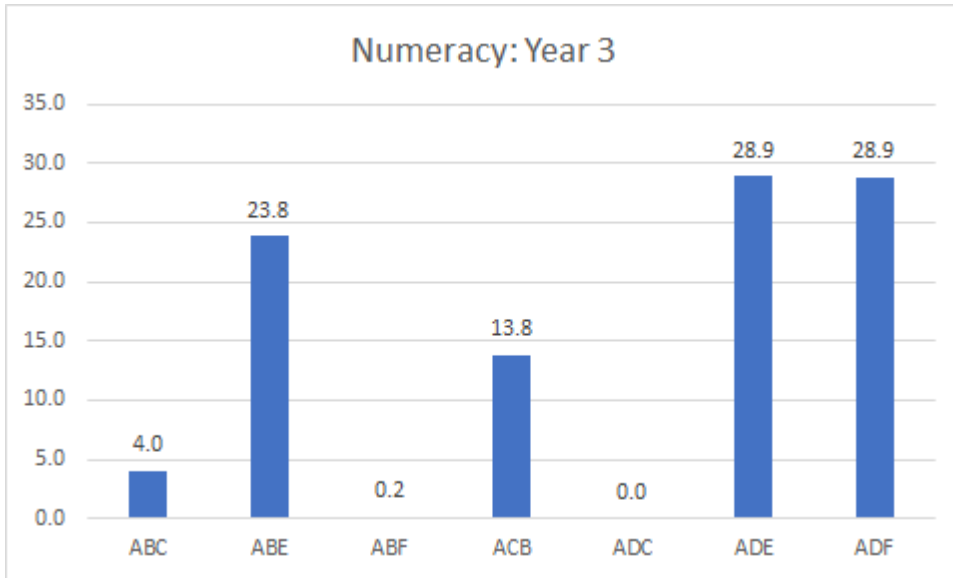


Figure 17: Percentage of students assigned to each pathway in Year 3 numeracy

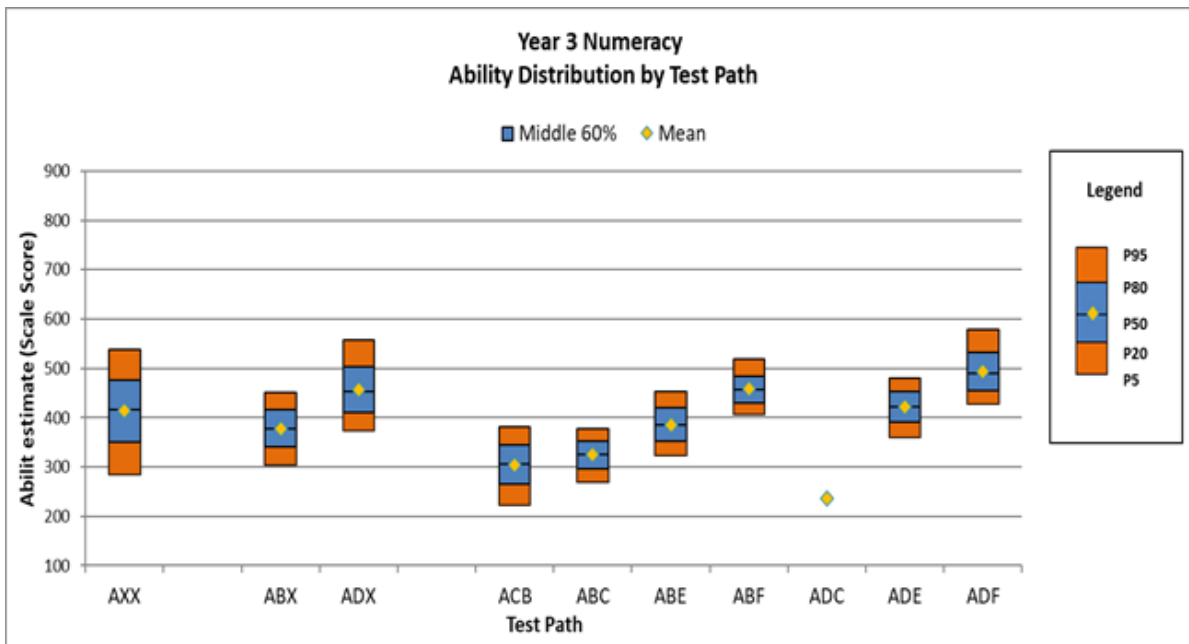


Figure 18: Ability distribution by pathway for Year 3 numeracy

As Figure 17 shows, the number of students assigned to each path varied from 0 per cent for ADC pathway to approximately 29 per cent in ADE and ADF pathways. 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.

Ability distributions by pathway are illustrated in Figure 18. 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 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 18 also shows that pathways overlapped in abilities.

Chapter 4: Sampling

There were two sampling activities associated with the NAPLAN 2019 administration:

- The national calibration sample: This was a sample of students who participated in the paper NAPLAN assessments. This sample was used for test calibration and scaling for the paper-based test, generating estimates of the key national outcomes to be used on parent reports and for the estimation of preliminary NAPLAN 2019 results on the existing NAPLAN scale. In previous cycle, when all students took the test on paper, this was also referred to as the calibration sample.
- Equating samples: These samples were required for equating the 2019 paper and online tests onto the historic NAPLAN reporting scales. Students in these samples took an equating test two weeks prior to the official NAPLAN test. Four equating samples were used for equating both the paper and online tests at the primary and secondary year levels respectively.

Sample frames

ACARA provided a list of Australian schools, which included the number of students enrolled at each school in Years 3, 5, 7 and 9 for 2019. Additional information such as state, sector, geographic location and aggregate NAPLAN performance data was also provided.

Calibration samples were drawn from the schools administering NAPLAN on paper so that analysis of data could start before data had been collected, scored and entered from all schools that administered NAPLAN on paper. Samples were drawn at each of the two levels of schooling for all jurisdictions, except for the ACT and Tasmania, because in those jurisdictions NAPLAN was administered online in nearly all schools. One sample was drawn based on the primary year levels (Years 3 and 5 for all jurisdictions, and Year 7 for South Australia) and a separate sample was drawn based on the secondary year levels (Year 7 for all jurisdictions except South Australia, and Year 9). Similarly, the equating samples were drawn at both primary and secondary levels.

The enrolment size for each target year level was based on the enrolment size in the equivalent year level for the previous calendar year. That is, the 2018 enrolment numbers for Years 3, 5, 7 and 9 given in the source data were used to estimate the 2019 enrolment sizes.

The calibration samples

A probability-based sample of students who participated in the paper NAPLAN assessments was selected and given priority in processing. This sample was used for the initial stage of the analysis process.

The following sections summarise aspects of the design of the calibration samples.

Exclusions

A number of school-level exclusions from the calibration sample were agreed with ACARA. These are summarised in Table 42.

Table 42: Calibration sample exclusions

Calibration sample exclusions
Primary and secondary levels
ACT and Tas. schools
2019 trial schools
Missing NAPLAN_Total_Cohort_3579 or 2017 NAPLAN performance data
Very remote schools in NSW or WA
Special schools
Distance schools
Schools with no enrolments in any relevant year levels
Schools sampled in the 2019 equating samples

For a state or territory to be included in the calibration sample, two criteria needed to be met:

1. At least 10 per cent of the schools in the state/territory population needed to be using the paper assessment mode.
2. At least 20 schools from the state/territory population were available for sampling.

The number and percentage of schools administering NAPLAN on paper and online are presented nationally and by jurisdiction in Table 43: Number and percentage of primary schools by assessment mode, nationally and for each jurisdiction and Table 44: Number and percentage of secondary schools by assessment mode, nationally and for each jurisdiction.

Table 43: Number and percentage of primary schools by assessment mode, nationally and for each jurisdiction

	Online		Paper	
	Number	Percentage	Number	Percentage
ACT	103	96%	4	4%
NSW	1,494	59%	1,046	41%
NT	29	18%	136	82%
Qld	404	27%	1,072	73%
SA	489	76%	158	24%
Tas.	214	97%	6	3%
Vic.	798	40%	1,173	60%
WA	823	88%	110	12%
Aus.	4,354	54%	3,705	46%

Table 44: Number and percentage of secondary schools by assessment mode, nationally and for each jurisdiction

	Online		Paper	
	Number	Percentage	Number	Percentage
ACT	40	95%	2	5%
NSW	484	49%	513	51%
NT	10	9%	96	91%
Qld	129	22%	465	78%
SA	171	68%	81	32%
Tas.	94	95%	5	5%
Vic.	263	34%	510	66%
WA	281	76%	90	24%
Aus.	1,472	46%	1,762	54%

Because of their high participation in the online mode of NAPLAN, ACT and Tas. did not contribute to the calibration sample at either the primary or the secondary levels. The other exclusion categories, relating to issues such as accessibility and remoteness, and the burden of schools participating in other survey work were similar to those used in previous NAPLAN test cycles.

Sample size

Table 45 shows the numbers of schools and students sampled for the calibration sample. The school sample sizes were very similar to the sizes used for this exercise in previous rounds of NAPLAN.

Table 45: Numbers of schools and students in the calibration sample

Year level	Source	NSW	Vic.	Qld	WA	SA	NT	Grand total
Year 3	Schools	53	51	53	44	48	25	274
	Students	2,246	2,065	2,259	1,302	1,442	319	9,633
Year 5	Schools	53	53	51	46	48	23	274
	Students	2,311	2,133	2,171	1,395	1,443	311	9,764
Year 7	Schools	58	56	57	36	48	24	279
	Students	6,631	6,633	7,566	2,268	1,551	597	25,246
Year 9	Schools	58	58	58	36	31	25	266
	Students	5,893	6,102	6,793	2,044	2,085	506	23,423
Total	Schools	222	218	219	162	175	97	1093
	Students	17,081	16,933	18,789	7,009	6,521	1,733	68,066

Stratification

The primary and secondary sample frames were explicitly stratified by state/territory and school sector. Because of high participation in the online study and following the application of exclusions according to the categories listed above, WA's participation in the primary calibration sample was limited to the independent sector. In total, the primary calibration sample frame was sampled from 16 explicit strata. For the same reasons, for the secondary calibration study, NSW's participation was restricted to the government and independent sectors, and WA's participation was restricted to the independent sector, leaving 15 explicit strata.

Within each explicit stratum, schools were implicitly stratified by the following four variables: NAPLAN performance quintiles, school type, geolocation, and mean enrolment. Each is briefly described below.

NAPLAN performance

Within each state/territory, NAPLAN performance values for schools provided by ACARA were used to determine quintiles. The quintile cut scores used to determine the performance levels within each state are shown in Table 46 for the primary sample and Table 47 for the secondary sample.

Table 46: Quintile cut points for primary school NAPLAN performance by jurisdiction

State/territory	20	40	60	80
NSW	-0.4573	-0.1372	0.1408	0.4367
NT	-3.1021	-2.3426	-1.9115	-0.7038
Qld	-0.5837	-0.3096	-0.0803	0.1860
SA	-0.5054	-0.2275	-0.1058	0.1344
Vic.	-0.2809	-0.0618	0.1372	0.3509
WA	-0.2270	-0.0317	0.1329	0.3757

Table 47: Quintile cut points for secondary school NAPLAN performance by jurisdiction

State/territory	20	40	60	80
NSW	-0.4642	-0.1727	0.1286	0.4271
NT	-3.2899	-2.8289	-2.2830	-1.4893
Qld	-0.5930	-0.2983	-0.0665	0.2295
SA	-0.4913	-0.1519	0.0539	0.2810
Vic.	-0.3926	-0.1483	0.0858	0.3169
WA	-0.2376	-0.0405	0.1548	0.4300

School type ID

This was an indicator denoting the inclusion (or absence) of Year 7 students in the sample. This variable was used to address the fact that Year 7 is a primary school level in South Australia but is a secondary level in the other jurisdictions.

Geolocation

The geolocation value was based on the Australian Bureau of Statistics (ABS) remoteness classification.

Table 48: ABS remoteness classification

Code	Description
0	Major cities of Australia
1	Inner regional Australia
2	Outer regional Australia
3	Remote Australia
4	Very remote Australia

Mean enrolment

This was the average of student enrolments across the target year levels within each school.

Sample selection

After sorting each explicit stratum by the implicit stratification variables, schools were selected using a random start, systematic sampling approach with equal probability selection of schools within each explicit stratum. This means that following the selection of the first school with a random start, every n^{th} school was selected from the stratum, with n being the number of schools in a stratum divided by the number to be selected from the stratum. After sampling the schools, all students from the target year levels in the selected schools were included in the sample. This sampling approach was applied so that all students from within the stratum had an equal chance of inclusion into the sample.

Substitution

To cater for the possibility that a sampled school could not be included, up to two substitute schools were selected for each sampled school. The substitute schools were those adjacent to the sampled schools at the time of sampling, and therefore were matched to the sampled school with respect to the key stratification variables. In some cases, due to limited availability of non-sampled schools within the region of the sampled school, only one substitute school could be selected. The chosen method for the selection of substitute schools is consistent with the methodology used in major international studies such as PISA and TIMSS.

Table 49 presents the participation rates by year level and state/territory for the original sampled schools, as well as for the combination of sampled and substitute schools.

Table 49: Percentage of sampled schools (%) included in the calibration sample (2019)

Year level	School type	NSW	Vic.	Qld	WA	SA	NT
Year 3	Sampled schools	85	93	96	90	94	59
	Sampled and substitute schools	98	94	98	90	98	83
Year 5	Sampled schools	85	94	93	94	94	62
	Sampled and substitute schools	98	96	94	94	98	72
Year 7	Sampled schools	97	97	86	82	92	47
	Sampled and substitute schools	100	97	98	82	98	77
Year 9	Sampled schools	97	100	84	82	83	47
	Sampled and substitute schools	100	100	97	82	83	67

Weighting

Weights were generated for the calibration sample so that the contributions of students reflected the population of students from each state and territory taking the paper assessments. The weights comprised components that reflected the sample selection probabilities, as well as adjustments for school- or student-level non-response. The individual components for the weighting are summarised below.

The School Base Weight (ScBWT)

The School Base Weight is the inverse of the probability of selection of the school. Because schools were sampled with equal probability, the School Base Weight in this case is equal to the total number of schools from the stratum divided by the number that were sampled from that stratum.

The Student Base Weight (StBWT)

For the calibration sample, the Student Base Weight equals 1 as all students at a sampled school who were in the appropriate year levels were selected.

The Preliminary Weight (Pwgt)

The product of ScBWT and StBWT, denoted the Preliminary Weight, reflects the selection probability of students into the calibration sample. Because schools were selected with equal probability within each stratum, and all students in the target grades from sampled schools were included, Pwgt equals ScBWT and has the same value for all students in the stratum. Because sample sizes varied across strata, Pwgt varies across strata.

The Senate Weight (Swgt)

For test calibration and scaling purposes, equal contributions from each participating jurisdiction were desired. To achieve this, the preliminary weights were scaled so that they summed to the same total as the sum of the preliminary weights from the jurisdiction with the largest number of participating students, to produce the so-called Senate Weight. This was done separately for each year level and test domain.

The School Non-response Adjustment (ScNRA)

If fewer schools (sampled or substitute) participated from a stratum than were sampled, a School Non-response Adjustment was applied, equal to the number of originally sampled schools divided by the number (sampled and substitute) that participated.

The Student Non-response Adjustment (StNRA)

If an eligible, non-excluded student from a sampled school did not participate (in any domain) in the NAPLAN assessment, then the respondents from that school were weighted up to represent all of the students sampled from that school. The Student Non-response Adjustment was calculated for each year separately, as the ratio of the expected number of participating students from the school at the relevant year level – the students who participated as well as the absent students – divided by the number of students who participated.

The Student Final Weight (Fwgt)

The student final weight was the product of the base weights, reflecting the sample design and probabilities of selection, and the school- and student-level non-response adjustments:

$$Fwgt = ScBWT * ScNRA * StBWT * StNRA.$$

Participation

The following tables present the distribution of participants in the calibration sample by gender, language background, indigenous status and location.

Table 50: Calibration sample: Distribution (%) by gender by year level (2019)

Year level	Gender	NSW	Vic.	Qld	WA	SA	NT
Year 3	Male	48.2	50.5	51.2	52.0	51.5	51.7
	Female	51.8	49.5	48.7	48.0	48.5	48.3
	Unspecified			0.1			
Year 5	Male	47.3	51.1	49.7	53.8	50.9	48.2
	Female	52.7	48.9	50.3	46.2	49.1	51.8
	Unspecified	0.0		0.0			
Year 7	Male	50.4	49.1	52.8	54.9	47.4	52.6
	Female	49.6	50.9	46.9	45.1	52.6	47.4
	Unspecified	0.0		0.2			
Year 9	Male	50.7	49.3	51.5	55.6	53.2	52.6
	Female	49.3	50.7	48.4	44.4	46.7	47.4
	Unspecified	0.0		0.2		0.0	

Table 51: Calibration sample: Distribution (%) by language background by year level (2019)

Year level	Background	NSW	Vic.	Qld	WA	SA	NT	Aus.
Year 3	Non-LBOTE	56.0	71.2	82.6	63.7	79.4	48.0	69.8
	LBOTE	43.3	28.8	17.2	34.5	19.5	48.8	29.4
	Not stated	0.7		0.2	1.8	1.1	3.1	0.7
Year 5	Non-LBOTE	55.8	74.0	85.0	68.6	80.2	49.0	71.6
	LBOTE	43.5	26.0	15.0	28.5	18.8	46.6	27.5
	Not stated	0.6		0.0	3.0	1.0	4.4	0.9
Year 7	Non-LBOTE	64.6	74.0	80.9	76.7	82.3	46.8	73.8
	LBOTE	32.6	26.0	19.1	21.5	16.8	37.1	24.8
	Not stated	2.8		0.0	1.8	0.9	16.2	1.4
Year 9	Non-LBOTE	65.0	74.4	80.3	77.7	87.6	42.7	74.6
	LBOTE	33	25.6	19.7	19.7	11.5	39.9	24.1
	Not stated	2.1		0	2.6	0.9	17.4	1.3

Table 52: Calibration sample: Distribution (%) by Indigenous status by year level (2019)

Year level	Indigenous status	NSW	Vic.	Qld	WA	SA	NT	Aus.
Year 3	Aboriginal and/or Torres Strait Islander	5.3	1.5	8.4	1.2	5.5	48.8	8.5
	Non-Aboriginal and/or Torres Strait Islander	93.3	98.3	86.4	98.4	92.4	51.2	89.6
	Not stated	1.4	0.2	5.2	0.4	2.1		1.8
Year 5	Aboriginal and/or Torres Strait Islander	5.4	1.5	8.4	1.3	5.6	49.9	8.2
	Non-Aboriginal and/or Torres Strait Isl.	93.2	98.3	85.7	98.2	92.5	49.9	89.8
	Not stated	1.4	0.2	5.8	0.5	2.0	0.3	2.0
Year 7	Aboriginal and/or Torres Strait Islander	5.7	1.6	9.5	2.2	4.6	48.6	7.4
	Non-Aboriginal and/or Torres Strait Islander	90.1	98.3	88.3	97.5	93.9	42.3	90.4
	Not stated	4.2	0.1	2.3	0.3	1.5	9.1	2.1
Year 9	Aboriginal and/or Torres Strait Islander	5.1	1.8	8.8	2.1	3.2	46.0	6.6
	Non-Aboriginal and/or Torres Strait Islander	90.9	98.0	89.0	96.6	95.4	52.2	91.5
	Not stated	4.0	0.2	2.2	1.3	1.4	1.9	1.9

Table 53: Calibration sample: Distribution (%) by geolocation by year level (2019)

Year level	Geolocation	NSW	Vic.	Qld	WA	SA	NT	Aus.
Year 3	Major cities	83.8	73.0	62.0	85.2	76.4		72.4
	Inner regional	9.9	20.8	18.3	9.2	12.9		14.2
	Outer regional	6.0	6.2	16.1	4.6	4.1	43.9	9.3
	Remote	0.4		2.6		6.3	13.3	2.1
	Very remote			1.0	1.0	0.3	42.8	2.0
Year 5	Major cities	83.8	73.1	64.0	85.2	75.4		73.0
	Inner regional	10.7	21.6	15.4	9.1	12.4		13.8
	Outer regional	5.3	5.3	16.3	5.3	4.4	47.1	9.1
	Remote	0.3		2.9		7.2	11.6	2.2
	Very remote			1.3	0.4	0.6	41.3	1.9
Year 7	Major cities	76.9	76.1	59.9	90.1	75.5		70.3
	Inner regional	18.4	17.6	17.8	4.9	14.7		16.0
	Outer regional	4.6	6.4	20.7	5.1	4.1	66.2	11.8
	Remote	0.1		0.4		5.4	4.2	0.6
	Very remote			1.2		0.2	29.6	1.2
Year 9	Major cities	76.9	77.9	61.1	89.6	71.7		70.8
	Inner regional	18.0	16.3	17.3	5.5	17.5		15.8
	Outer regional	5.0	5.9	20.4	4.9	3.1	64.9	11.4
	Remote	0.1		0.4		7.3	9.6	1.1
	Very remote			0.8		0.4	25.4	1.0

Equating samples

Equating samples were administered a secure equating test two weeks prior to the official NAPLAN tests. These results were required for equating the 2019 paper and online tests onto the historic NAPLAN reporting scales. As in earlier administrations, a common person equating method was used. Four equating samples were used, for equating both the paper and online tests at the primary and secondary year levels respectively.

The following sections summarise the design of the equating samples.

Exclusions

A number of school level exclusions from the equating samples were agreed with ACARA. These are summarised in Table 54.

Table 54: Equating samples (paper and online) – exclusions

Equating samples (paper and online) exclusions
2019 trial schools
2018 equating schools
Missing NAPLAN_Total_Cohort_3579 or 2018 NAPLAN performance data
Steiner, Waldorf, Montessori and Brethren (M.E.T) schools
Remote and very remote schools in all jurisdictions
Special schools
Distance schools
Enrolments < 15 in any relevant year levels
Schools sampled in the NAP-CC field trial and main study

Sample sizes

All jurisdictions contributed to the equating schools where possible. Because of the differential participation in the paper and online modes of assessment, and exclusions as noted above, state and territory contributions to the equating studies were limited in some cases.

For the online study, an additional 18 schools were sampled at both primary and secondary levels, with 9 schools (3 for each domain) identified from the top NAPLAN performance quintile and 9 identified from the bottom NAPLAN performance quintile. This was to boost the amount of response data for students at the top and bottom ends of the performance distribution. The schools were distributed across state and sector.

Table 55 shows the target school sample sizes for the four equating studies.

Table 55: Target school sample sizes, 2019 equating studies

State	Primary online	Primary paper	Secondary online	Secondary paper
ACT	4	0	2	0
NSW	34	26	35	33
NT	0	3	1	1
Qld	13	26	12	30
SA	12	3	11	6
Tas.	6	0	6	0
Vic.	21	27	24	35
WA	18	5	17	9
Aus.	108	90	108	114

Table 56 to Table 59 show the achieved number of schools and students in the equating sample for each domain by year level and state/territory.

Table 56: Achieved number of schools and students in the equating sample (reading) (2019)

Year level		NSW	Vic.	Qld	WA	SA	NT	Aus.
Year 3	School	8	9	9	2	1	1	30
	Student	201	212	180	43	26	21	683
Year 5	School	8	9	9	2	1	1	30
	Student	202	222	207	50	25	24	730
Year 7	School	11	12	10	3	1		37
	Student	311	276	239	73	22		921
Year 9	School	11	12	10	3	2		38
	Student	298	252	244	76	44		914

Table 57: Achieved number of schools and students in the equating sample (spelling) (2019)

Year level		NSW	Vic.	Qld	WA	SA	NT	Aus.
Year 3	School	9	9	9	1	1	1	30
	Student	229	207	214	19	27	21	717
Year 5	School	9	9	9	1	1	1	30
	Student	246	200	223	17	26	16	728
Year 7	School	9	11	9	3	1	1	34
	Student	227	259	203	75	28	24	816
Year 9	School	9	11	10	3	2	1	36
	Student	223	240	241	73	52	24	853

Table 58: Achieved number of schools and students in the equating sample (grammar and punctuation) (2019)

Year level		NSW	Vic.	Qld	WA	SA	NT	Aus.
Year 3	School	9	9	9	1	1	1	30
	Student	223	207	212	19	27	21	709
Year 5	School	9	9	9	1	1	1	30
	Student	245	201	224	17	26	16	729
Year 7	School	9	11	9	3	1	1	34
	Student	226	260	203	75	28	24	816
Year 9	School	9	11	10	3	2	1	36
	Student	224	241	242	73	51	24	855

Table 59: Achieved number of schools and students in the equating sample (numeracy) (2019)

Year level		NSW	Vic.	Qld	WA	SA	NT	Aus.
Year 3	School	9	9	7	2	1	1	29
	Student	223	219	152	50	22	22	688
Year 5	School	9	9	7	2	1	1	29
	Student	227	217	148	53	27	25	697
Year 7	School	10	12	10	3	1		36
	Student	260	274	246	69	22		871
Year 9	School	10	12	10	3	2		37
	Student	238	256	225	68	40		827

Stratification

For the equating samples, the sample frames were explicitly stratified by state/territory. For the online equating study at both primary and secondary level and for the paper equating study at the primary level, all eight jurisdictions were included. For the paper equating study at the secondary level secondary level, all jurisdictions except for ACT and Tas. were involved.

Within state and territory, schools were implicitly stratified by NAPLAN performance quintiles, sector, school type, geolocation and mean enrolment. Refer to the notes on the stratification for the calibration sample for a description of these variables.

Table 60 to Table 63 show cut points for the classification of schools into NAPLAN performance quintiles for the paper and online equating studies at the primary and secondary levels of schooling.

Table 60: Quintile cut points for the online primary school equating sample by jurisdiction

State/territory	20	40	60	80
ACT	-0.1451	-0.0365	0.1449	0.2234
NSW	-0.4076	-0.1403	0.0634	0.3228
NT	-1.5191	-1.1230	-0.3481	-0.3055
Qld	-0.4478	-0.1651	-0.0056	0.1755
SA	-0.5465	-0.3381	-0.1377	0.0991
Tas.	-0.7027	-0.3834	-0.1553	0.0855
Vic.	-0.1963	-0.0020	0.1737	0.3862
WA	-0.4210	-0.1689	0.0343	0.2905

Table 61: Quintile cut points for the paper primary school equating sample by jurisdiction

State/territory	20	40	60	80
NSW	-0.5231	-0.2142	0.0418	0.3417
NT	0.0000	0.0000	-1.4641	-0.3859
Qld	-0.5718	-0.2983	-0.0749	0.1642
SA	-0.7562	-0.3561	-0.1326	0.0297
Vic.	-0.3167	-0.0879	0.1090	0.3431
WA	0.0000	-0.2968	-0.0509	0.2098

Table 62: Quintile cut points for the online secondary school equating sample by jurisdiction

State/territory	20	40	60	80
NSW	-0.5208	-0.2179	0.0467	0.3425
Qld	-0.5874	-0.3450	-0.0886	0.1652
SA	-0.5769	-0.3163	-0.1346	0.0660
Tas.	-0.6713	-0.3823	-0.2064	0.1277
Vic.	-0.3184	-0.0943	0.1082	0.3450
WA	-0.6284	-0.2814	-0.0448	0.1667

Table 63: Quintile cut points for the paper secondary school equating sample by jurisdiction

State/territory	20	40	60	80
NSW	-0.5189	-0.2181	0.0427	0.3417
Qld	-0.5854	-0.3044	-0.0768	0.1438
SA		-0.3620	-0.1326	0.0241
Vic.	-0.3309	-0.0989	0.1037	0.3428
WA		-0.2968	-0.0509	0.2009

Sample selection

Following stratification and sorting, schools were selected using a random-start, systematic sampling where schools within each explicit stratum were selected using probability proportional to size (PPS). The total measure of size for the stratum was divided by the number of schools to be sampled to determine the stratum sampling interval. The selection probability for a school was equal to its measure of size divided by the sampling interval.

From each sampled school, one class per target year level was randomly selected for inclusion into the sample. The combined sampling approach across schools and classes within schools achieves the desirable outcome that students from a stratum are selected into the sample with equal probability.

Assignment to cognitive domains

Schools sampled for the equating studies were assigned to one of the three assessment domains: reading (R), language conventions (LC) or numeracy (N), as shown in Table 64 and Table 65.

Table 64: School domain assignment, online equating 2019

	Primary			Secondary		
	R	LC	N	R	LC	N
Year 3	36	36	36			
Year 5	36	36	36			
Year 7	4	4	4	33	33	33
Year 9				36	36	36

Table 65: School Domain Assignment, paper equating 2019

	Primary			Secondary		
	R	LC	N	R	LC	N
Year 3	30	30	30			
Year 5	30	30	30			
Year 7	1	1	1	36	36	36
Year 9				38	38	38

Chapter 5: Data collection and preparation

This chapter describes data collection and delivery, data validation and data preparation for NAPLAN 2019. 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. 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 and providing it to ACARA. ACARA and contractor then perform quality assurance on the final data received from each jurisdiction.

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

The purpose of the Data Standards Manual: Student Background Characteristics⁴ 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
- socio-educational background (parental occupation and education)
- language background.

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

- staggered delivery of online test data including both scored and raw response data (May)
- 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 (June)
- delivery of online test participation list, schools reverting to paper list, and mixed mode student data files from jurisdictions
- delivery of the paper test data from the national calibration sample from jurisdictions by year level for calibration and equating purposes for reading, spelling, grammar & punctuation, numeracy, and writing data (June)

⁴ www.acara.edu.au/reporting/data-standards-manual-student-background-characteristics

- delivery of the stage 1 census data (as near complete as possible, on paper and online tests) for analysis to produce the national summary report (July)
- delivery of the stage 2 complete census data (on paper and online tests) to produce the NAPLAN 2019 National Report (September).

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 national calibration samples of paper schools, both described in Chapter 4. The third round involved delivery of nearly complete stage 1 full cohort NAPLAN paper-based test data of Years 3, 5, 7 and 9 students in mid-July 2019. These data were used for the generation of the NAPLAN 2019 national summary information. The complete (with background data) full cohort data used for the production of the national report were delivered in September 2018. With each round of data delivery, the datasets were cleaned and recoded in preparation for analysis. A systematic process involving data checking was used to ensure that each dataset was consistent with national code frames and data dictionaries.

Online tests

The Education Services Australia (ESA) managed the online national assessment platform (platform) on which the NAPLAN 2019 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 started in April. 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.

Mixed mode tests

Data collection for mixed mode tests was undertaken by the test administration authorities (TAAs) in the jurisdictions. There were two rounds of data delivery for mixed mode tests. The first round of delivery of mixed mode student data was in mid-July 2019, after the delivery of complete stage 1 full cohort NAPLAN paper-based test data. The second round of delivery of mixed student data was in September 2019, around the same time of the delivery of the complete (with background data) full cohort data. With each delivery of mixed mode student data, the datasets were cleaned and combined with paper and online tests data for analysis and reporting.

Data cleaning validation process

All data files were checked for invalid codes and inconsistencies. Data were cleaned and recoded as part of the central data analysis process. 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.

In order to achieve a high-quality standard and eliminate errors, ACER implemented dual independent data processing: two data analysts processed the data independently for all test levels and subjects, one analyst using SAS software, and another analyst using IBM SPSS software.

ACER adopted the following quality assurance checks in its data cleaning procedures:

- identify duplicate records using unique student identifier information
- validate the categories of item responses, with invalid responses referred back to jurisdictions for further verification
- validate students' test participation against students' responses
- validate students' DOBs inside the range of the expected NAPLAN year at a year level
- check the frequencies of student background variable categories.

Data preparation

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

In 2019, data for multiple-choice items were indicated by the number of the chosen response options for each item; that is, 1, 2, 3, 4, or 5. Responses for students not participating on a particular test were recoded to 'R' and treated as *not administered*. Multiple responses ('7') were treated as *incorrect*. Embedded missing responses were coded as '9' by the TAAs 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 recoding are provided in Table 66.

Table 66: Rules for data recoding

Participation code	Data recoding rule
P – present	<p>Data string (i.e. item responses) expected. Any embedded missing responses are indicated with a 9 by the TAA, invalid responses with a 7.</p> <p>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.</p> <p>Students who are present but do not attempt any question will have a string of Ms.</p> <p>Additionally, for the online tailored test data, responses for items in those testlets that were not administered to the students are coded as an R.</p>
A – absent	A data string of all 9s 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	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. Response data are coded as an R.
W – withdrawn	A data string of all 9s for that test was expected from the TAA. See National Protocols for Test Administration, section 5.4. Response data are coded as an R.
E – exempt	<p>A data string of all 9s for that test was expected from the TAA. See National Protocols for Test Administration, section 5.2.</p> <p>These students are not included in the calibration or in the calculation of means. Item data are recoded as a string of Rs.</p>

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 67. 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 67: Pathway assignment rules to incomplete online tests

Domain	Last item attempted		Assigned pathway
N & R	None		ACB
N & R	Stage 1	A	ABE
N & R	Stage 2	B	ABE
N & R	Stage 2	C	ACB
N & R	Stage 2	D	ADE
S	None		SASBPB
S	Stage 1	A	SASBPB
S	Stage 2	B	SASBPB
S	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 the vast majority of items was an important consideration. This section provides relevant information, reported in terms of the percentage of trailing missing responses across all students for a given paper test or online test pathway.

Not reached items in paper tests

Figure 19 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, reading and spelling had the highest trailing missing rates, and grammar & punctuation had the lowest trailing missing rates. This test was also the shortest test and it was administered before the spelling test on the same day. 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, 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. A steep increase in the proportion of trailing missing responses was observed amongst the non-calculator items.

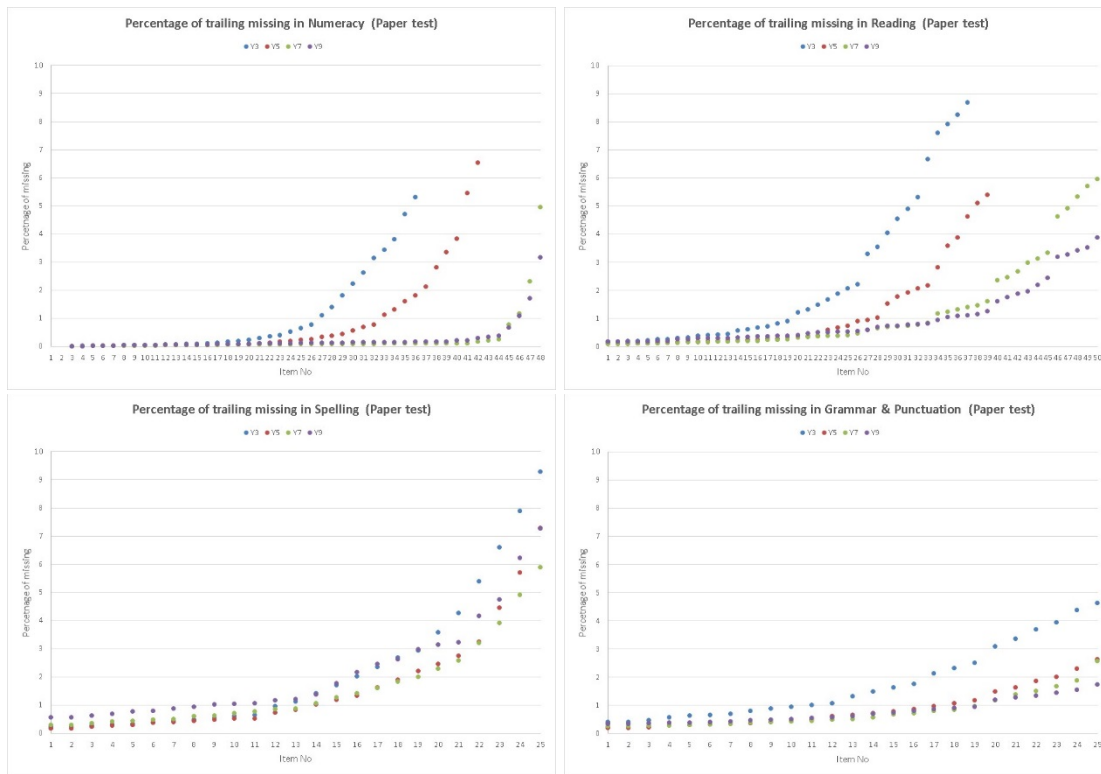


Figure 19: Trailing missing percentage in numeracy, reading, grammar & punctuation, and spelling

Not reached items in online tests

Figure 20 to Figure 23 show the percentage of trailing missing responses by year levels and test pathways in numeracy, reading, spelling and grammar & punctuation for the online tests. In these charts, the trailing missing responses were shown for one of parallel testlets (for example, testlets A1 to F1 for numeracy and reading, testlets C1 to F1 for grammar & punctuation, and testlets SA1 to PD1 for spelling). Across domains, grammar & punctuation had the lowest trailing missing rates, but it was also the shortest test. This test is the shortest test and it is also administered before the spelling test on the same day. In numeracy, reading 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 testpath A1D1F1 had the highest trailing missing rates in Years 5, 7 and 9 numeracy and reading. In spelling, the easiest testpath SA1SB1PB1 had the highest trailing missing rates in Years 3, 5, 7 and 9. Similar patterns of trailing missing responses were found in other parallel testlets.

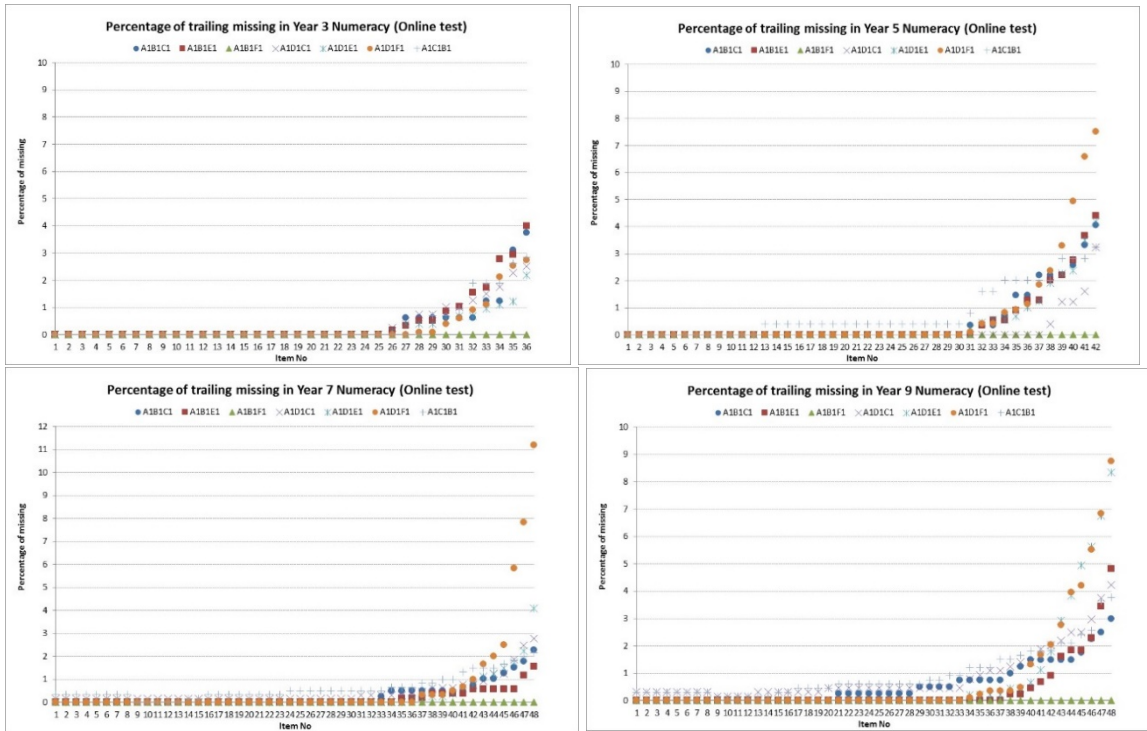


Figure 20: Trailing missing percentage in numeracy

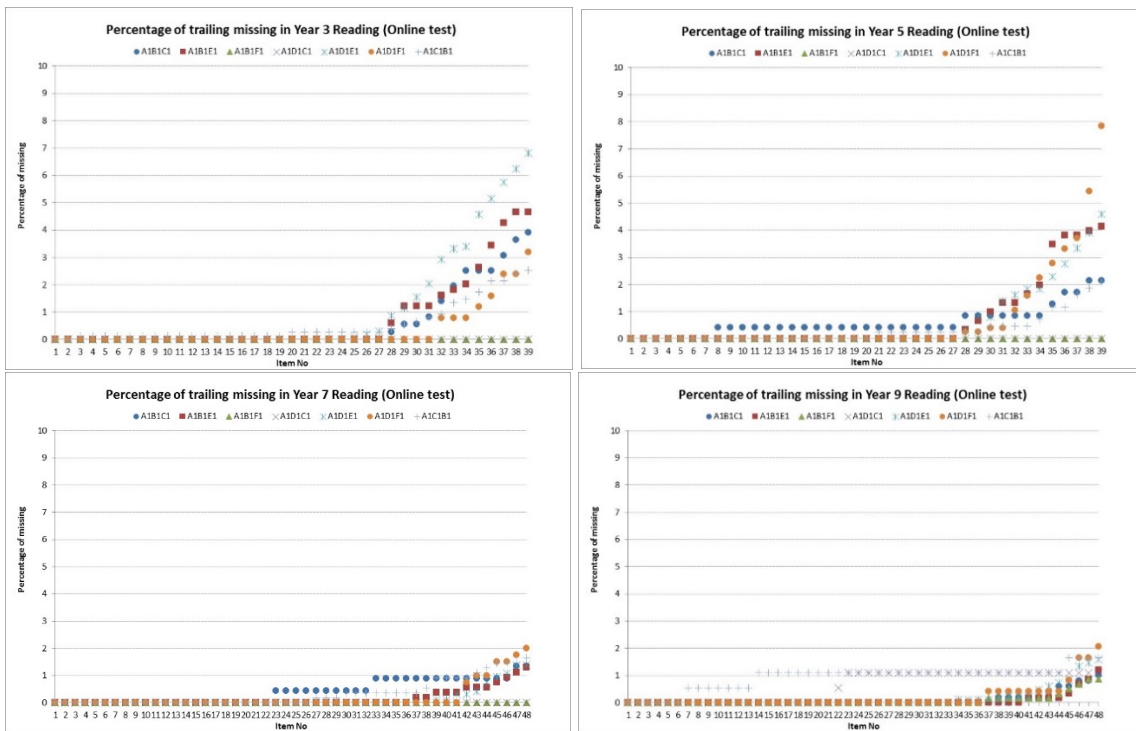


Figure 21: Trailing missing percentage in reading

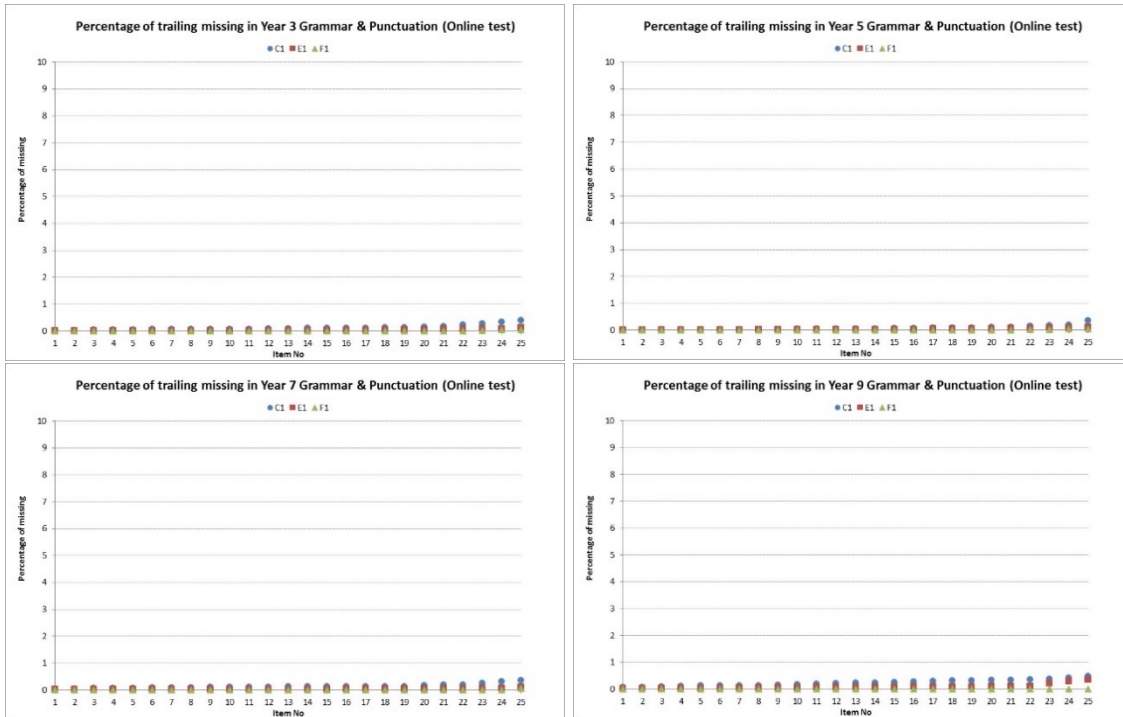


Figure 22: Trailing missing percentage in grammar & punctuation

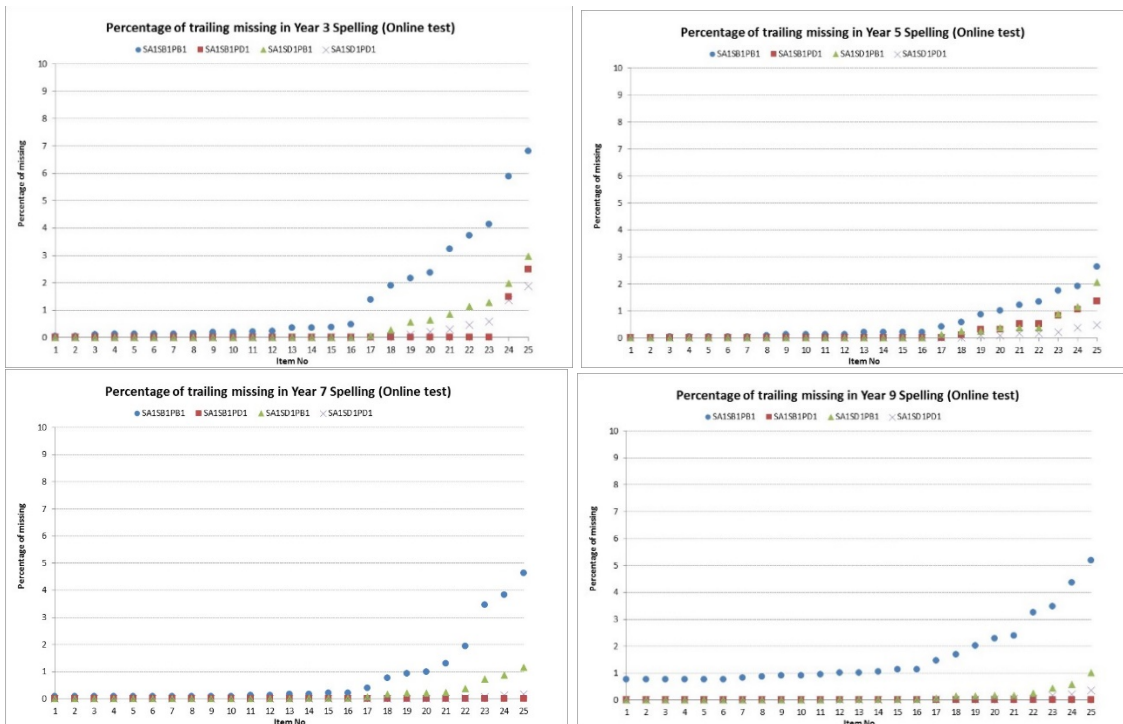


Figure 23: trailing missing percentage in spelling

Final student participation rates

Final student participation rates of NAPLAN 2019 are recorded in Table 68.

Table 68: Student participation rates by year level and domain, nationally and for each jurisdiction

TAA	Year level	Numeracy (%)	Reading (%)	Writing (%)	Spelling (%)	Grammar and punctuation (%)
NSW	3	96.3	96.8	96.2	96.7	96.7
Vic.	3	94.6	95.0	94.2	94.7	94.7
Qld	3	92.7	93.2	92.8	93.1	93.1
WA	3	95.0	96.1	95.0	95.4	95.4
SA	3	93.6	94.4	92.8	93.8	93.8
Tas.	3	96.1	96.9	95.3	96.2	96.2
ACT	3	94.8	95.0	94.6	94.8	94.8
NT	3	81.8	83.9	84.9	84.8	84.8
Aus.	3	94.6	95.2	94.5	94.9	94.9
NSW	5	96.5	97.1	97.1	97.0	97.0
Vic.	5	95.0	95.6	95.3	95.3	95.3
Qld	5	92.7	93.4	93.1	93.3	93.3
WA	5	95.4	96.6	96.5	95.9	95.9
SA	5	93.7	94.6	94.4	94.1	94.1
Tas.	5	95.5	96.9	96.3	96.1	96.1
ACT	5	94.9	95.5	95.4	95.0	95.0
NT	5	84.1	86.1	87.1	87.0	87.0
Aus.	5	94.8	95.5	95.4	95.3	95.3
NSW	7	95.4	96.3	96.4	96.1	96.1
Vic.	7	94.2	94.8	94.7	94.6	94.6
Qld	7	89.6	90.4	90.6	90.6	90.6
WA	7	94.6	96.1	96.2	95.2	95.2
SA	7	93.0	94.4	94.4	94.0	94.0
Tas.	7	93.9	96.3	96.2	95.1	95.1
ACT	7	93.0	94.2	94.3	93.5	93.5
NT	7	81.8	83.3	83.8	83.8	83.8
Aus.	7	93.4	94.3	94.4	94.1	94.1
NSW	9	92.2	93.5	93.9	93.4	93.4
Vic.	9	89.7	90.3	90.5	90.5	90.5
Qld	9	84.1	85.2	85.6	85.5	85.5
WA	9	92.4	94.3	94.6	93.0	93.0
SA	9	88.0	89.6	89.8	88.7	88.7
Tas.	9	89.0	92.5	91.9	91.0	91.0
ACT	9	87.4	88.4	89.2	88.2	88.2
NT	9	74.3	76.4	77.6	77.3	77.3
Aus.	9	89.2	90.4	90.7	90.3	90.3

Chapter 6: Scaling methodology and outcomes

This chapter describes the processes and methodologies used in the NAPLAN 2019 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).

The NAPLAN 2019 test calibrations for the paper tests were based on the item calibration sample. The NAPLAN 2019 test calibrations for the online tests were based on all available online data.

Scaling model

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

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

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

where $P(1|\theta)$ 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 $P(x|\theta_n)$ is the probability of person n to score x on item i . θ_n denotes the person's latent trait, 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

For paper tests, the reading, spelling, grammar and punctuation, and numeracy tests were calibrated separately by domain and year level, resulting in 16 separate calibrations. For each of the online reading, spelling and numeracy tests, items from all testlets within a domain and a year level were calibrated in a concurrent analysis, resulting in 12 separate calibrations. However, for Grammar and Punctuation test, items were calibrated by each generic testlet, that was by testlet Cs, E1&E2, E3 and Fs. The generic testlet Cs contain two testlets C1 and C2. The generic testlet Fs contain two testlets F1 & F2. Because there were very few common items between testlet E3 and testlets E1&E2, testlet E3 was calibrated separately from testlets E1&E2. Thus at each year level, there were four calibrations and a total of 16 calibrations were carried out for online Grammar and Punctuation tests. Thus, a total of 28 calibrations were carried out for online tests.

For 2019 writing, the resulting scripts from students who responded on paper or online from different tasks were rated using the same marking rubric based on the ten criteria. The writing test 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 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 (7, including *absent*, *withdrawn* and *exempt*) were treated as *not administered*, and embedded-missing (9) and invalid response (8) were treated as *incorrect* responses. The senate weight was used for item calibration to ensure each jurisdiction was equally represented. 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.

For each jurisdiction, the senate weight was calculated according to the following equation:

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

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

Review of test and item characteristics

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

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. Any statistics shown at the end of the item analysis results for the online reading, spelling and numeracy tests are to be ignored as these were not for any one test but were for the whole item pool at each year level.

The Rasch item parameter estimates and statistics are summarised in Appendix C for each of the 16 paper tests (numeracy, reading, spelling and grammar & punctuation for four year levels), the online items in each of the 12 item pools for the reading, spelling and numeracy tests for four year levels and each of the 16 grammar and punctuation calibrations (C, E1 & E2, E3, F for 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 (paper-based and online) are shown in Appendix D. The ICC plot shows a comparison of the empirical ICC based on observations from 10 equal-size ability groupings (broken line joining 10 dots) and the expected model-based ICC (smooth line). 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 distractor response curves, the performance of distractors can be examined using the item analysis results and the response curves in the ICC plots.

Test reliability

Table 69 shows the IRT-based reliability (WLE) of each paper test and online test.

For the online tests, the reliabilities were between 0.88 and 0.89 for the reading tests, between 0.92 and 0.93 for the spelling tests and between 0.90 and 0.94 for the numeracy tests. The reliability for the writing test was 0.95. For grammar and punctuation, the tests were calibrated by four testlets – testlet C, testlet E1&E2, testlet E3 and testlet F – because of the lack of links between them. The reliabilities were between 0.56 and 0.79.

For the paper tests, the reliabilities were between 0.82 and 0.86 for reading, between 0.88 and 0.90 for spelling, between 0.70 and 0.76 for grammar and punctuation, and between 0.87 and 0.92 for the numeracy tests. The reliability for the writing test was 0.94. In general, the reliability of online tests was somewhat higher than the reliability of the paper tests.

Table 69: Reliability (WLE) for NAPLAN 2019 paper tests

Test mode	Year level	Reading	Spelling	Grammar and punctuation	Numeracy	Writing*
Online	3	0.89	0.93	C: 0.79 E1&E2: 0.61 E3: 0.68 F: 0.66	0.90	0.95
	5	0.88	0.92	C: 0.75 E1&E2: 0.62 E3: 0.70 F: 0.68	0.92	
	7	0.89	0.92	C: 0.68 E1&E2: 0.63 E3: 0.62 F: 0.56	0.93	
	9	0.88	0.92	C: 0.67 E1&E2: 0.57 E3: 0.74 F: 0.58	0.94	
Paper	3	0.82	0.90	0.75	0.87	0.94
	5	0.84	0.89	0.70	0.90	
	7	0.86	0.89	0.75	0.92	
	9	0.86	0.88	0.76	0.92	

*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 variable map. Appendix E provides the variable 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, with the exception of grammar and punctuation 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 per cent 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 variable 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 24 displays the variable map for Year 3 numeracy paper test. That variable map indicates that the current test targeted the average numeracy achievement level of the student group quite well. The distribution of student abilities (each X represents approximately 16 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 25 with the latent ability distribution on the left-hand side. Figure 26 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

Chapter 6: Scaling methodology and outcomes

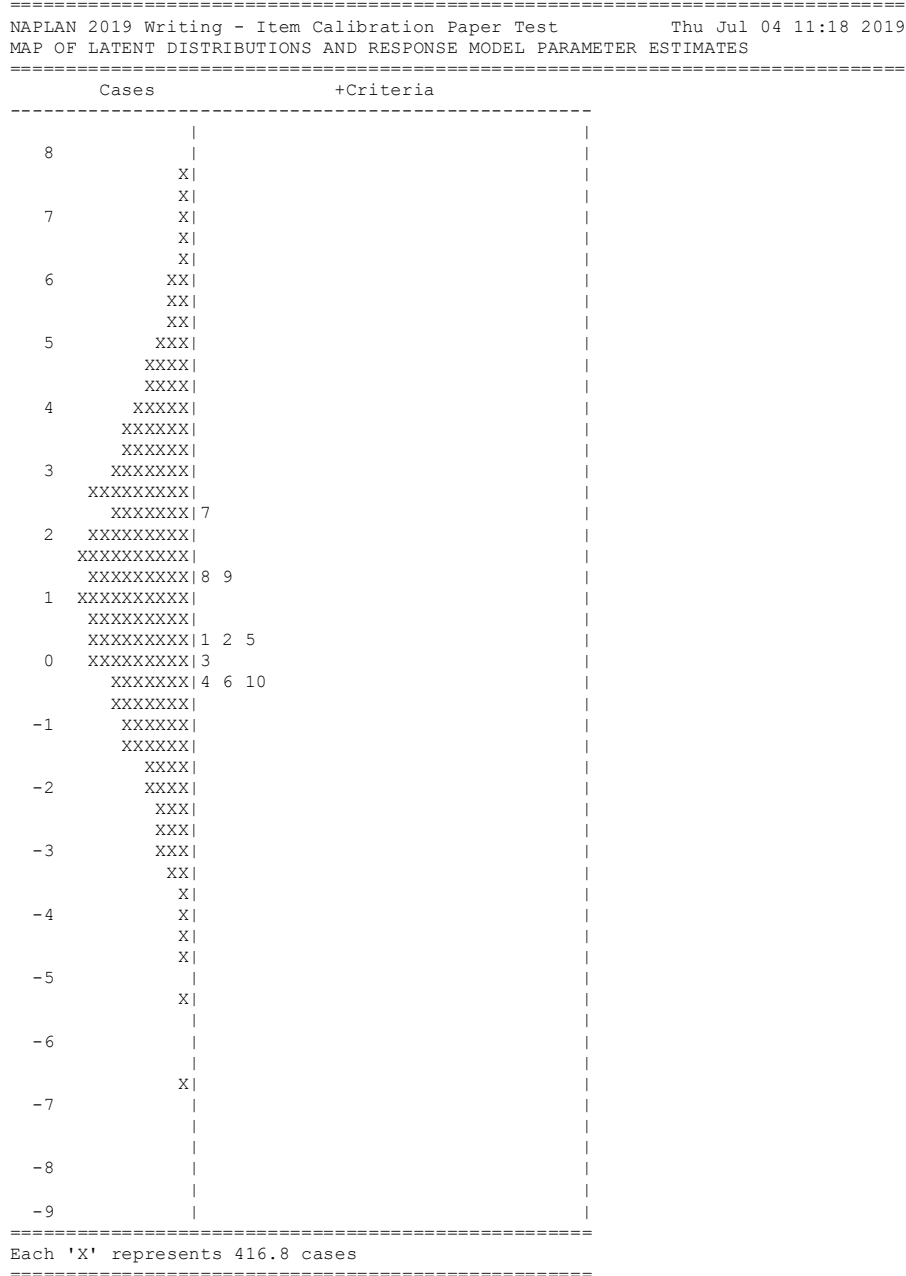


Figure 25: Wright map for paper writing test (a polytomous example)

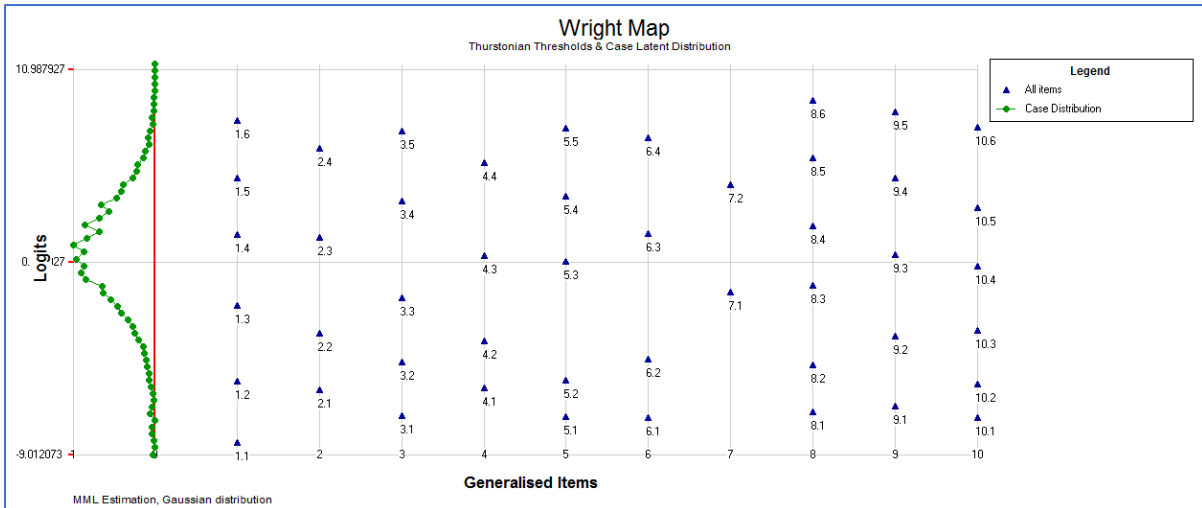


Figure 26: Thurstonian thresholds for writing test

Since the online tailored test design improves test targeting to students' abilities, it is not necessary for item spread within a tailored test to be as wide as a fixed test designed for all students. Each tailored test caters for a narrower range of student abilities. Figure 27 presents an illustrative example. The figure shows the information function for the NAPLAN 2019 Year 5 reading test path (A1B1E1). The information function is plotted in a blue curve and the standard error of measurement in an orange curve. The peak of the information function corresponds to the lowest standard error. This peak shows the range on the scale at which the test information is the highest and the standard error of measurement is the lowest. When moving away from this peak in either direction along the scale, the test information decreases and the standard error of measurement increases. The student ability distribution is shown in a bar below the horizontal axis with the mean, the fifth and ninety-fifth percentiles indicated on the bar. The peak of information function for this path (A1B1E1) corresponds to the middle of the ability distribution, indicating that the test is well targeted to students allocated to this pathway.

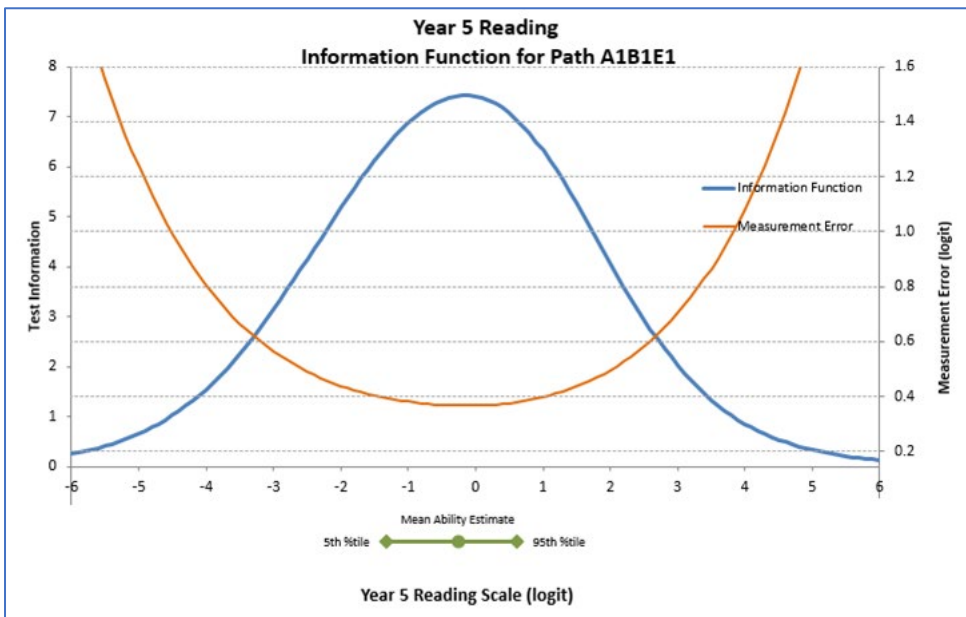


Figure 27: Year 5 reading information function for test path A1B1E1

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. A value greater than the expected value of 1.0 indicates that the item responses contain a greater amount of variability than expected by the model, and a value below 1.0 indicates that the consistency between observed data and model-based predictions is better than expected. 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 underfit). We also calculated classical item statistics (that is, item-rest score correlation and facility) for the purpose of item fit evaluation, specifying criterion values for discrimination (based on item-rest score correlation) less than 0.25 and facility outside the range of 0.10 to 0.90. Values of the infit mean square and classical item statistics of each item can be found in appendices B and C for the paper-based tests and online tests.

As mentioned above, the ICC of each item shows a comparison of the empirical ICC based on observations from 10 equal-size 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 reading, spelling, grammar and punctuation and numeracy at each of the four year levels. As all items were trialled 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 70 presents a summary of item statistics in NAPLAN 2019 paper tests with the number of items falling into two infit mean square ranges of less than, or equal to, 1.20, and greater than 1.20. It also presents the number of items with discrimination less than 0.25 and the number of items with facility outside the range of 0.10 to 0.90. As seen from the table, there were 15 items across 16 tests having infit greater than 1.20. Regarding classical test statistics, there was a total of 92 items (out of a total 550 items) across the 16 tests with discrimination less than 0.25. There were 38 items with facility higher than 0.90 and 16 items with facility less than 0.10. Figure 28 shows the ICC of one reading Year 3 item (item x00073229) with an infit statistic close to 1.01. In contrast, Figure 29 shows the ICC of one grammar and punctuation item (item x00017700) with an infit statistic (1.25) 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 paper tests (with writing) and for each of the 29 online 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 paper writing, the criteria punctuation and spelling exhibited misfit to the Rasch partial credit model (that is, infit are 1.42 and 1.34, respectively). For online writing, there were two additional criteria, paragraphing and punctuation, exhibiting misfit, (infit are 1.33 and 1.50, respectively). None of the other criteria exhibited misfit to the Rasch partial credit model. Inspection of the ICCs did not reveal large differences between the empirical and the expected curves for each of the ten criteria. The ICCs of the 10 writing criteria for both paper and online writing are included in Appendix D.

Table 70. Summary of item statistics in NAPLAN 2019 paper tests

Domain	Year level	Total number of items	Number of items with item-rest correlation <0.25	Number of items with		Number of items with	
				Infit ≤ 1.2	Infit > 1.2	Facility > 0.90	Facility < 0.10
Reading	3	37	12	37	0	1	1
	5	39	7	39	0	7	1
	7	50	14	50	0	4	2
	9	50	12	50	0	4	1
Spelling	3	25	0	24	1	1	1
	5	25	0	23	2	1	2
	7	25	1	24	1	0	1
	9	25	0	24	1	1	1
Grammar and punctuation	3	25	7	25	0	1	0
	5	25	11	25	0	2	1
	7	25	8	25	0	3	1
	9	25	8	24	1	2	0
Numeracy	3	36	4	35	1	2	1
	5	42	3	39	3	4	0
	7	48	2	44	4	3	1
	9	48	3	47	1	2	2

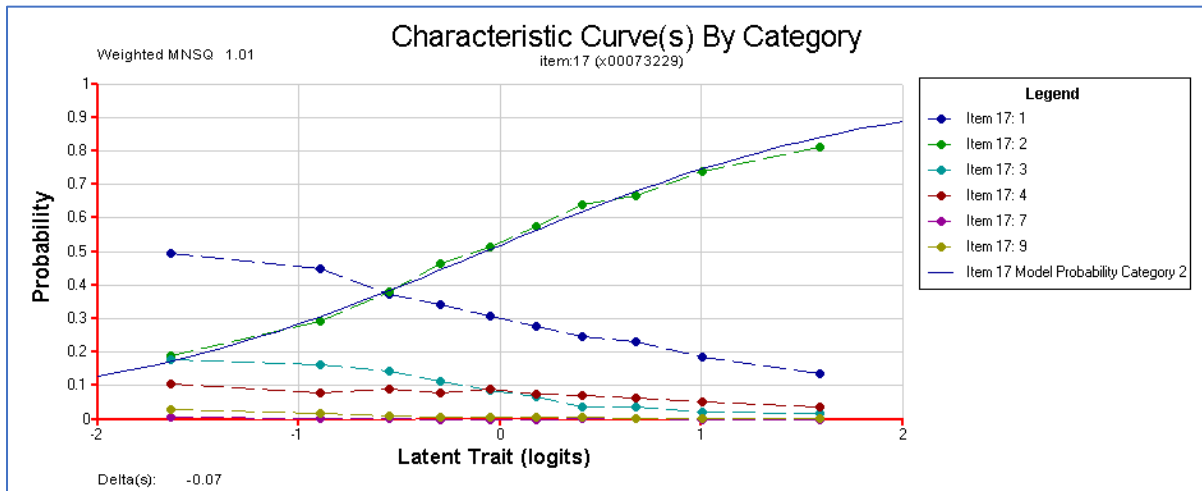


Figure 28: Item characteristic curves for an item with infit =1.01

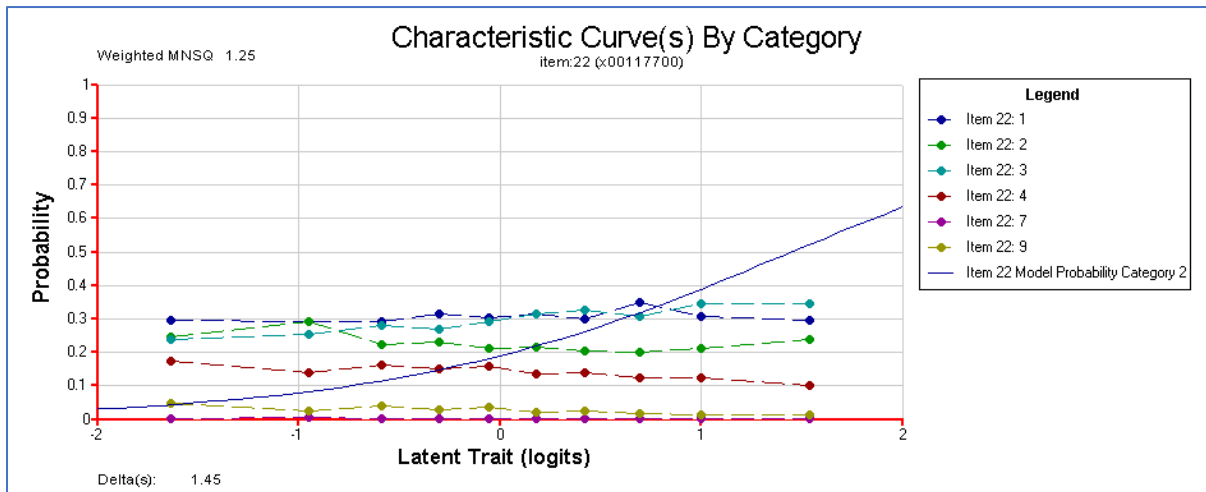


Figure 29: Item characteristic curves for an item with $infit = 1.25$

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 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. A diagonal line 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.

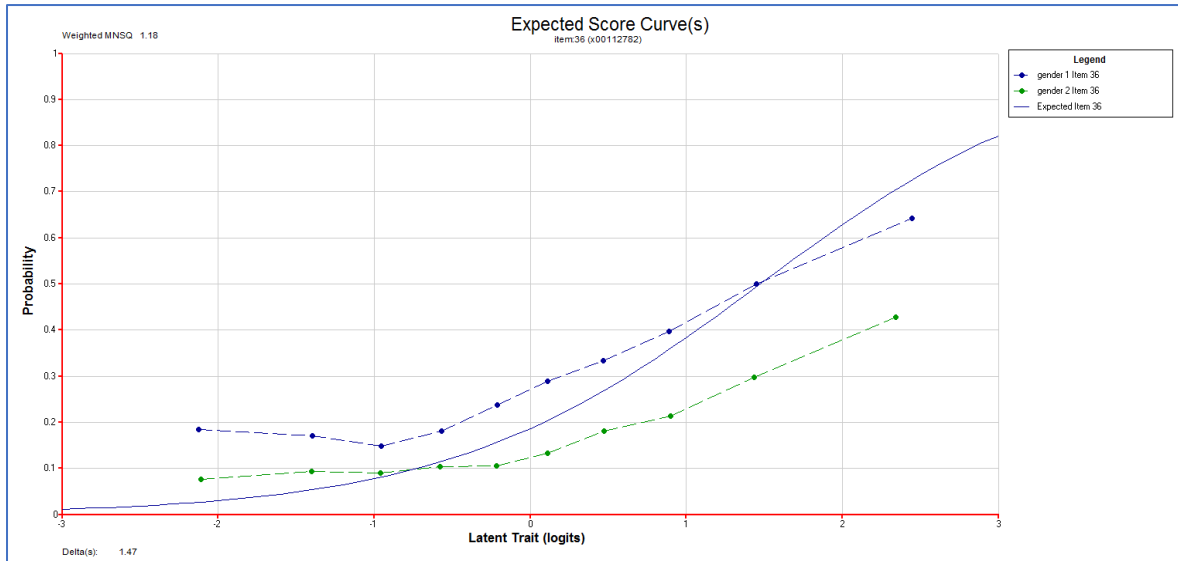
Gender DIF

Appendix F presents the scatter plots for examining gender DIF in the five domains for both paper and online tests. The plots for reading, spelling, grammar and punctuation, and numeracy 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 71 identifies the number of items that show gender DIF with an absolute adjusted difference of 0.50 or greater for reading, spelling, grammar and punctuation, and numeracy. Figure 30 shows as an example, one Year 9 numeracy paper test item (Item x00112782) with an absolute adjusted difference of 0.50 or greater. This item with a positive adjusted difference indicates that the item was relatively easier (adjusted difference = 0.80) 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 71. Number of items showing gender DIF by domain by year level

Test mode	Year level	Reading		Spelling		Grammar and punctuation								Numeracy	
		Total of test items	Total of DIF items	Total of test items	Total of DIF items	Total of test items				Total of DIF items				Total of test items	Total of DIF items
Paper	3	37	2	25	0	25				0				36	2
	5	39	2	25	1	25				1				42	5
	7	50	2	25	0	25				0				48	3
	9	50	3	25	2	25				0				48	4
Online						GC	GE1 /GE2	GE3	GF	GC	GE1 /GE2	GE3	GF		
	3	200	8	94	4	31	25	25	31	0	0	0	0	178	19
	5	193	6	96	11	31	25	25	30	0	0	0	0	215	21
	7	236	14	96	8	30	25	25	31	0	0	1	0	225	21
	9	246	14	95	10	31	25	25	29	0	0	3	1	221	17



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

Figure 30: Example of item characteristic curves displaying gender DIF†

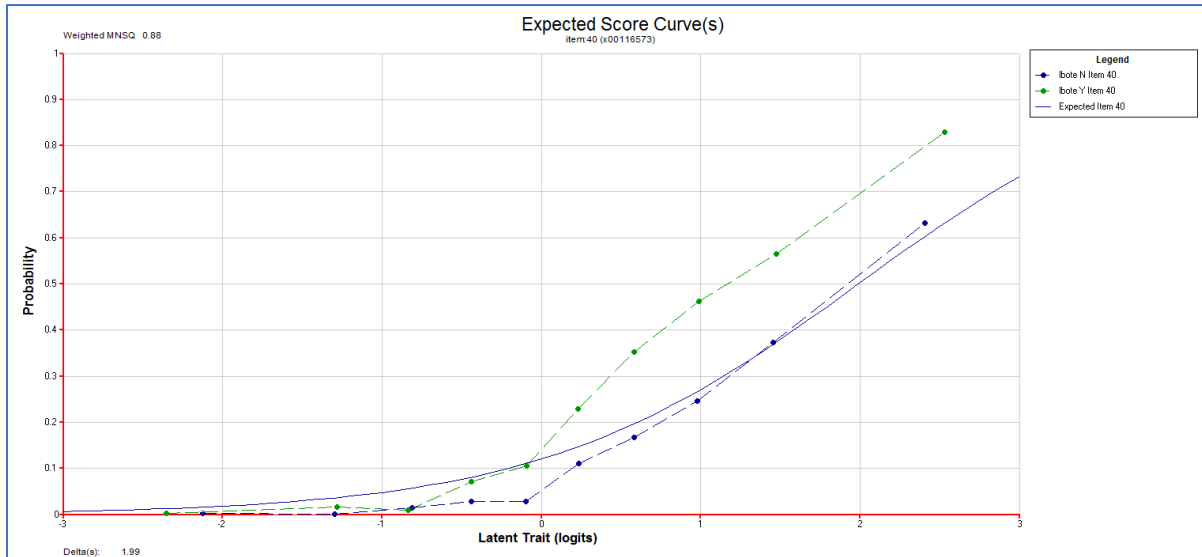
Language background DIF

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 72 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 31 depicts one Year 5 numeracy paper test item (item x00116573) with an absolute mean difference of 0.50 or greater. This item was relatively easier (mean difference = -0.88) for LBOTE students.

Table 72. Numer of Items Showing LBOTE DIF by Domain by Year Level

Test mode	Year level	Reading		Spelling		Grammar and punctuation				Numeracy					
		Total of test items	Total of DIF items	Total of test items	Total of DIF items	Total of test items		Total of DIF items		Total of test items		Total of DIF items			
Paper	3	37	0	25	4	25		1		36		1			
	5	39	0	25	1	25		1		42		2			
	7	50	0	25	0	25		0		48		3			
	9	50	1	25	2	25		4		48		4			
Online						GC	GE1/GE2	GE3	GF	GC	GE1/GE2	GE3	GF		
	3	200	2	94	9	31	25	25	31	1	4	3	5	178	7
	5	193	1	96	11	31	25	25	30	3	0	1	0	215	13
	7	236	3	96	7	30	25	25	31	0	0	3	1	225	10
	9	246	4	95	10	31	25	25	29	3	4	0	1	221	12



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

Figure 31: Example of item characteristic curves displaying LBOTE DIF†

Indigenous status DIF

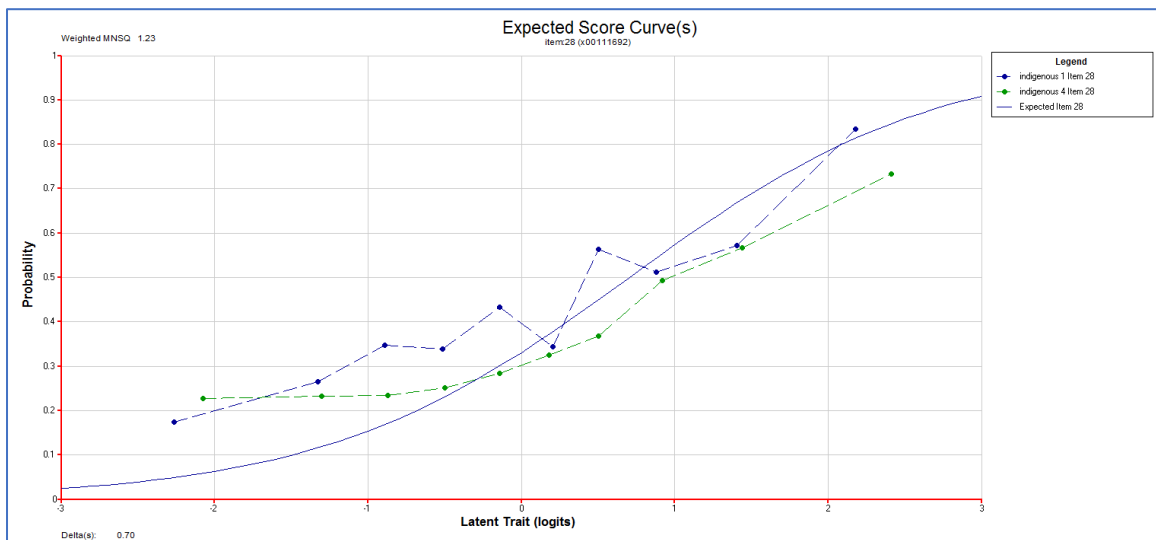
Appendix H includes scatter plots for examining Indigenous DIF in the five domains for both paper and online tests. Writing Indigenous DIF was performed by combining all four grades. These plots showed that there were not many items that showed notable differences in the relative item difficulties.

Table 73 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 32 depicts one reading item (item x00116936) with an absolute mean difference of 0.60 or greater. This item was relatively easier (mean difference = -1.22) for Indigenous students.

Appendix H provides the item DIF plots for items listed in Table 73. 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 73. Numer of items showing Indigenous DIF by domain by year level

Test mode	Year level	Reading		Spelling		Grammar and punctuation				Numeracy					
		Total of test items	Total of DIF items	Total of test items	Total of DIF items	Total of test items		Total of DIF items		Total of test items	Total of DIF items				
Paper	3	37	9	25	0	25		4		36	5				
	5	39	8	25	0	25		5		42	5				
	7	50	8	25	0	25		5		48	4				
	9	50	3	25	0	25		5		48	3				
Online						GC	GE1 /GE2	GE3	GF	GC	GE1 /GE2	GE3	GF		
	3	200	6	94	0	31	25	25	31	0	0	0	1	178	1
	5	193	2	96	0	31	25	25	30	0	2	0	1	215	3
	7	236	4	96	0	30	25	25	31	0	1	3	3	225	2
	9	246	4	95	0	31	25	25	29	0	1	0	4	221	6



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

Figure 32: Example of item characteristic curves displaying Indigenous DIF†

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

Jurisdictional DIF

The number of items showing statistically significant state/territory related DIF in paper and online reading, spelling, grammar and punctuation, and numeracy are shown in Table 74. In the headings of Table 74, 'E' indicates that the item is relatively easier for the jurisdiction, and 'H' indicates that the item is relatively harder for the jurisdiction. For paper tests, there were 73 potential DIF in reading, 24 in spelling, 16 in grammar and punctuation, and 25 in numeracy across all four year levels across the jurisdictions. Table 74 can be read in conjunction with Appendix I, from which the items showing

state/territory related DIF can be identified. For example, from Table 74, there was one item in Year 3 reading showing DIF in Vic. when compared with the national level, with this item (x00073229) being easier for VIC, as seen from Appendix I.

Table 74. Number of items showing state/territory DIF by domain by year level

Domain	Year level	ACT		NSW		NT		Qld		SA		Tas.		Vic.		WA	
		E	H	E	H	E	H	E	H	E	H	E	H	E	H	E	H
Paper																	
Reading	3	-	-	0	0	0	0	0	0	0	0	-	-	1	0	0	0
	5	-	-	0	0	0	0	0	0	0	0	-	-	0	0	0	0
	7	-	-	1	0	0	1	0	0	0	0	-	-	2	3	0	0
	9	-	-	2	3	2	2	43	2	2	0	-	-	2	0	2	5
Spelling	3	-	-	0	0	0	0	1	0	1	0	-	-	0	0	0	0
	5	-	-	0	5	0	0	0	0	0	0	-	-	0	0	1	0
	7	-	-	0	7	0	0	1	0	0	0	-	-	0	1	0	1
	9	-	-	1	1	0	0	0	3	0	0	-	-	0	0	1	0
Grammar and punctuation	3	-	-	0	0	0	1	0	0	0	0	-	-	0	2	0	0
	5	-	-	0	0	0	0	0	0	0	0	-	-	0	0	0	0
	7	-	-	0	0	1	0	1	0	0	0	-	-	0	2	0	1
	9	-	-	0	1	1	2	0	0	0	0	-	-	1	0	1	2
Numeracy	3	-	-	0	1	1	0	0	0	0	0	-	-	0	1	0	0
	5	-	-	1	1	0	0	1	1	0	0	-	-	0	0	0	0
	7	-	-	0	2	0	0	2	1	0	0	-	-	1	1	1	1
	9	-	-	3	1	0	0	1	2	0	0	-	-	1	1	0	0
Online																	
Reading	3	0	0	3	5	1	0	1	1	0	0	0	0	4	2	0	2
	5	0	0	1	2	0	0	1	1	0	0	0	0	1	2	0	0
	7	0	0	1	2	0	0	1	1	0	0	0	0	2	0	1	0
	9	0	0	3	8	0	0	1	0	1	0	0	0	1	1	0	1
Spelling	3	1	0	12	15	1	0	4	2	1	2	0	0	6	5	10	9
	5	0	1	9	9	0	0	7	2	0	0	0	0	3	0	9	3
	7	1	0	10	6	0	0	1	1	0	2	2	2	1	0	4	1
	9	0	0	11	5	1	0	2	0	0	1	1	0	1	0	3	6

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	
Grammar and punctuation	3	C	0	0	1	1	0	0	1	0	0	0	1	0	0	0	0	0
		E1E2	0	0	4	3	0	0	0	1	0	1	0	0	1	1	4	1
		E3	0	0	3	0	0	0	1	1	1	0	0	0	1	1	0	2
		F	0	0	3	2	0	0	1	0	0	0	0	0	1	0	2	0
	5	C	0	0	1	0	0	0	0	1	0	0	0	0	1	1	1	0
		E1E2	0	0	3	4	0	0	3	1	0	0	0	0	4	3	1	0
		E3	0	1	1	0	0	0	0	2	0	0	0	0	3	0	2	0
		F	0	0	8	4	1	0	2	0	0	0	0	0	2	0	2	0
	7	C	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
		E1E2	0	0	5	2	0	0	2	1	0	0	0	1	1	1	1	1
		E3	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
		F	0	0	3	3	1	0	1	1	0	0	0	0	0	0	1	0
	9	C	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
		E1E2	0	0	2	1	0	0	0	1	0	0	0	1	0	0	3	0
		E3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
		F	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Numeracy	3	0	0	9	14	0	0	1	0	0	0	0	0	2	5	5	1	
	5	0	0	22	19	0	0	3	2	0	1	0	0	5	2	8	1	
	7	0	0	24	23	0	0	7	1	0	1	0	1	3	1	1	2	
	9	0	0	20	15	0	0	2	1	0	1	0	1	3	1	10	10	

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

To examine jurisdictional DIF for the writing test, the expected score curves of the ten rating criteria were plotted for the eight jurisdictions in Appendix J. None of the criteria showed notable differences across jurisdictions.

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 value 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 of the paper tests or for each test path of the online tests by domain by year level 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 just as was done for paper tests.

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 with latent regression (Wu et al., 2007) for students in each of the year levels for each of reading, spelling, grammar and punctuation and numeracy. The plausible values for writing were drawn based on a unidimensional model for Years 3, 5, 7 and 9 concurrently, conditioning on year level.

In drawing the plausible values, conditioning variables were used as regressors in the model. The regression model used in 2019 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, school geolocation, school sector, 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 33. The school writing WLE mean was used in the conditioning instead of the reading mean for drawing plausible values for writing.

The categorical variables (gender, LBOTE status, Indigenous status, parental education, parental occupation, school geolocation and school sector) 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.

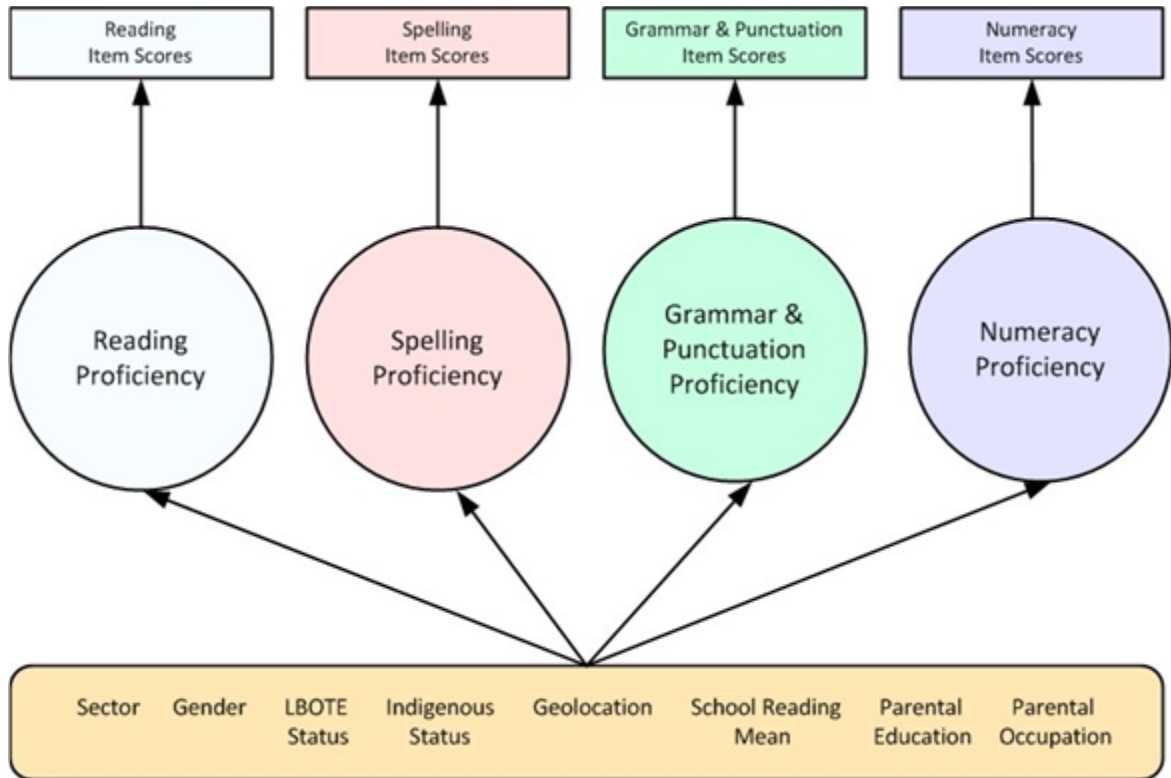


Figure 33: Conditioning variables for the multidimensional item response model with latent regression model

Chapter 7: Equating procedures

In 2019, about 50 per cent of students sat the online tests and another 50 per cent of students sat the paper tests. This chapter describes the process of equating the 2019 tests onto the NAPLAN historic scales for both the paper tests and the online tests in turn.

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

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 paper and online tests is represented schematically in terms of data matrix in Table 75. The 2019 year level NAPLAN tests were linked to each other by a set of common items between adjacent year levels. The 2019 tests were 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 is based on a common student equating design.

Table 75: Equating design for both assessment modes

	NAPLAN test items (paper or online)						
Students	Y3	Y3 & 5	Y5	Y5&7	Y7	Y7&9	Y9
Y3 population	█						
Y5 population		█					
Y7 population				█			
Y9 population						█	
	Equating test items (paper)						
Students	Y3		Y5		Y7		Y9
Y3 equating sample	█						
Y5 equating sample			█				
Y7 equating sample					█		
Y9 equating sample							█

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 2019 scale for each domain needs to be horizontally equated to the historic NAPLAN reporting scale. The horizontal equating of the NAPLAN 2019 scale to the NAPLAN historical scale was achieved by a common person equating design. The secure, paper-based equating tests used for horizontal equating in 2009–2018 were used for common-person equating again in 2019 for selected students of both paper and online assessment modes (the equating samples), noting that some of the secure forms were modified or updated. Students from Years 3, 5, 7 and 9 in the equating sample were administered the secure equating tests at their year level two weeks prior to the NAPLAN

2019 tests. The response data on the equating test were used to equate the 2019 tests onto the existing NAPLAN reporting scales.

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 test. However, vertical link items were used in all subsequent years to check and, if needed, 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 horizontal and vertical link items were reviewed. Link items that differed too much in relative difficulty between the two tests were broken and therefore excluded when calculating horizontal or vertical equating shifts. Once these shifts were calculated, HVR shifts were estimated and used to equate the NAPLAN 2019 results onto the 2008 historical scale.

Horizontal equating shifts

Calculation of horizontal equating shifts for reading, spelling, grammar and punctuation, and numeracy involved a common-person equating method. The common-person equating was achieved through the equating sample. The equating was carried out using secure equating tests that were administered with the NAPLAN 2019 online and paper tests for reading, spelling, grammar and punctuation and numeracy. Each student in the equating sample completed an equating test two weeks prior to the NAPLAN 2019 paper tests. Table 75 also shows the horizontal equating design for each of reading, spelling, grammar and punctuation and numeracy at each year level.

In 2009, the equating test for each domain at each year level had been equated to the historic NAPLAN scale, which was established in 2008. The first step in 2019 was to place the NAPLAN 2019 test and the equating test on a newly calibrated scale using 2019 response data, for each domain at each year level by test mode, using the common person equating data. This was achieved through a concurrent calibration, separately by domain and year level, of the data from the NAPLAN 2019 tests and the corresponding equating tests, with the NAPLAN 2019 tests anchored to its 2019 delta-centred item parameters. This step provided a set of item parameters on the newly calibrated NAPLAN 2019 scales. The set of 2019 equating test item difficulties were then compared with the item difficulties of the 2009 equating tests that were on the historic NAPLAN scale. Items were considered to be broken as links if they functioned poorly (based on Mean Square indices and ICCs) and/or if their relative locations differed by an absolute value of greater than 0.3, compared to their relative locations on the original 2009 scale for the relevant secure form. A slightly different equating routine was used for grammar and punctuation online test equating. Each year level grammar and punctuation online test consisted of four generic testlets: C, E1 & E2, E3 and F. Because of the lack of link items between testlets, grammar and punctuation online tests were equated by generic testlets in two steps: the equating tests based on the 2019 equating sample were placed the NAPLAN scale by year level, and each generic testlet was placed on the 2019 equating test.

Figure 34 to Figure 77 show the comparisons of the 2009 item parameter estimates with the 2019 item parameter estimates, either from the paper test equating test or the online test equating test, for each of the 32 equating tests. Each figure shows a pair of scatterplots of the linked items, before and after breaking unsatisfactory link items. For link items that did not change in relative item difficulty, the bivariate points were on the identity line, which is shown in each figure for reference as a thick red line. A second, thinner grey line is the

linear line of best fit through the dots in each scatterplot. Equating items that were modified over time were omitted from the horizontal linking process.

Horizontal link item review of paper tests

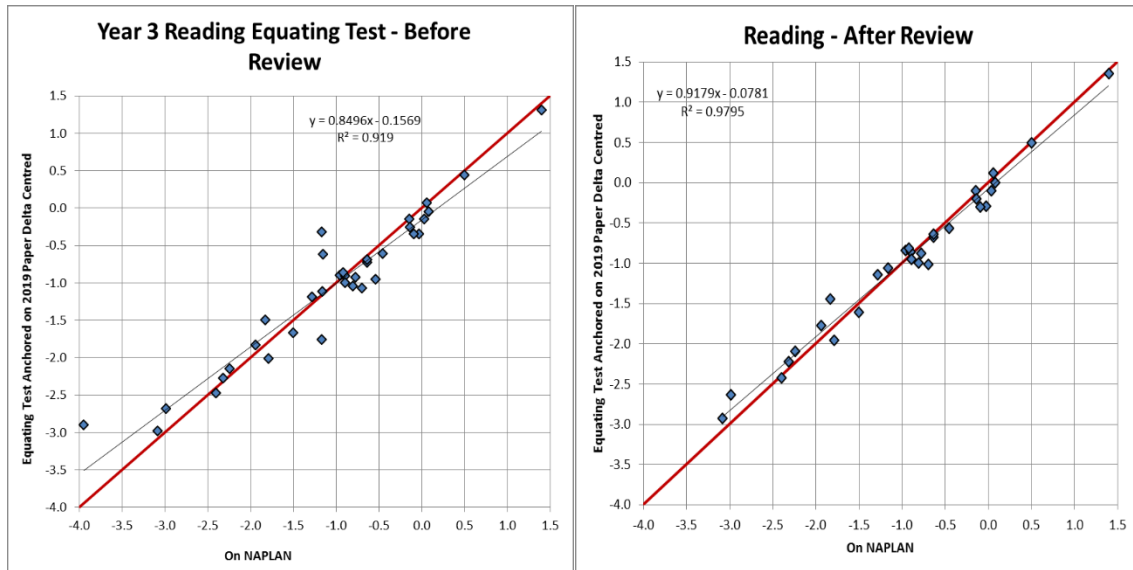


Figure 34: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 3 paper students

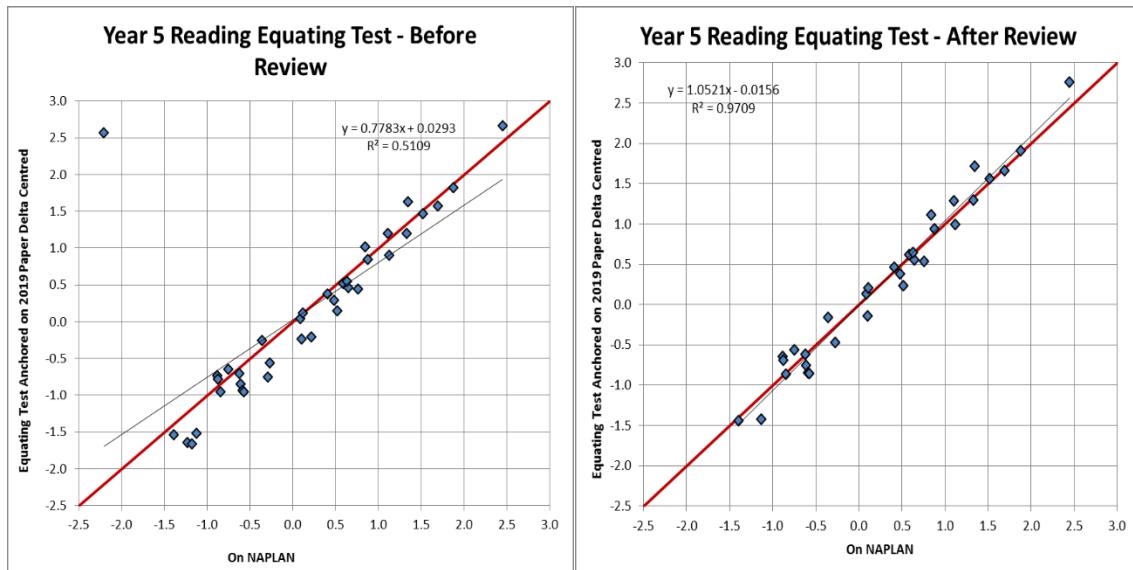


Figure 35: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 5 paper students)

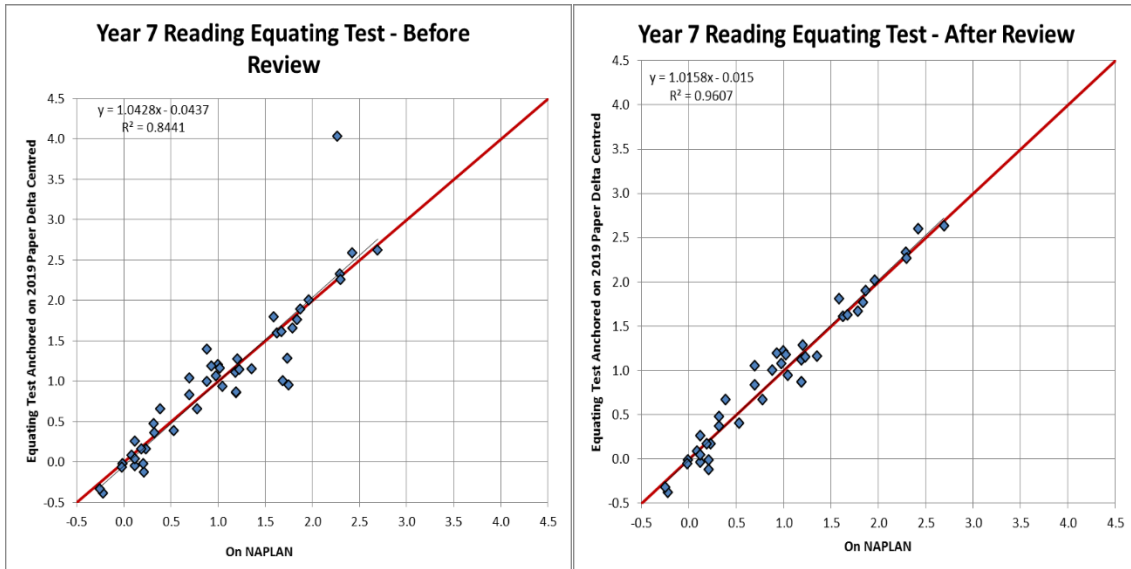


Figure 36: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 7 paper students

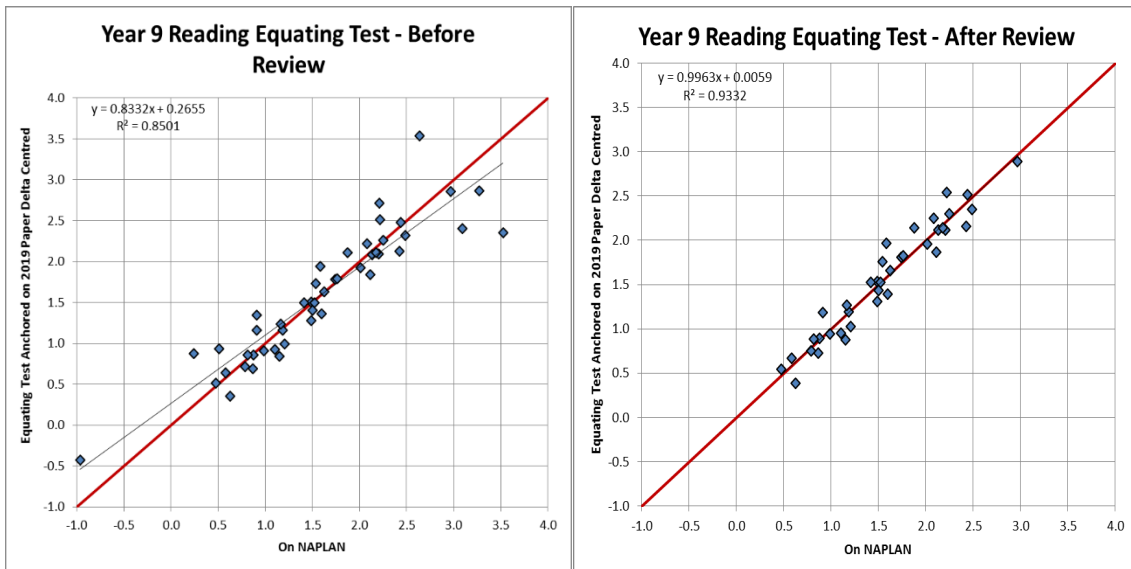


Figure 37: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 9 paper students

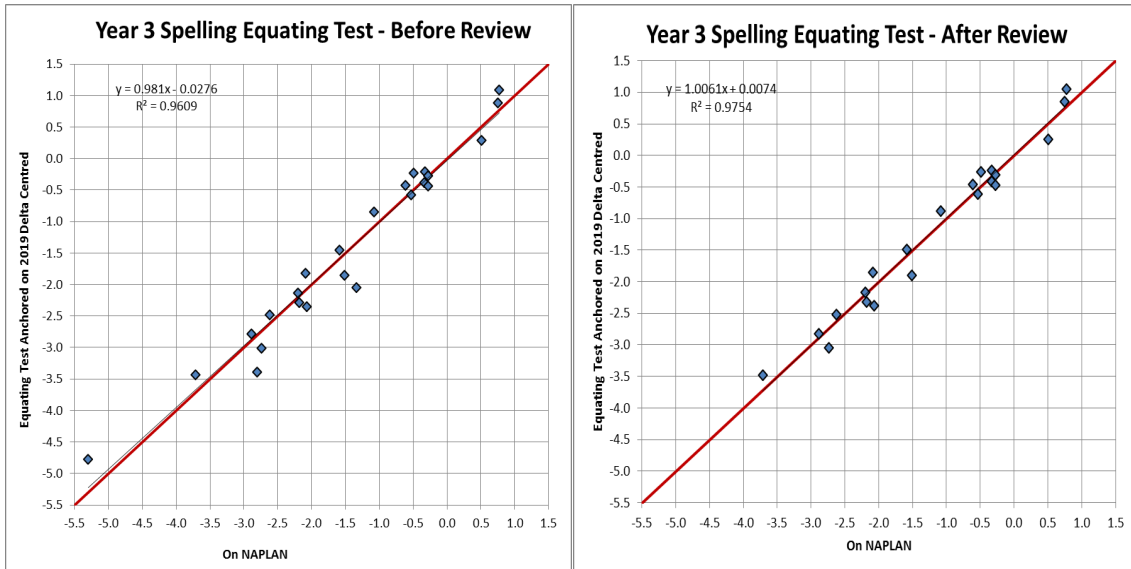


Figure 38 Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 3 paper students

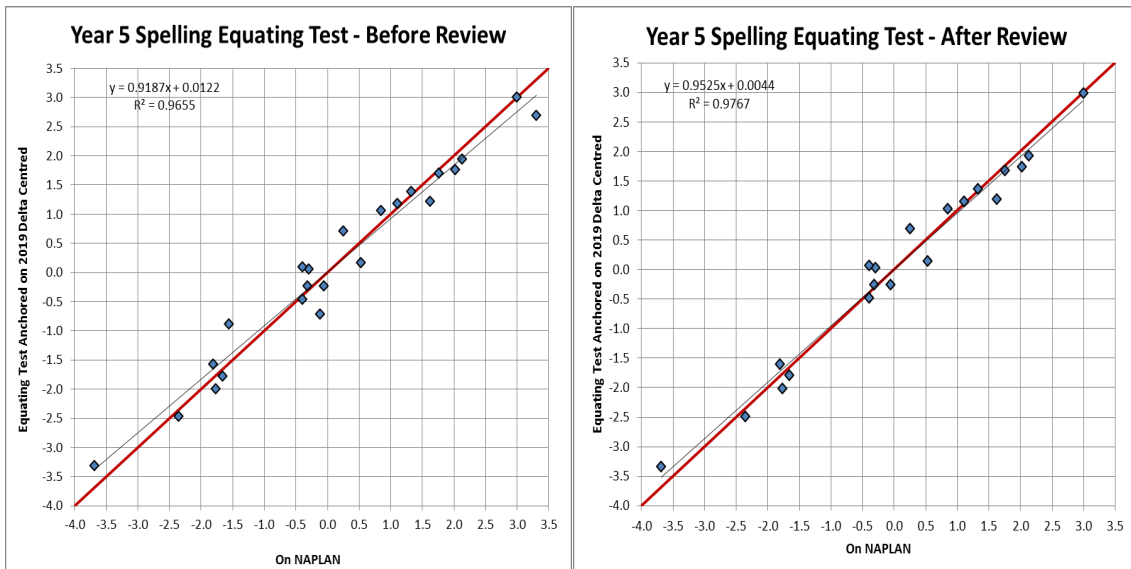


Figure 39: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 5 paper students

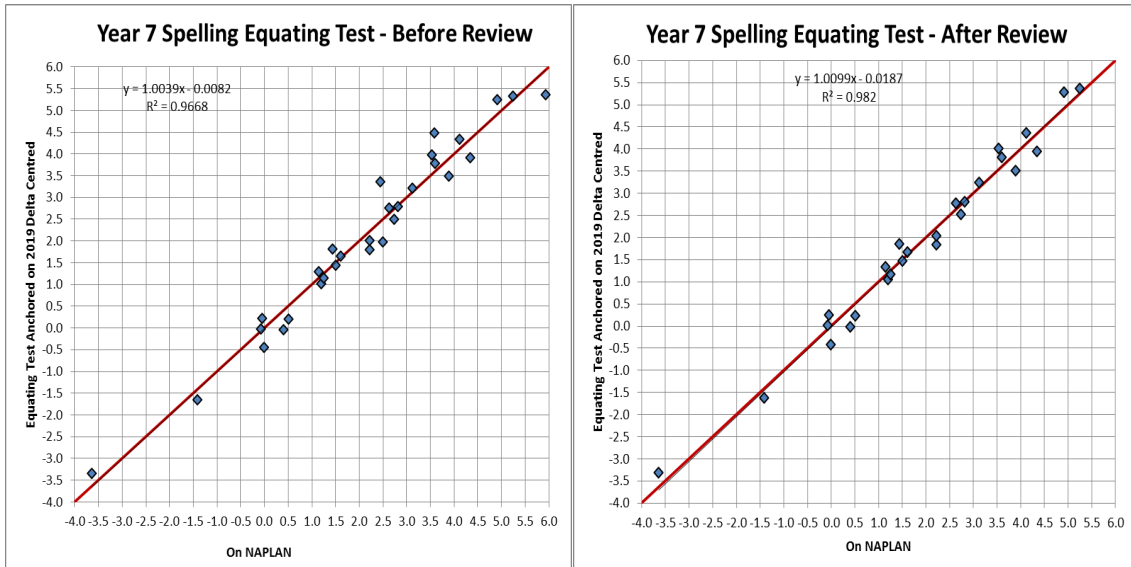


Figure 40: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 7 paper students

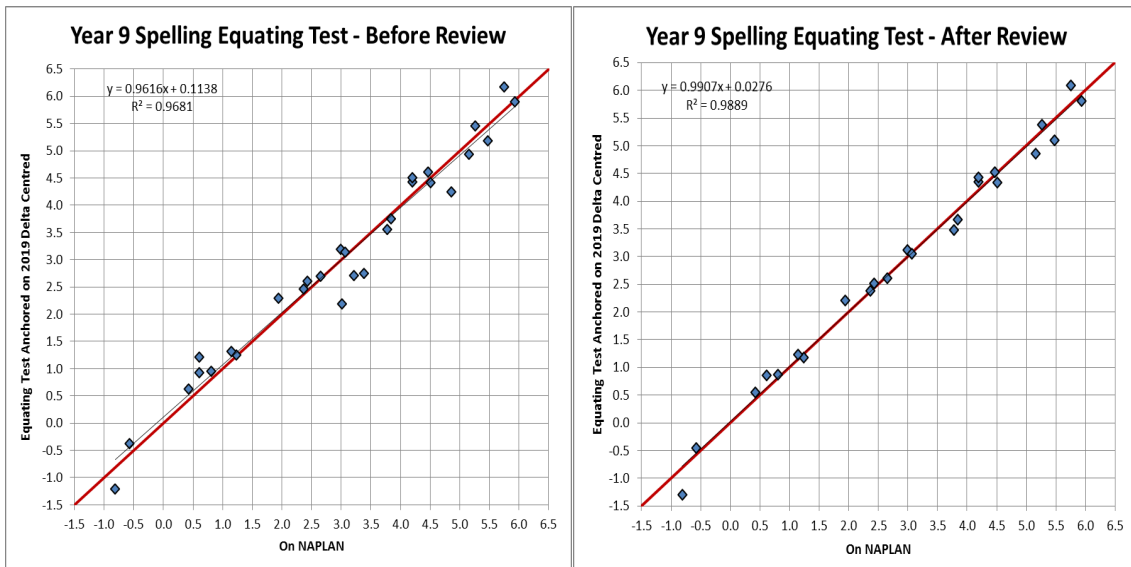


Figure 41: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 9 paper students

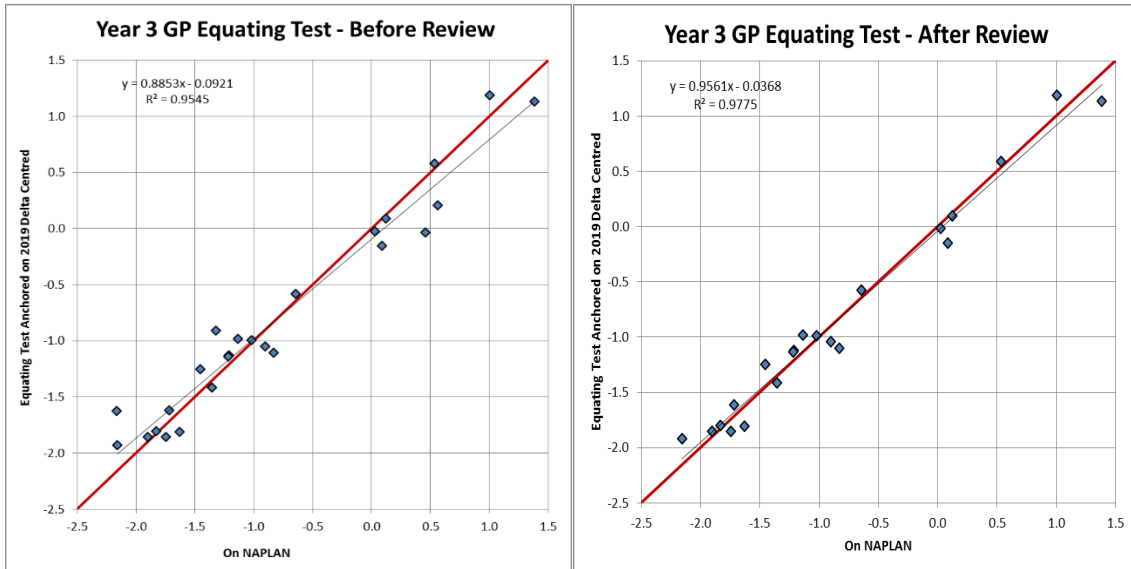


Figure 42: Scatterplot of grammar and punctuation, horizontal equating items between 2019 and 2009 for Year 3 paper students

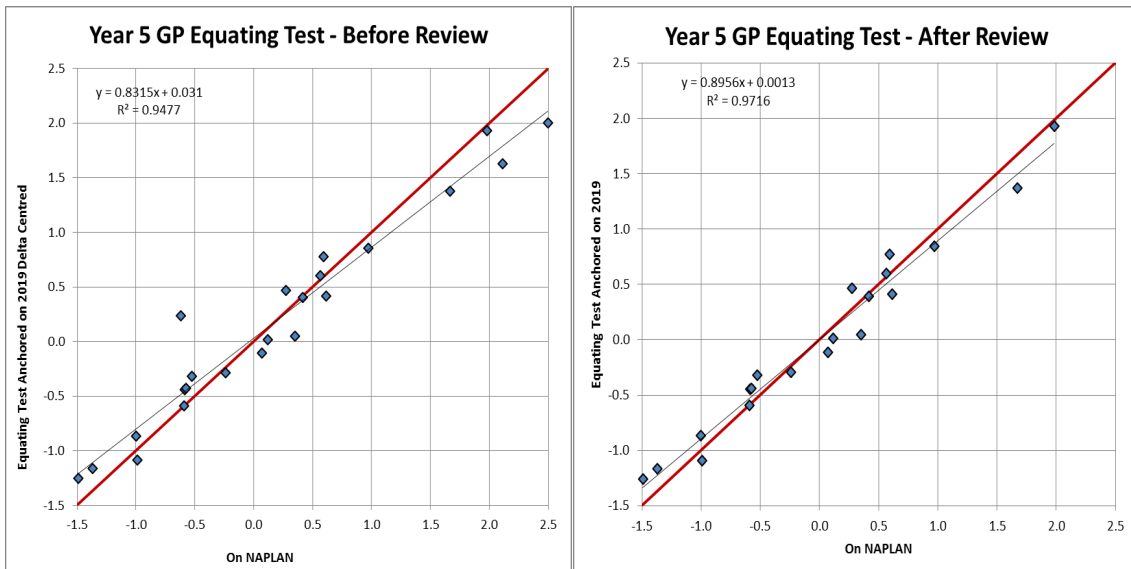


Figure 43: Scatterplot of grammar and punctuation, horizontal equating items between 2019 and 2009 for Year 5 paper students

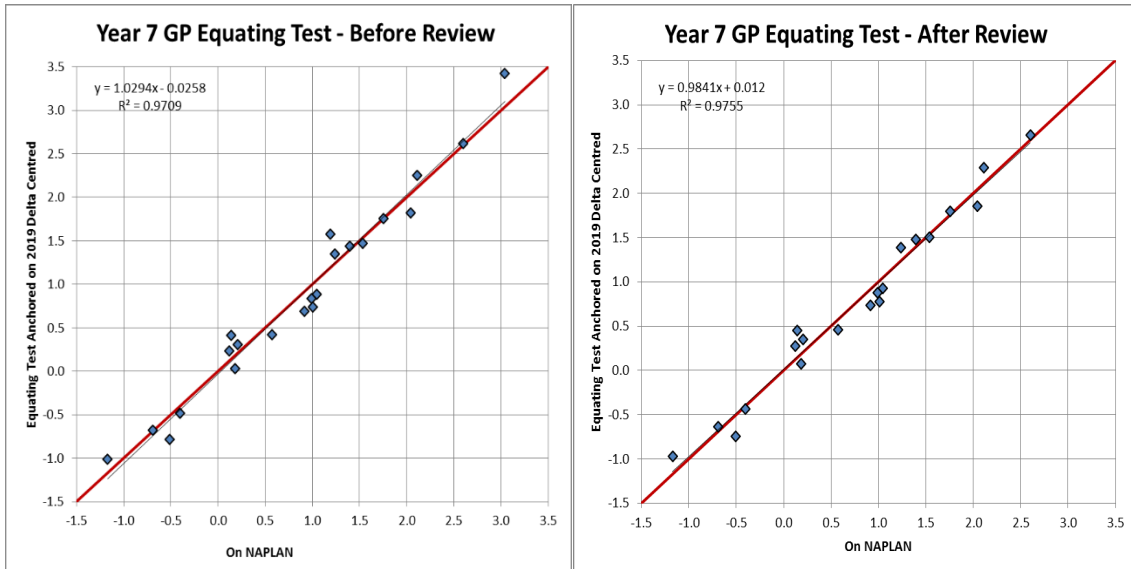


Figure 44: Scatterplot of grammar and punctuation, horizontal equating items between 2019 and 2009 for Year 7 paper students

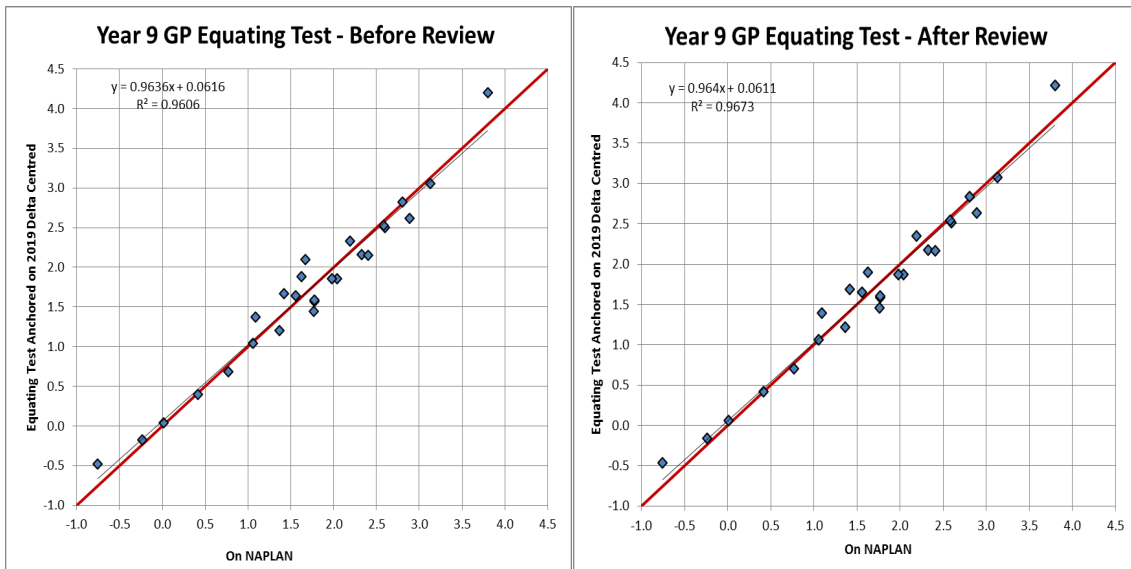


Figure 45: Scatterplot of grammar and punctuation, horizontal equating items between 2019 and 2009 for Year 9 paper students

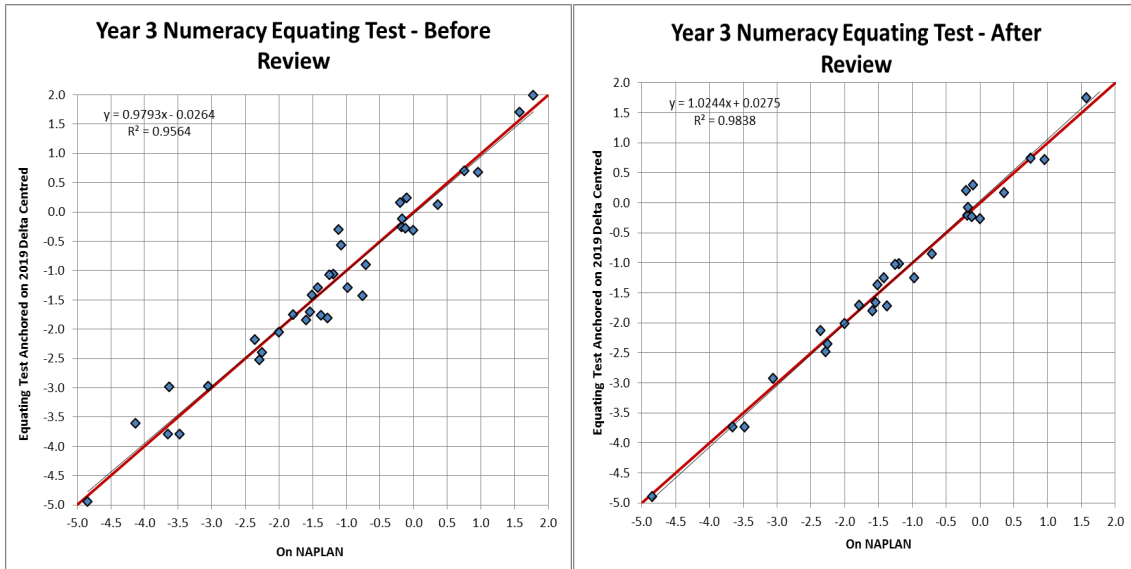


Figure 46: Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 3 paper students

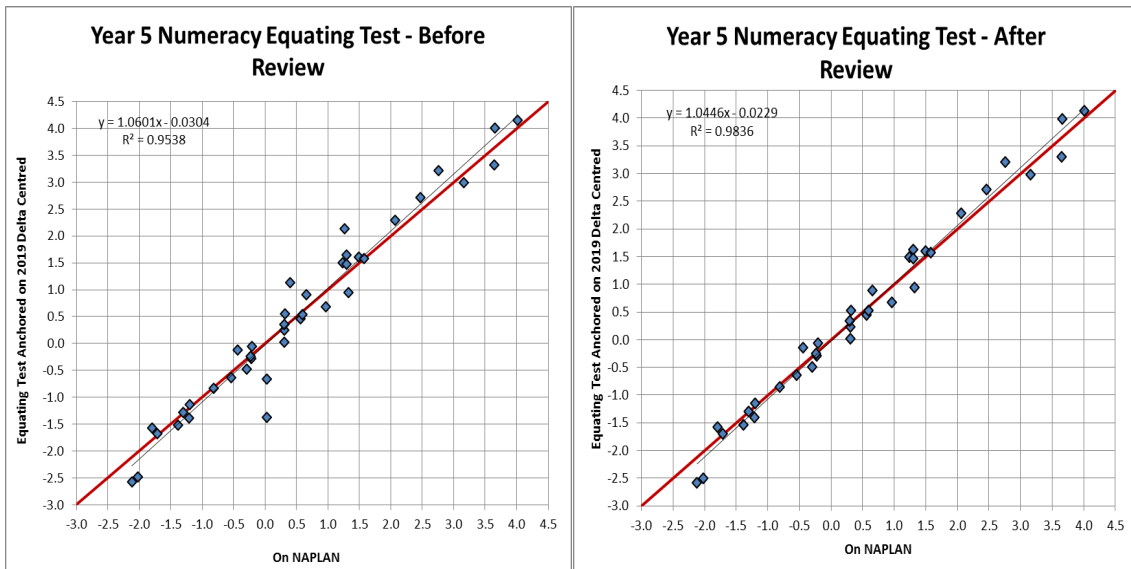


Figure 47: Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 5 paper students

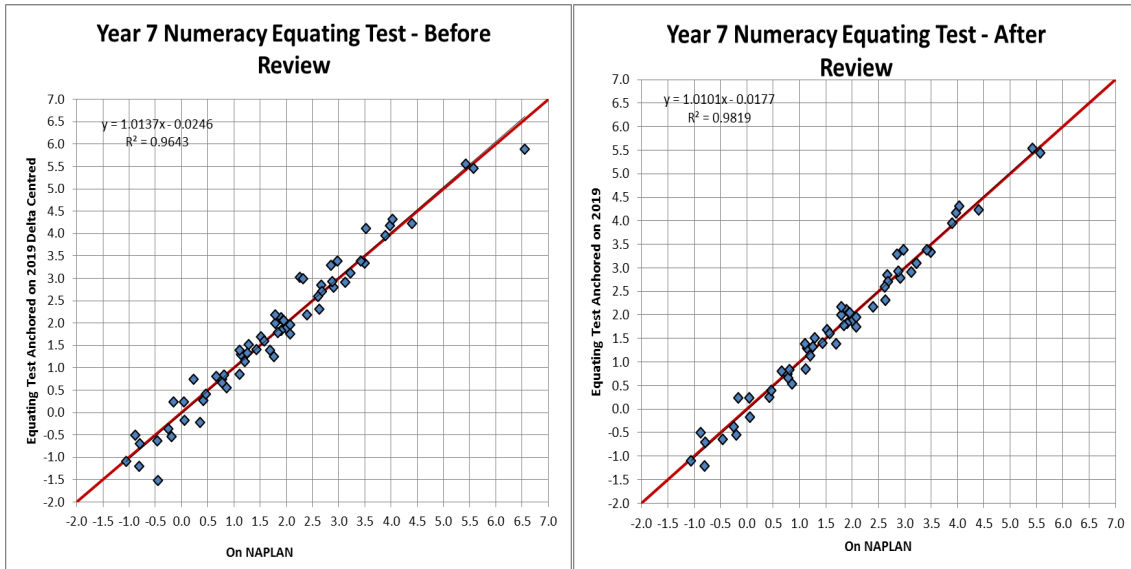


Figure 48: Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 7 paper students

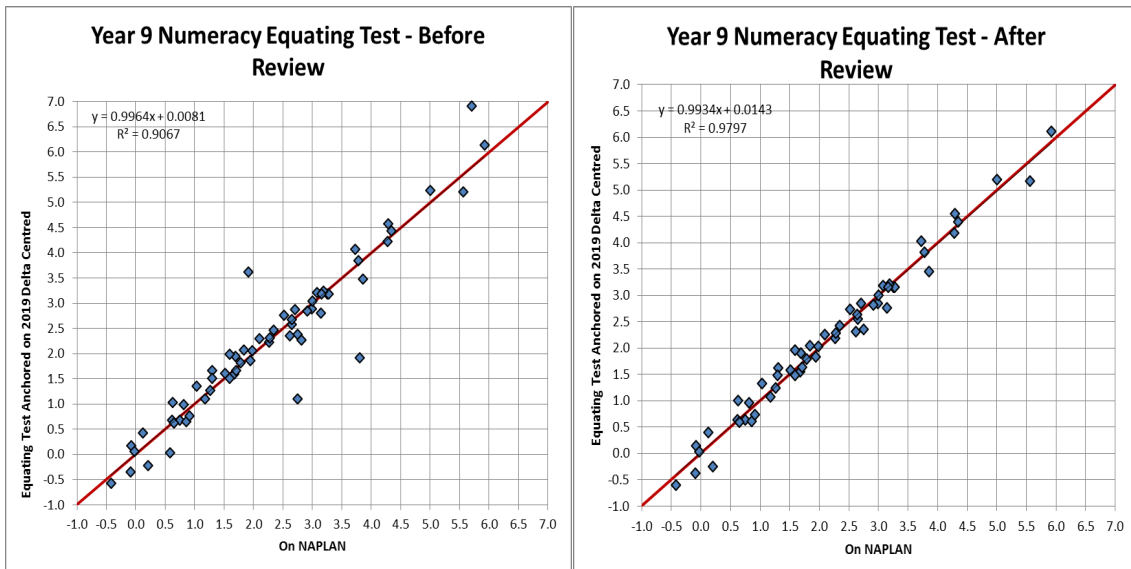


Figure 49: Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 9 paper students

Horizontal link item review of online tests

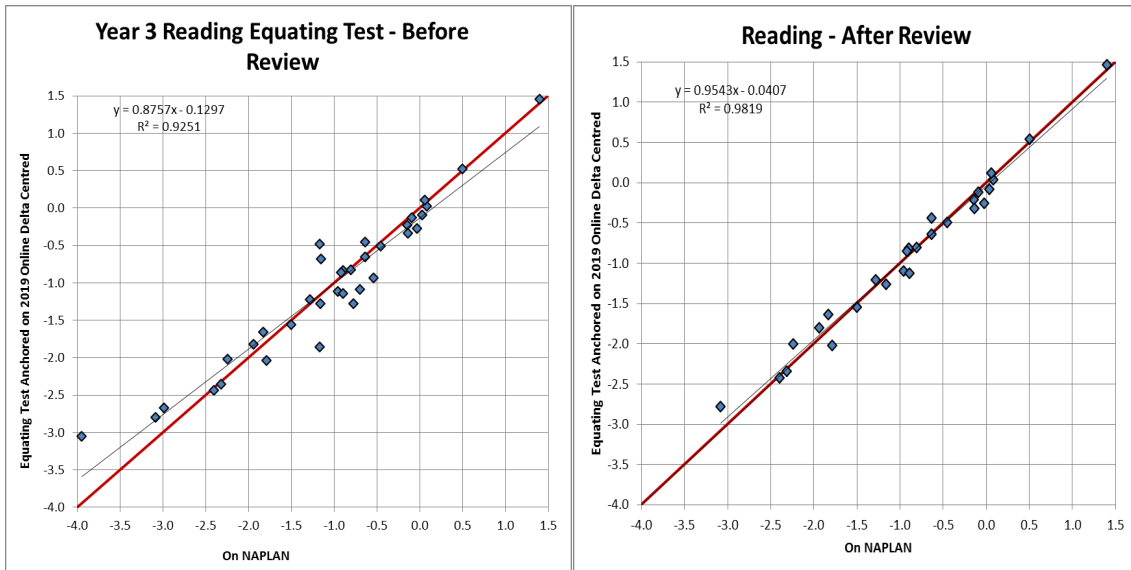


Figure 50: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 3 online students

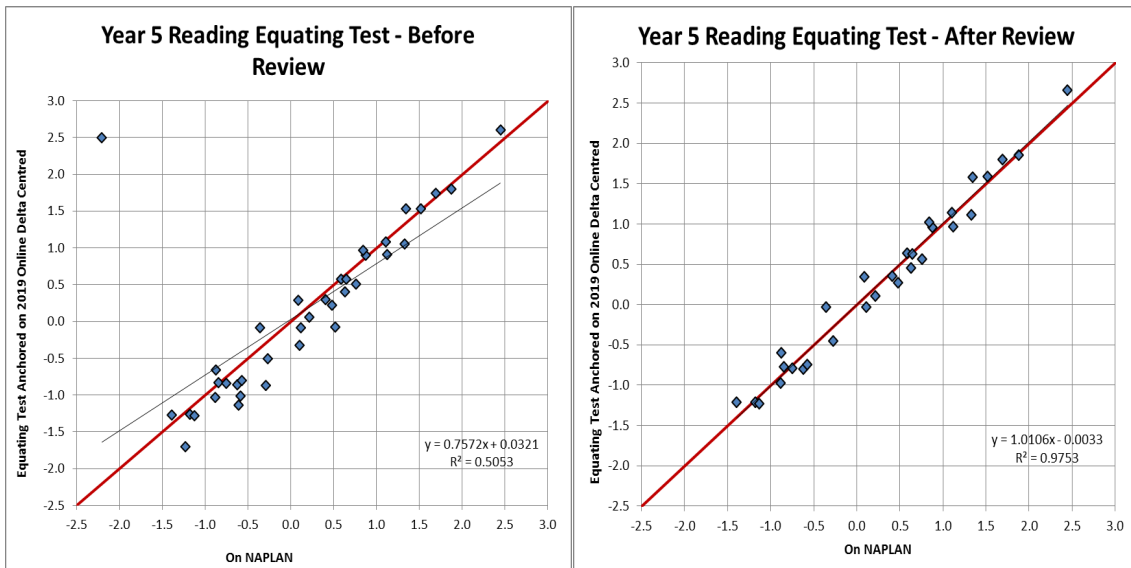


Figure 51: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 5 online students

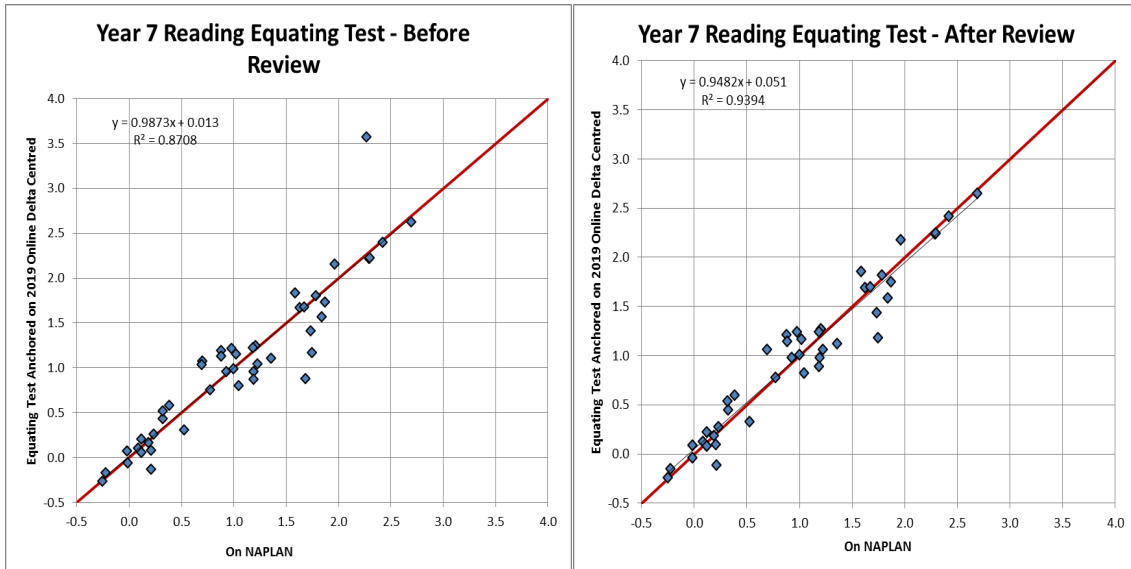


Figure 52: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 7 online students

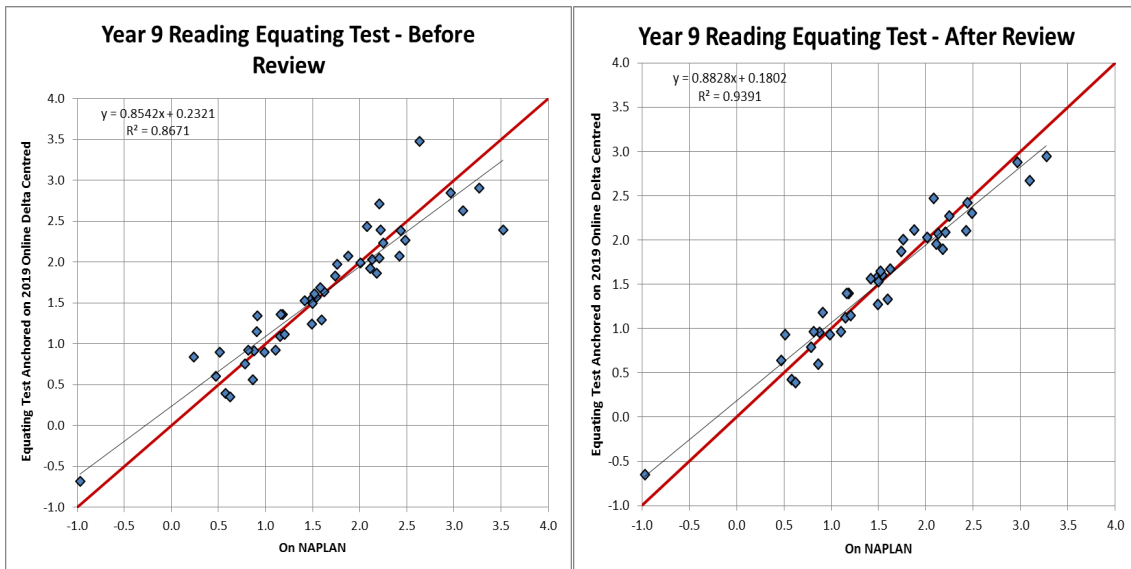


Figure 53: Scatterplot of reading, horizontal equating items between 2019 and 2009 for Year 9 online students

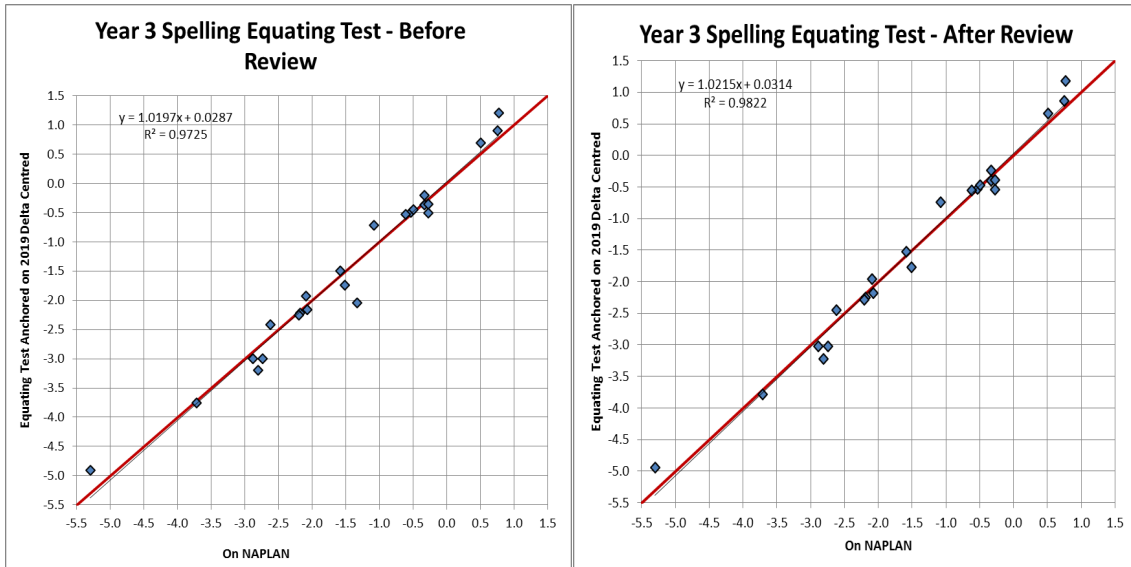


Figure 54: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 3 online students

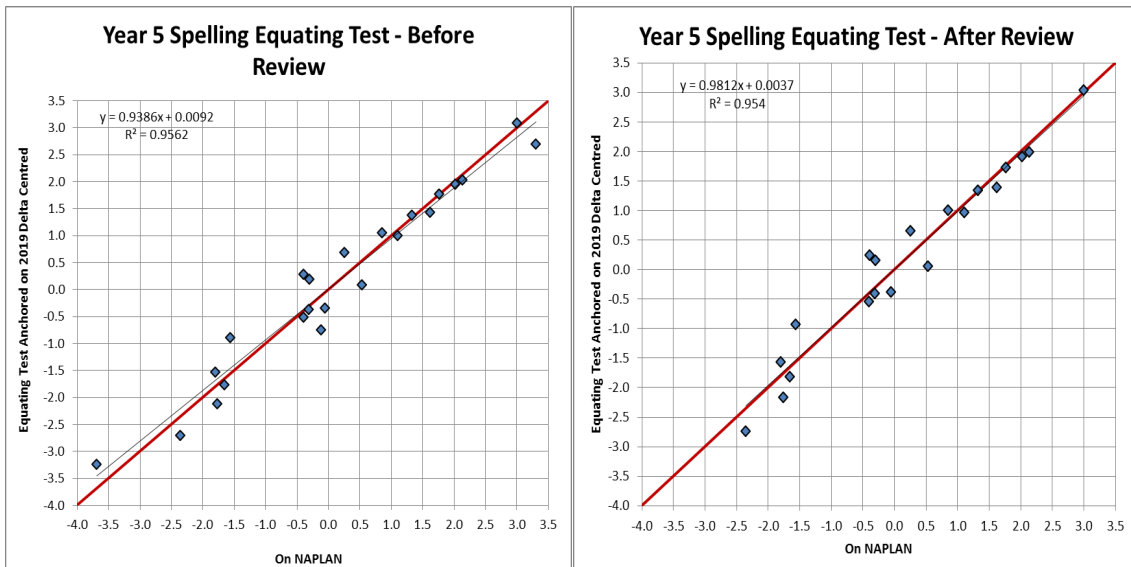


Figure 55: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 5 online students

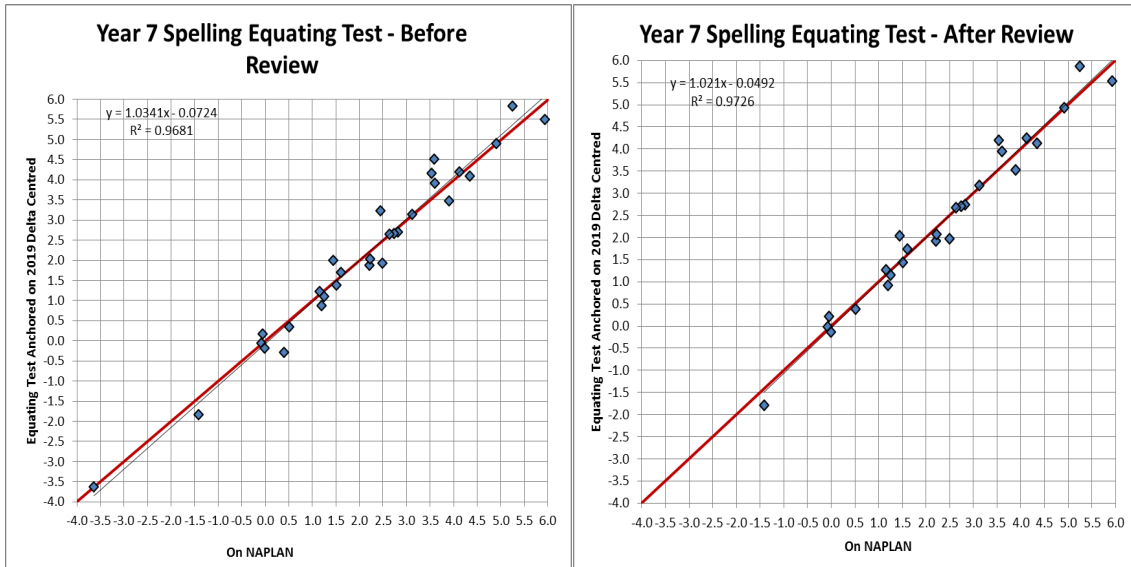


Figure 56: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 7 online students

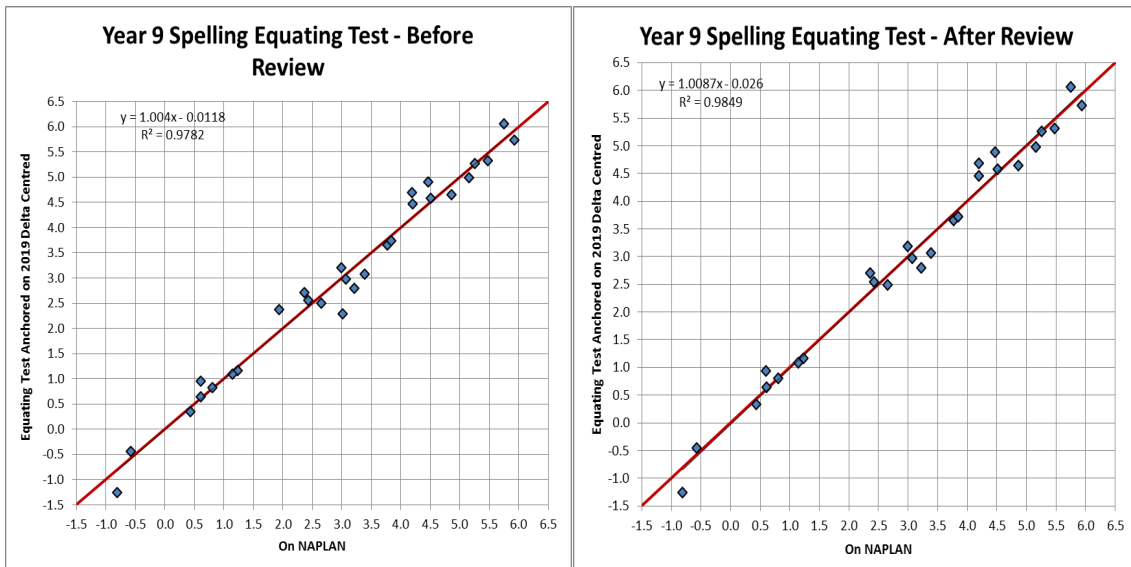


Figure 57: Scatterplot of spelling, horizontal equating items between 2019 and 2009 for Year 9 online students

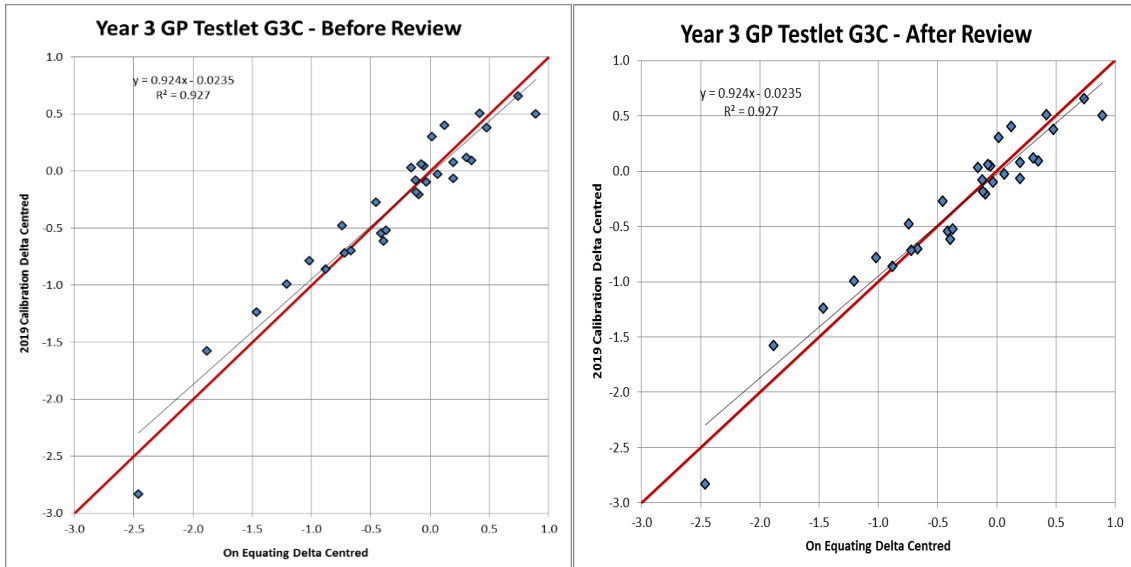


Figure 58: Scatterplot of GP testlet C, horizontal equating items between 2019 and 2009 for Year 3 online students

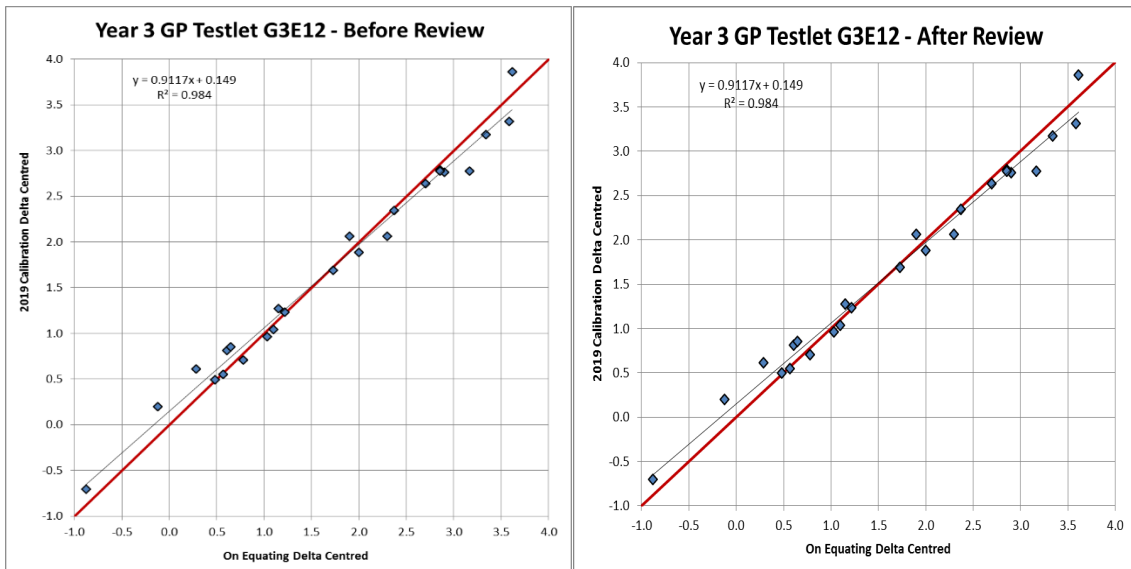


Figure 59: Scatterplot of GP testlet E1&E2, horizontal equating items between 2019 and 2009 for Year 3 online students

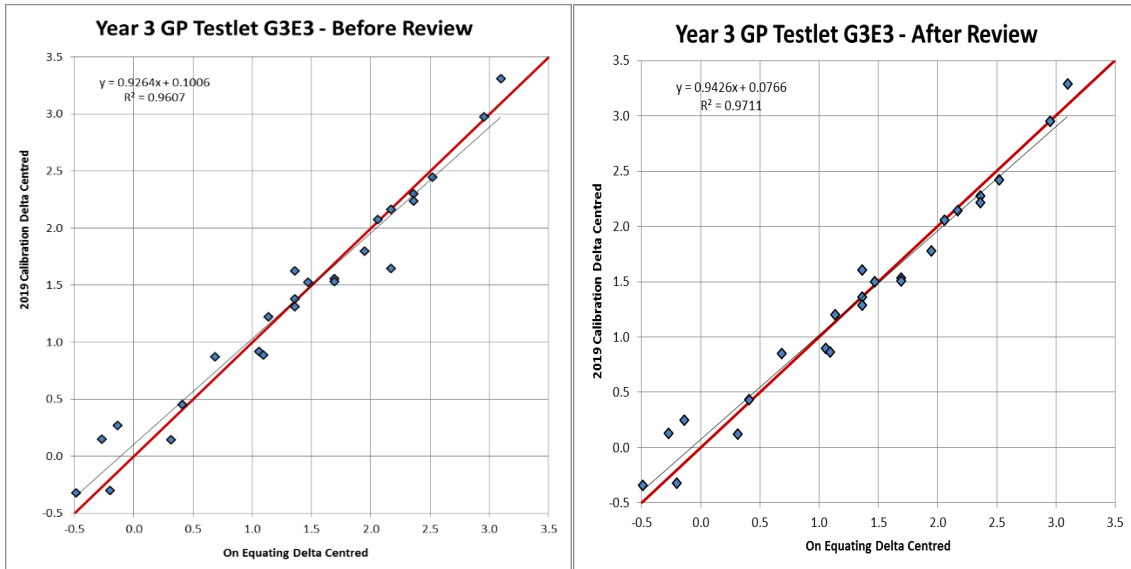


Figure 60. Scatterplot of GP testlet E3, horizontal equating items between 2019 and 2009 for Year 3 online students

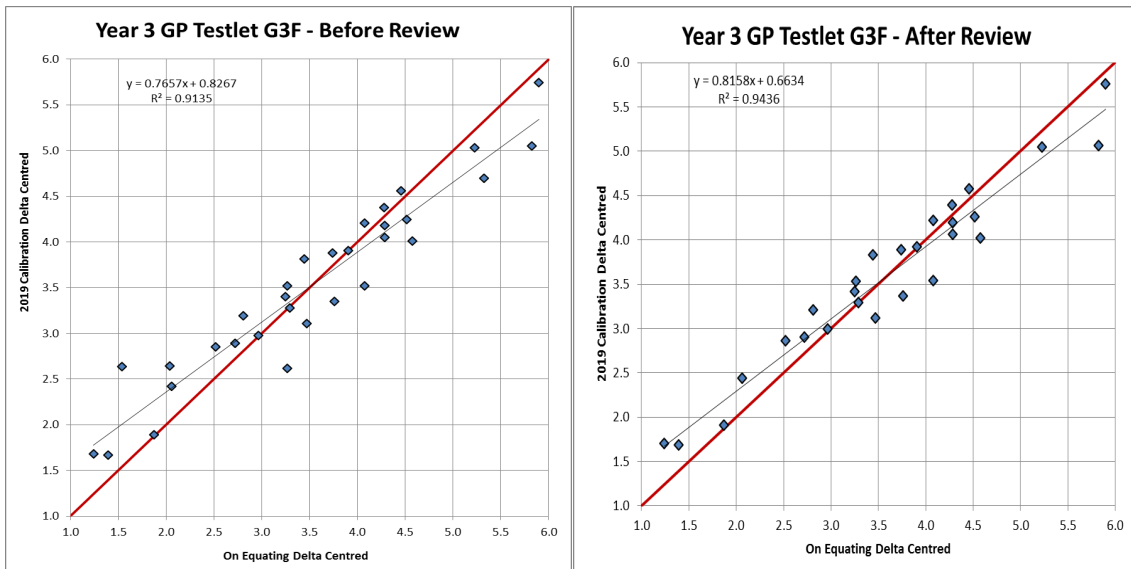


Figure 61. Scatterplot of GP testlet F, horizontal equating items between 2019 and 2009 for Year 3 online students

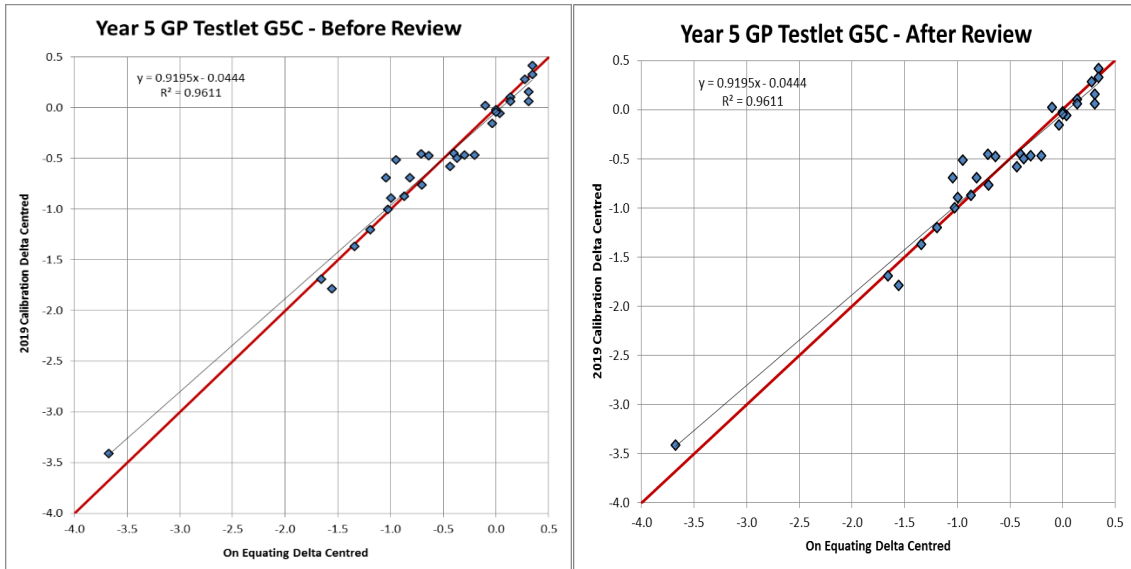


Figure 62. Scatterplot of GP testlet C, horizontal equating items between 2019 and 2009 for Year 5 online students

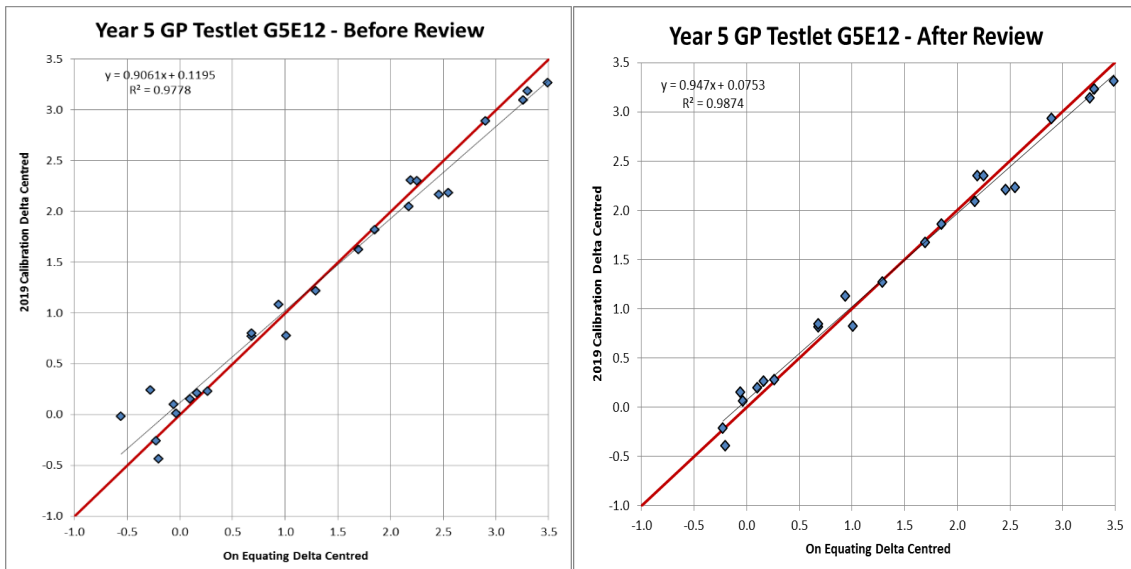


Figure 63. Scatterplot of GP testlet E1&E2, horizontal equating items between 2019 and 2009 for Year 5 online students

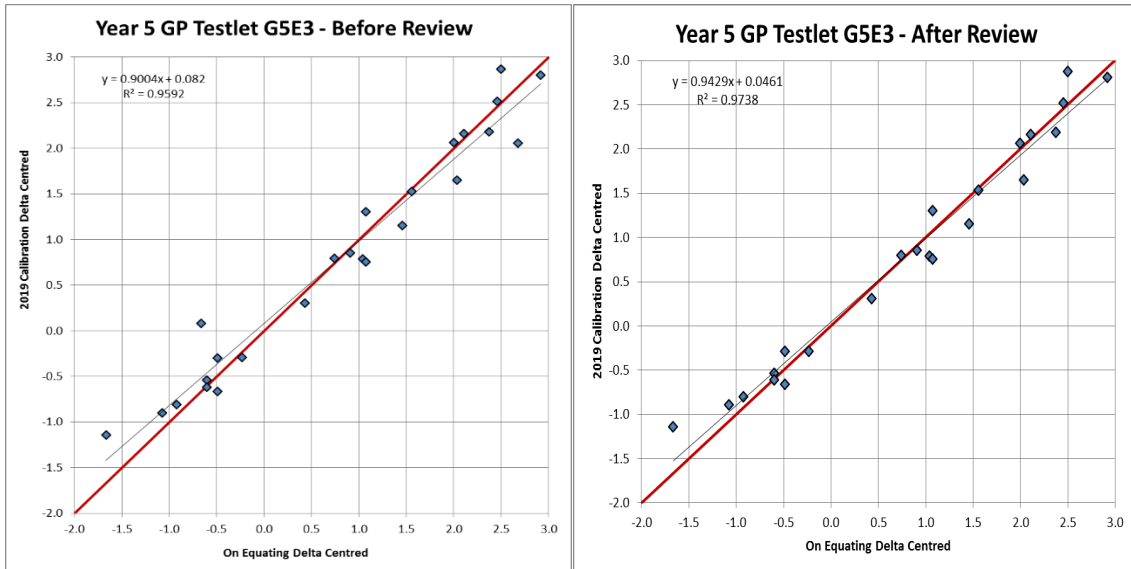


Figure 64. Scatterplot of GP testlet E3, horizontal equating items between 2019 and 2009 for Year 5 online students

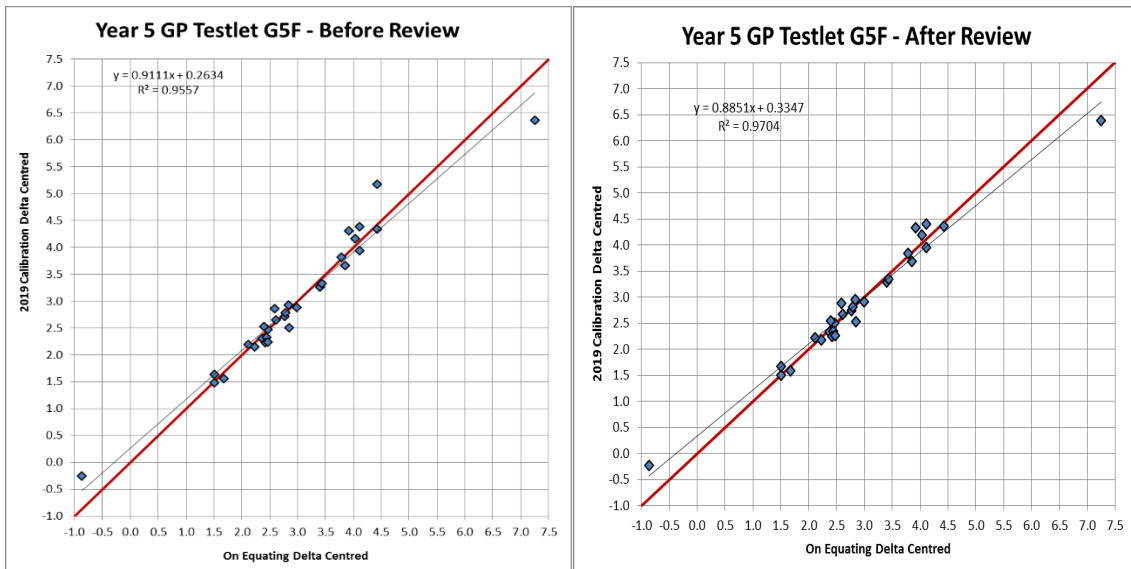


Figure 65 Scatterplot of GP testlet F, horizontal equating items between 2019 and 2009 for Year 5 online students

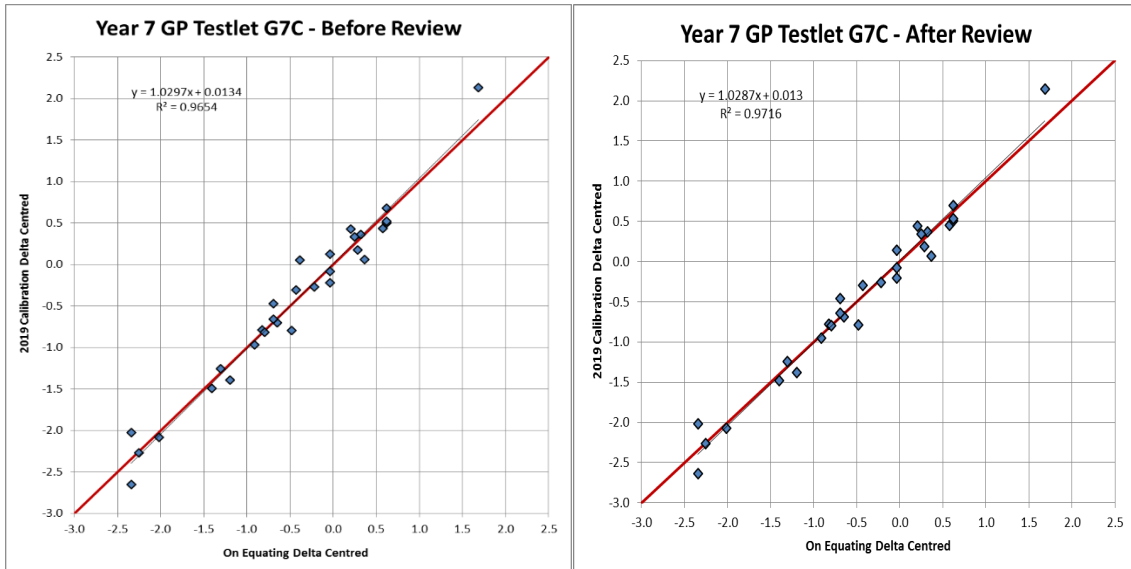


Figure 66. Scatterplot of GP testlet C, horizontal equating items between 2019 and 2009 for Year 7 online students

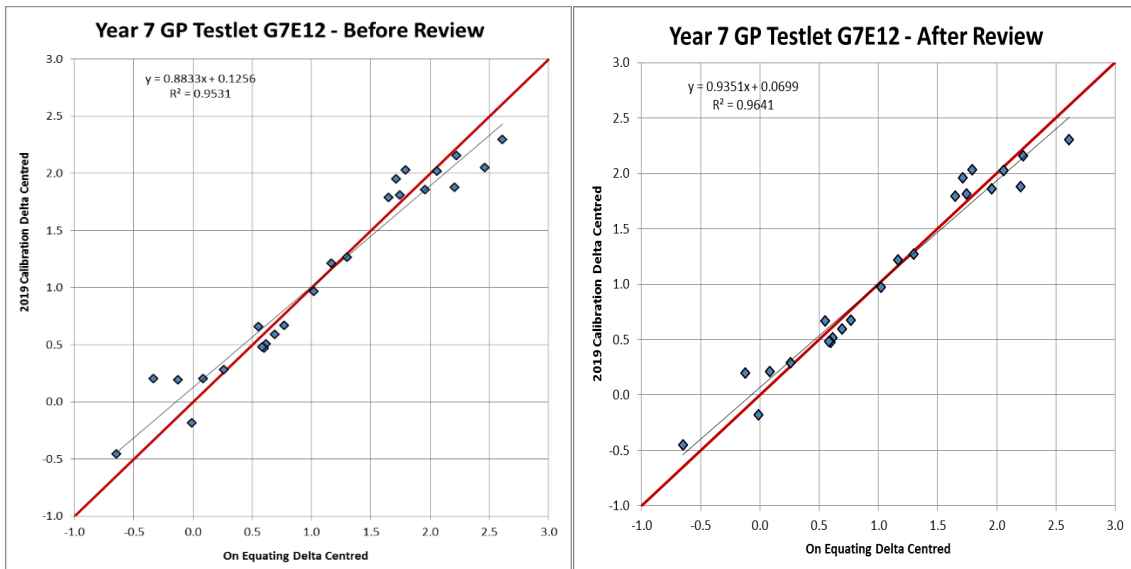


Figure 67. Scatterplot of GP testlet E1&E2, horizontal equating items between 2019 and 2009 for Year 7 online students

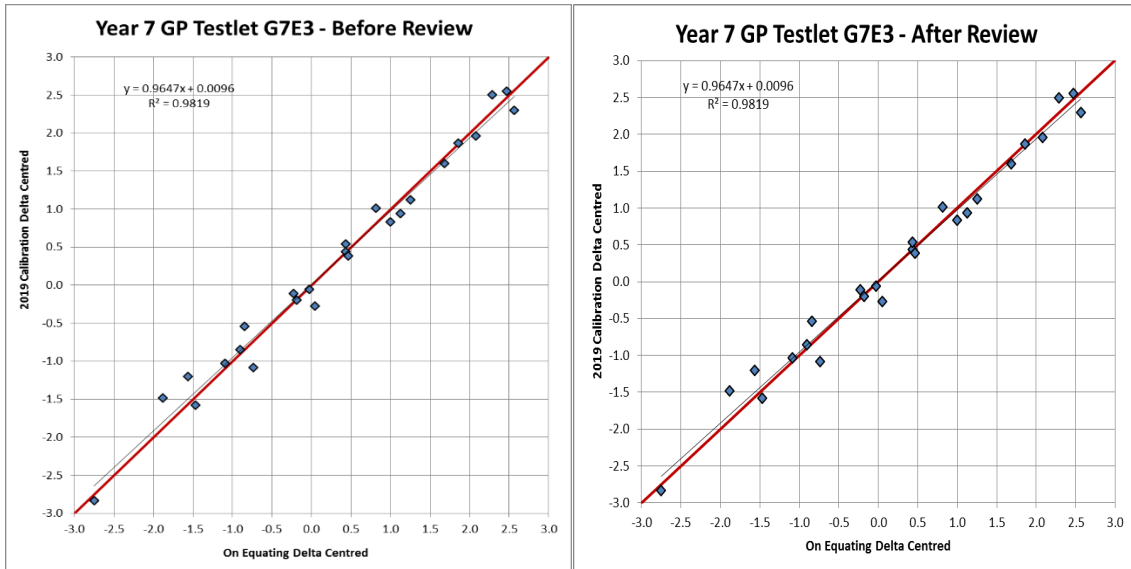


Figure 68. Scatterplot of GP testlet E3, horizontal equating items between 2019 and 2009 for Year 7 online students

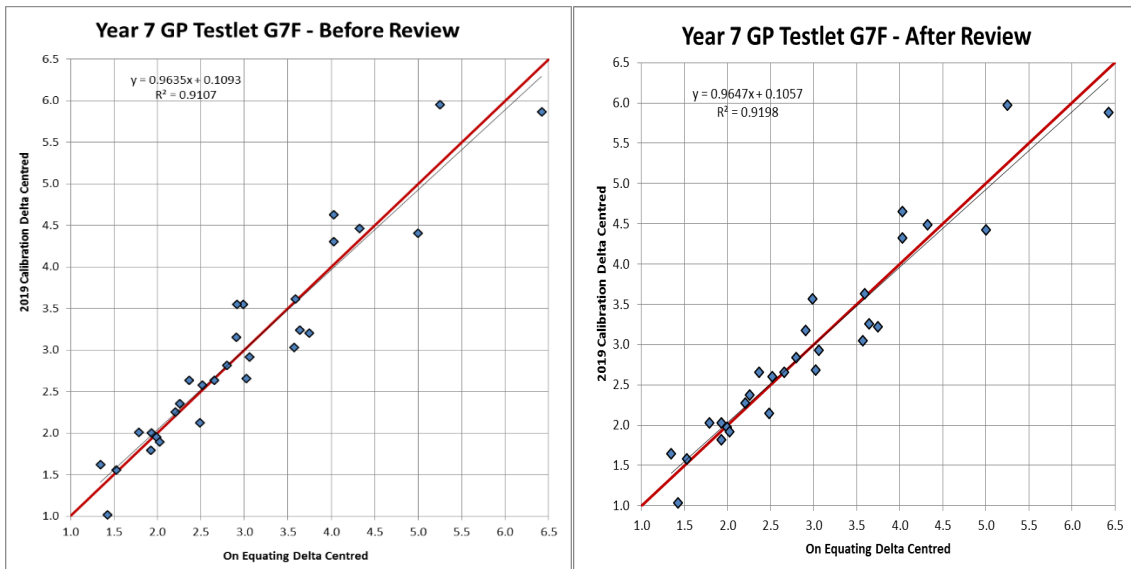


Figure 69. Scatterplot of GP testlet F, horizontal equating items between 2019 and 2009 for Year 7 online students

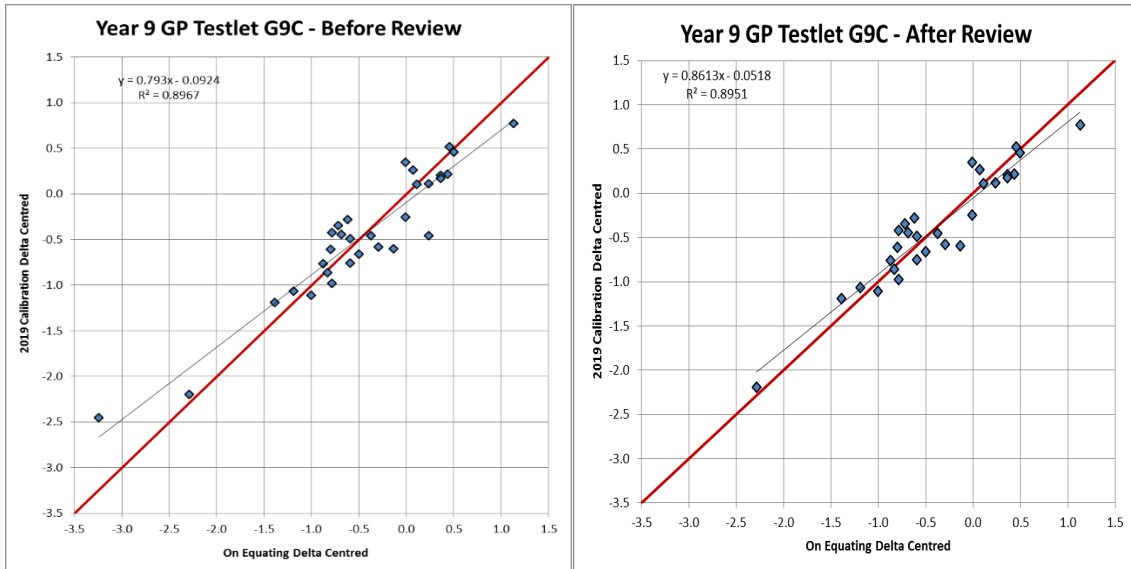


Figure 70. Scatterplot of GP testlet C, horizontal equating items between 2019 and 2009 for Year 9 online students

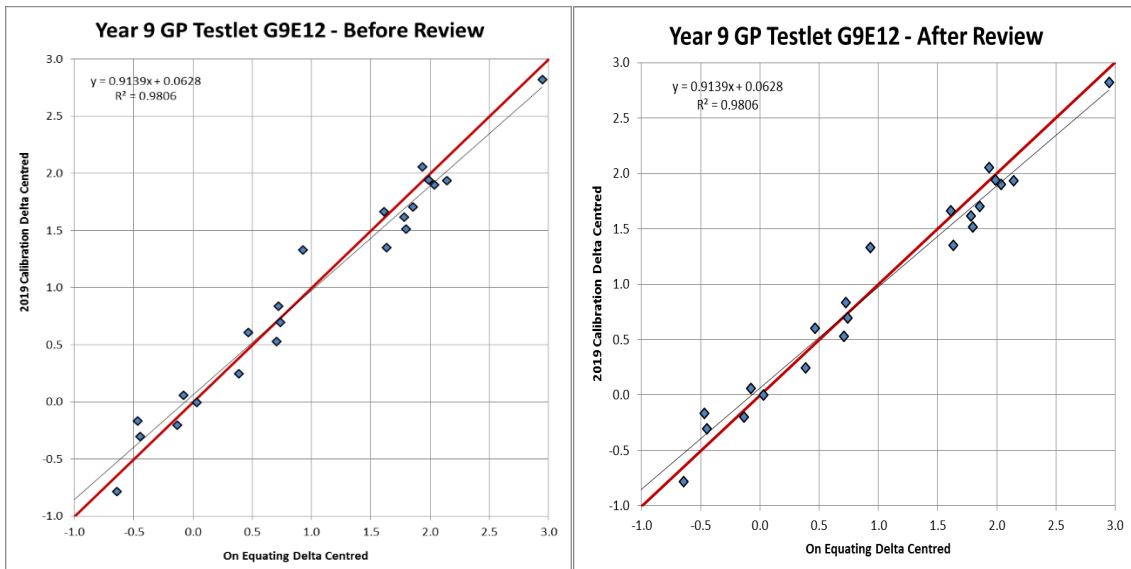


Figure 71. Scatterplot of GP testlet E1&E2, horizontal equating items between 2019 and 2009 for Year 9 online students

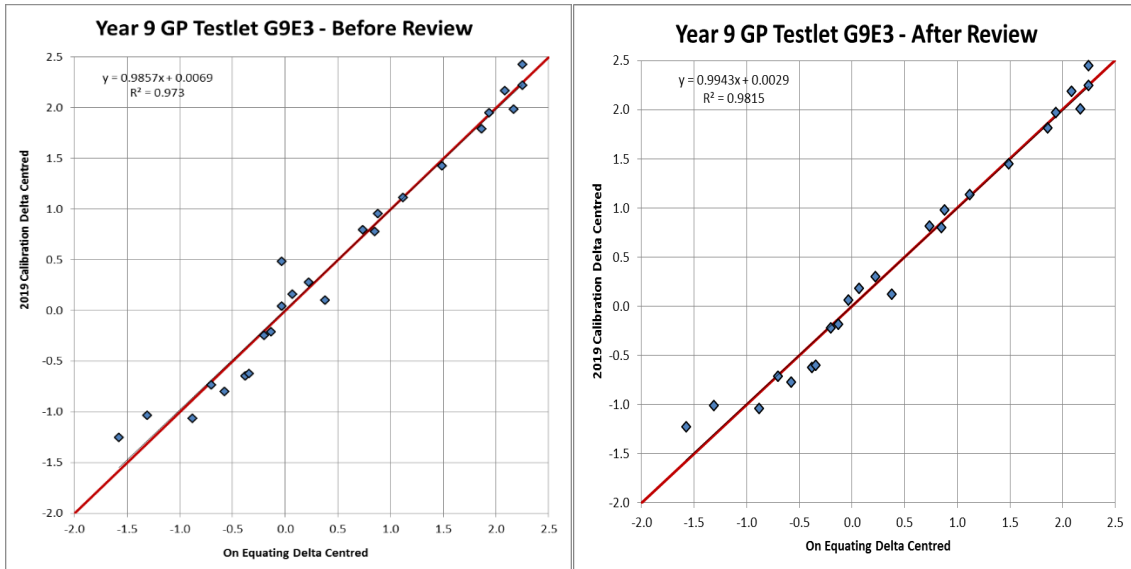


Figure 72. Scatterplot of GP testlet E3, horizontal equating items between 2019 and 2009 for Year 9 online students

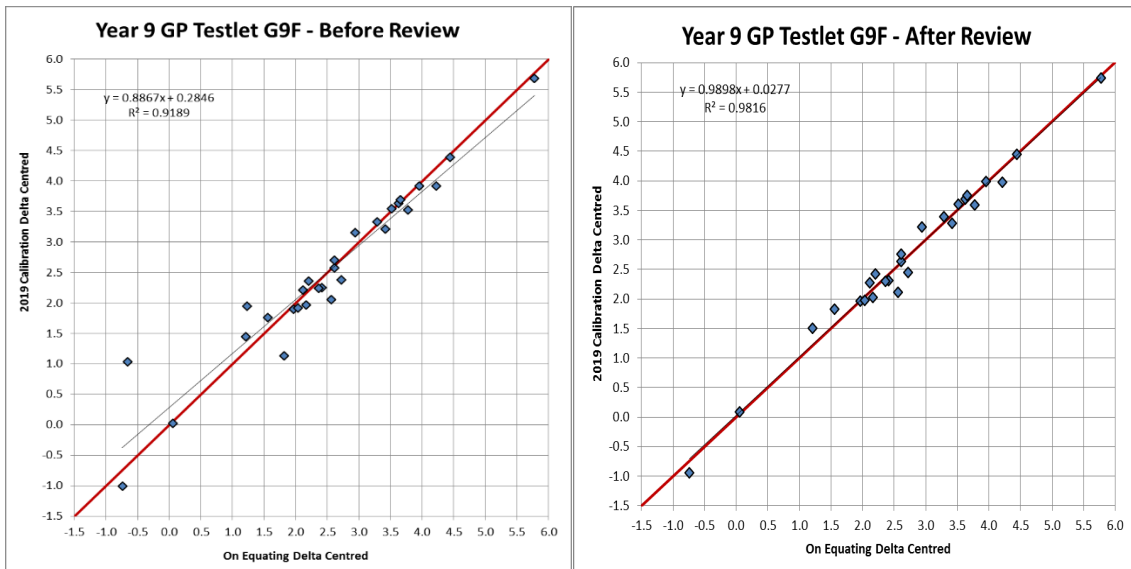


Figure 73. Scatterplot of GP testlet F, horizontal equating items between 2019 and 2009 for Year 9 online students

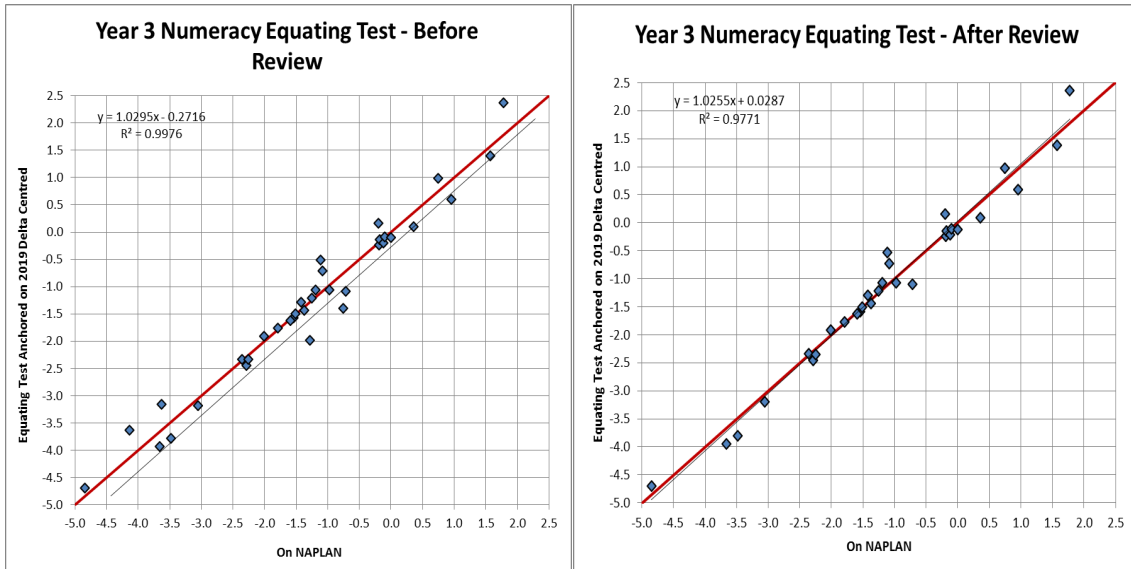


Figure 74. Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 3 online students

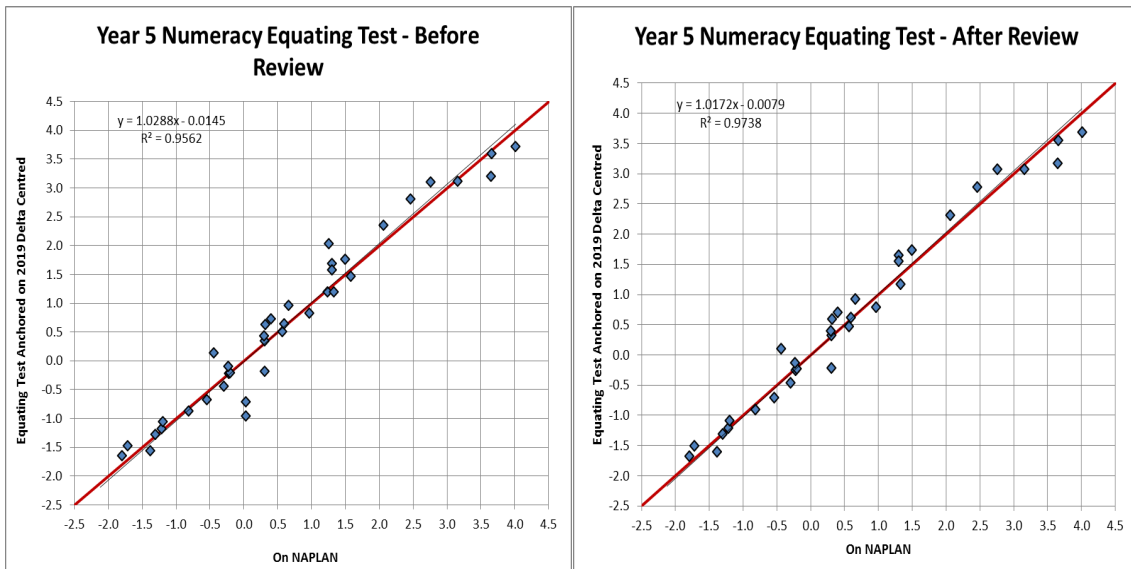


Figure 75. Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 5 online students

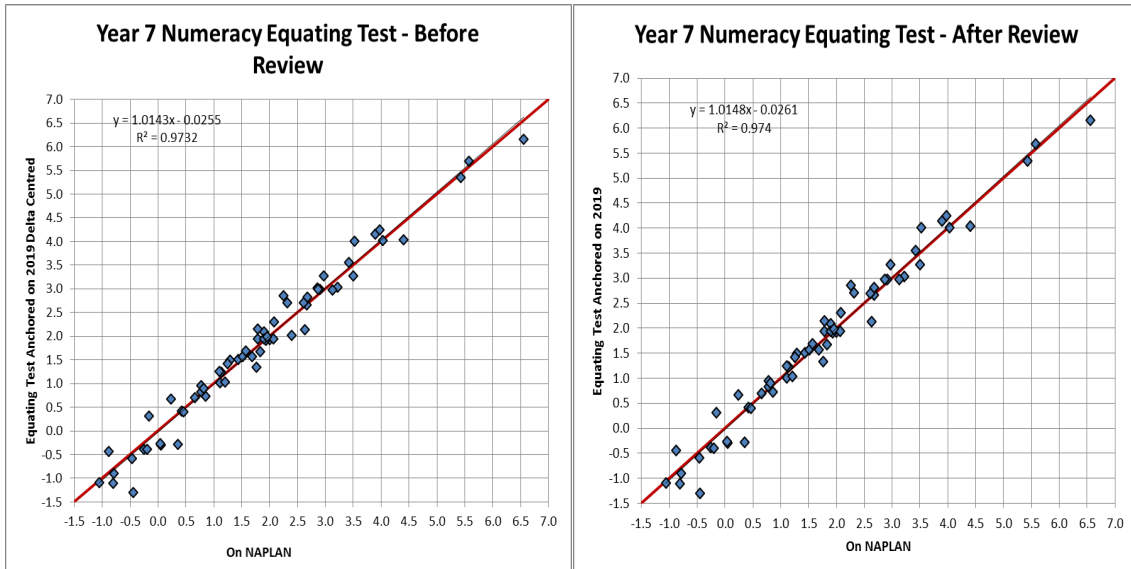


Figure 76. Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 7 online students

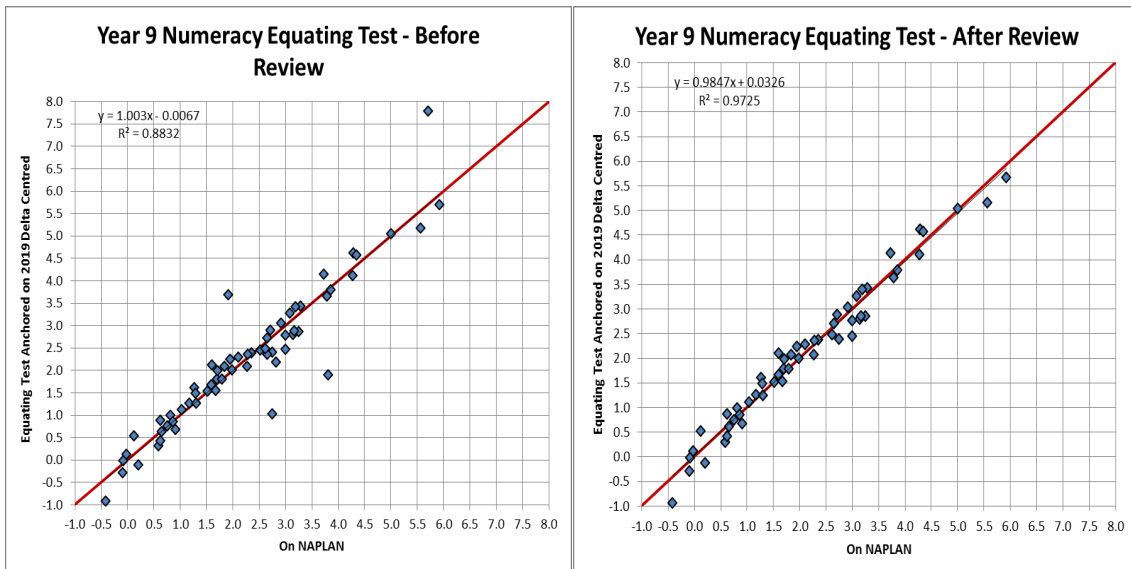


Figure 77. Scatterplot of numeracy, horizontal equating items between 2019 and 2009 for Year 9 online students

After the review and evaluation of the equating items, 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 2019 to 2008. These were not the final shifts to equate the 2019 results onto the historical scale. Instead, these horizontal shifts were subsequently adjusted, using the vertical equating shifts, resulting in the final HVR shifts. The numbers of horizontal links used and retained for each test are shown in Table 76. Table 77 shows the horizontal shift-constants for each domain at each year level by test mode. Appendix K presents the 2019 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 76. Horizontal link review summary

Test mode	Domain	Year 3	Test mode	Domain	Year 3	
Paper	Reading	30/35	32/37	42/47	38/47	
	Spelling	21/24	20/23	26/30	24/29	
	Grammar and punctuation	21/25	20/23	20/22	25/26	
	Numeracy	29/35	36/40	56/64	56/62	
Online	Reading	27/35	30/37	44/47	40/47	
	Spelling	23/24	20/23	26/30	27/29	
	Grammar and punctuation	GC	31/31	31/31	29/30	29/31
		GE1/GE2	25/25	23/25	23/25	25/25
		GE3	24/25	23/25	25/25	24/25
		GF	27/31	29/30	29/30	26/29
	Numeracy	31/35	35/40	62/64	55/62	

Table 77. Horizontal equating shifts between 2019 item locations and 2009 item locations by test mode

Test mode	Year level	Reading	Spelling	Grammar and punctuation	Numeracy	
Paper	Year 3	-0.085		-0.921	0.144	-1.102
	Year 5	0.865		1.381	1.169	0.257
	Year 7	1.815		2.348	1.538	1.364
	Year 9	2.237		3.325	2.199	2.174
Online	Year 3	0.182		-1.667	C: -0.309 E1&E2: 1.687 E3: 1.345 F: 3.544	-1.115
	Year 5	1.007		0.348	C: -0.552 E1&E2: 1.320 E3: 0.829 F: 2.990	0.396
	Year 7	1.746		1.488	C: -0.435 E1&E2: 1.082 E3: 0.272 F: 3.130	1.552
	Year 9	2.310		2.868	C: -0.443 E1&E2: 0.730 E3: 0.506 F: 2.578	2.550

Vertical equating shifts

As in previous years of testing, the NAPLAN 2019 Reading, Spelling, Grammar and Punctuation and Numeracy 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 paper tests and online tests. There was no vertical equating carried out for online grammar and punctuation due to a very small number of common items between testlets.

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 2019 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.

The quality of these common items in functioning as equating links between year levels was systematically reviewed for each domain. Only items that showed satisfactory and similar psychometric properties in the adjacent year levels were used as link items when scaled separately for each year level.

A common item was considered for omission (that is, not to be used for vertical linking purposes) based on the fit of the item and evidence for Differential Item Functioning (DIF) between year levels. As there were usually not many common items in the paper tests between year levels, it was generally agreed to maximize the number of links retained, where possible, as in previous years. Review of the vertical link items was undertaken in steps outlined below:

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

Step 2. Each item was checked for misfit at each year level 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 year level 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.

Step 3. Items were omitted if they showed cross year-level DIF. The impact was judged based upon changes in the shift constant and the slope of the best fit line. Items were considered for exclusion from being used as common links based on DIF if the absolute adjusted difference was greater than 0.3 and if the absolute standardised difference was greater than the criterion set. In such cases, these items were treated as different items in each of the year levels.

After each stage, the cross-year level scatterplot was evaluated 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 year level (that is, Year 3 SD / Year 5 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.

The evaluate year level DIF, difficulties of the set of common items were centred around zero for each year level. For each pair of adjacent tests, one set of item difficulties (for example, of Year 3 link items) was then plotted against the other set of item difficulties (of Year 5 link items). Two plots are presented below 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. The 24 plots of the vertical equating for the paper tests are shown in Figure 78 to Figure 89. Another 18 plots of the vertical equating for the online tests are shown in Figure 90 to Figure 98. For each set of adjacent year level scales, mean item parameters of the link items were calculated for each of the two-year levels. The vertical shift is the difference between the two means.

Vertical link item review of paper tests

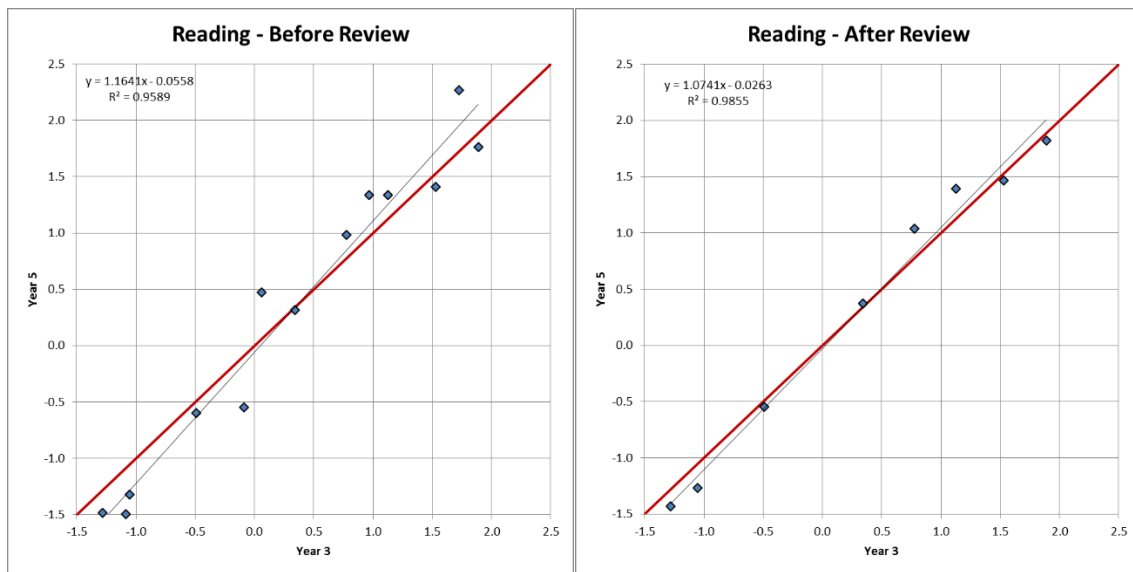


Figure 78. Scatterplot for vertical link item review for reading between Year 3 and Year 5 paper tests

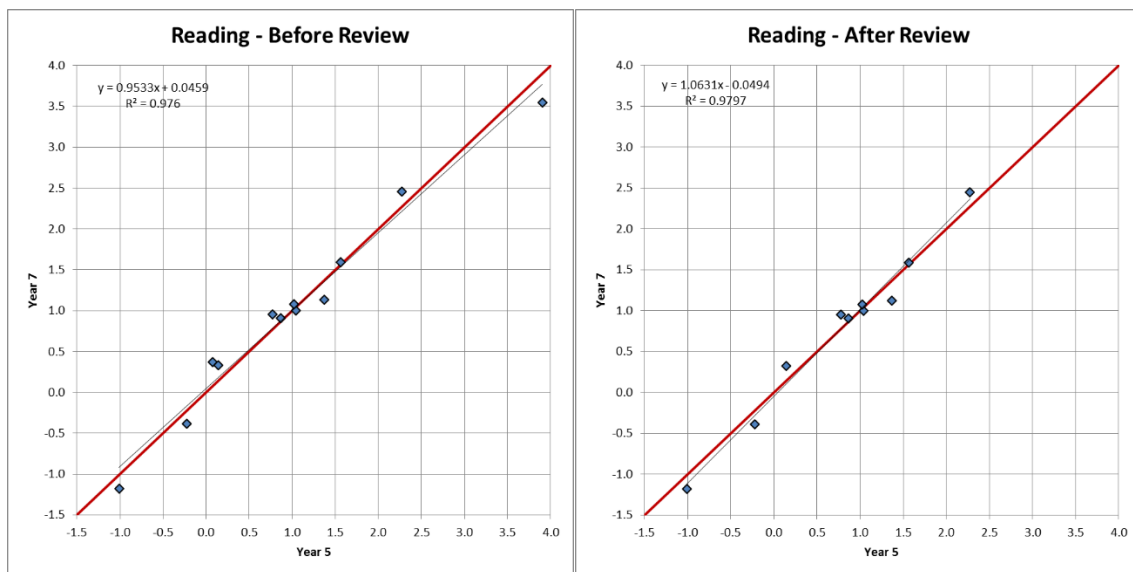


Figure 79. Scatterplot for vertical link item review for reading between Year 5 and Year 7 paper tests

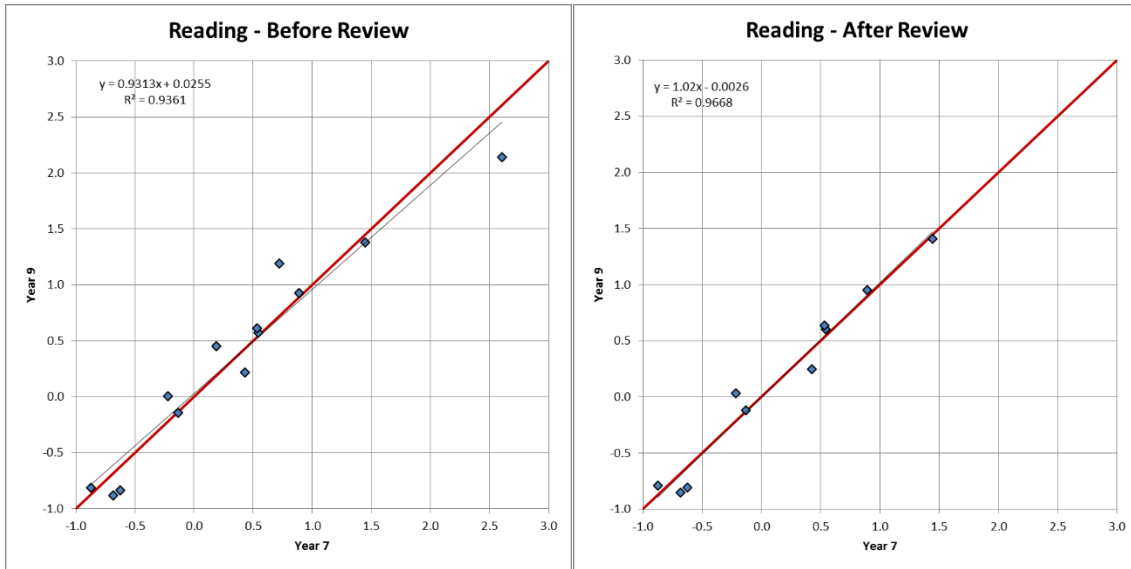


Figure 80. Scatterplot for vertical link item review for reading between Year 7 and Year 9 paper tests

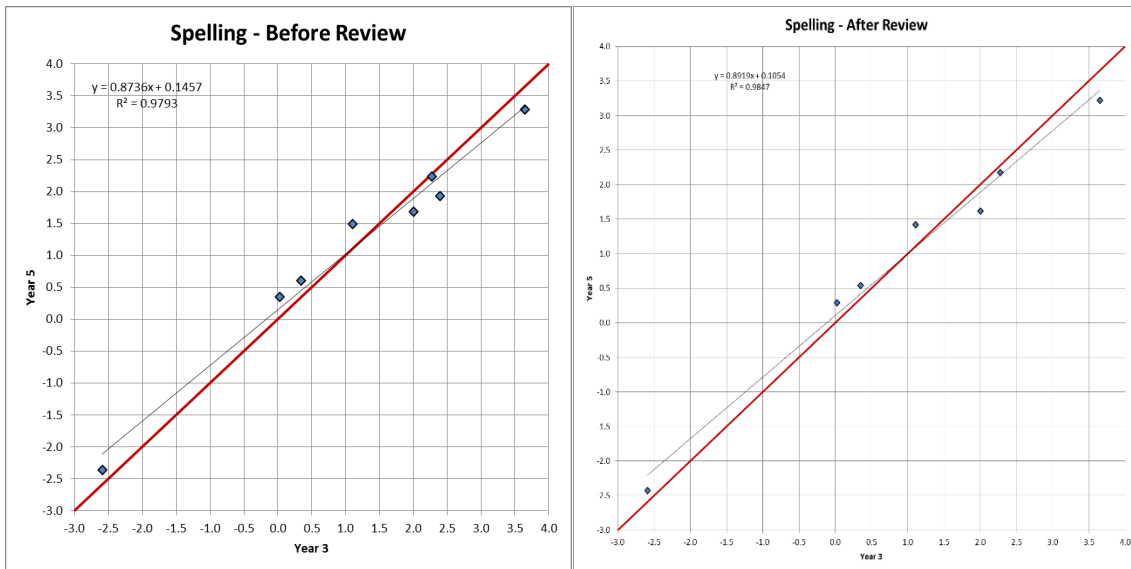


Figure 81. Scatterplot for vertical link item review for spelling between Year 3 and Year 5 paper tests

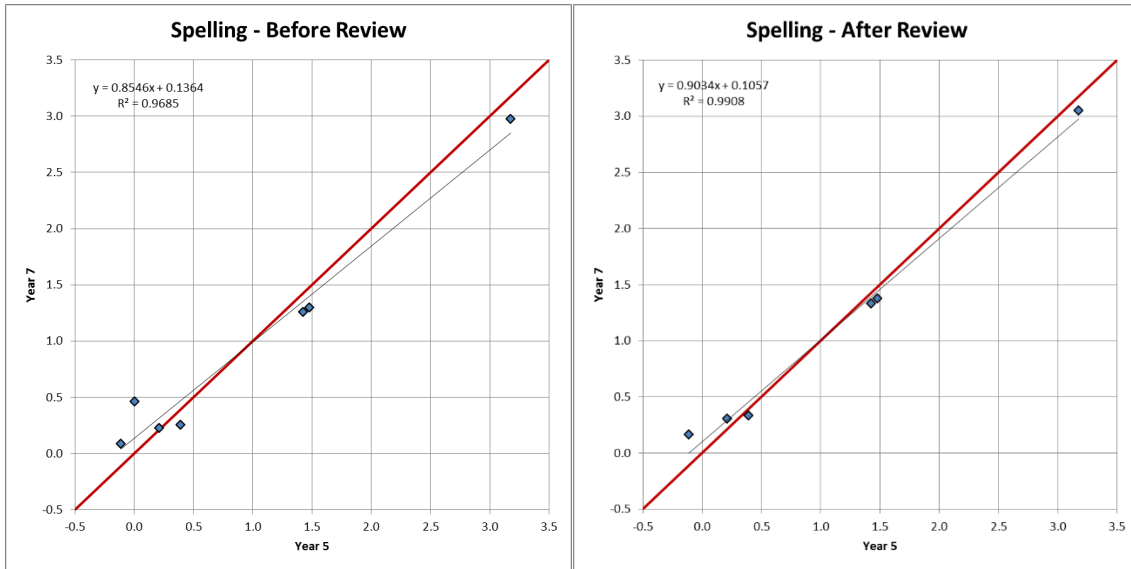


Figure 82. Scatterplot for vertical link item review for spelling between Year 5 and Year 7 paper tests

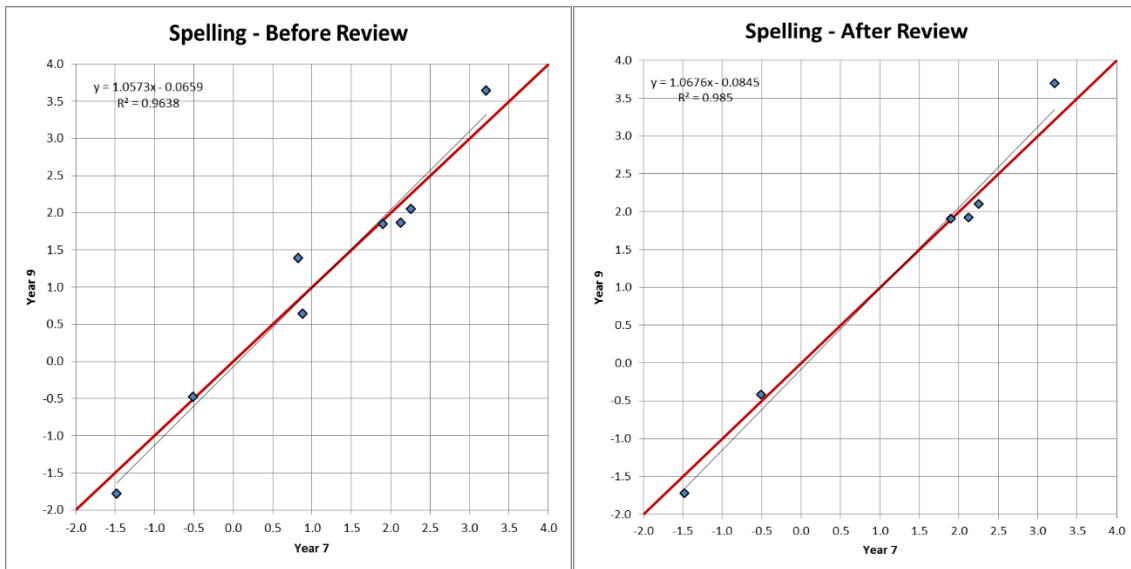


Figure 83. Scatterplot for vertical link item review for spelling between Year 7 and Year 9 paper tests

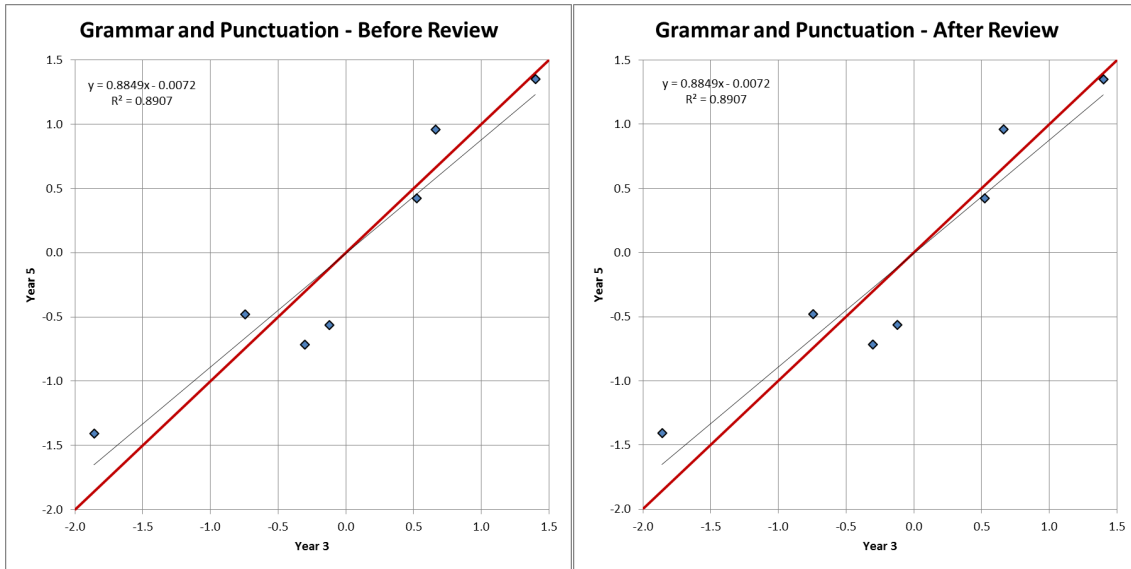


Figure 84. Scatterplot for vertical link item review for grammar and punctuation between Year 3 and Year 5 paper tests

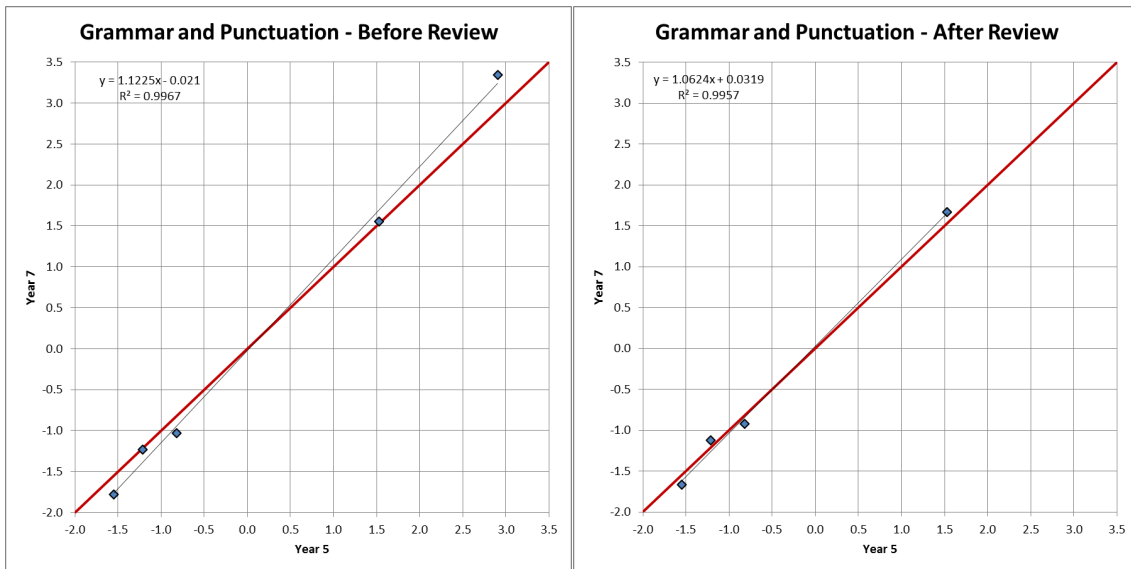


Figure 85. Scatterplot for vertical link item review for grammar and punctuation between Year 5 and Year 7 paper tests

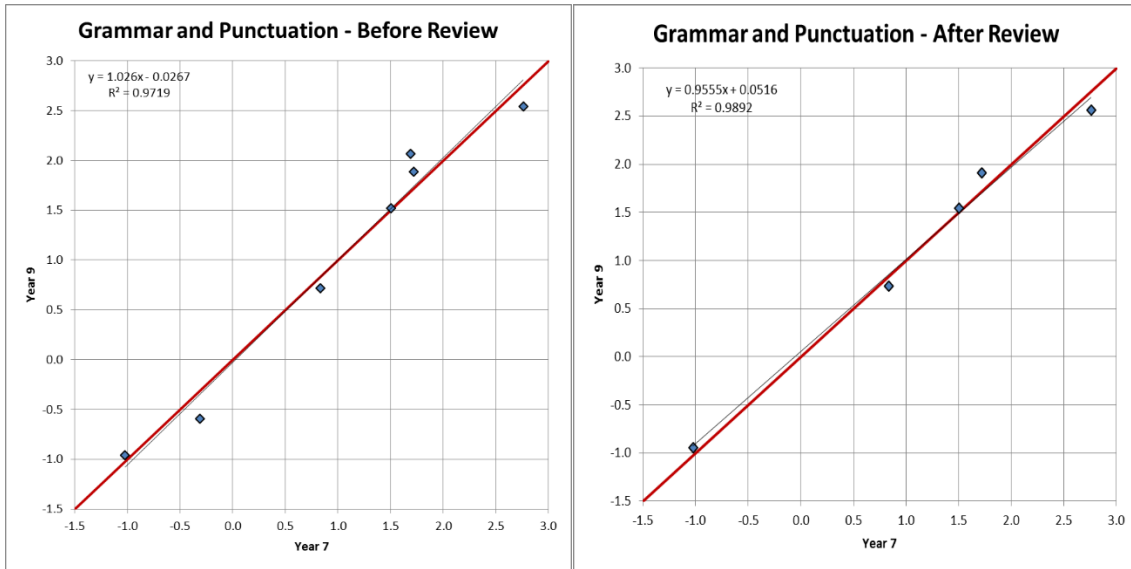


Figure 86. Scatterplot for vertical link item review for grammar and punctuation between Year 7 and Year 9 paper tests

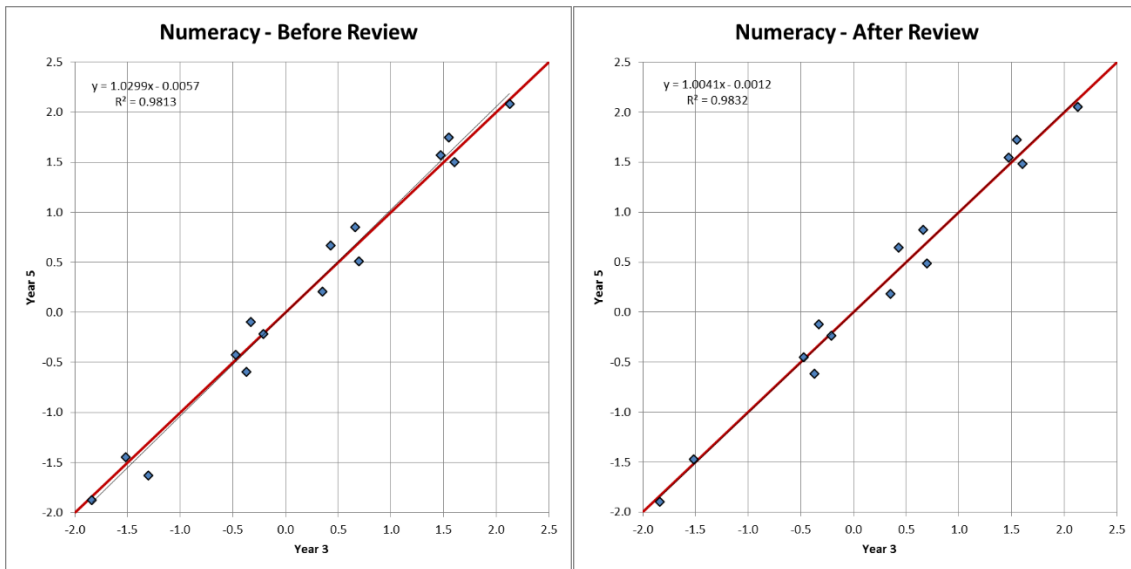


Figure 87. Scatterplot for vertical link item review for numeracy between Year 3 and Year 5 paper tests

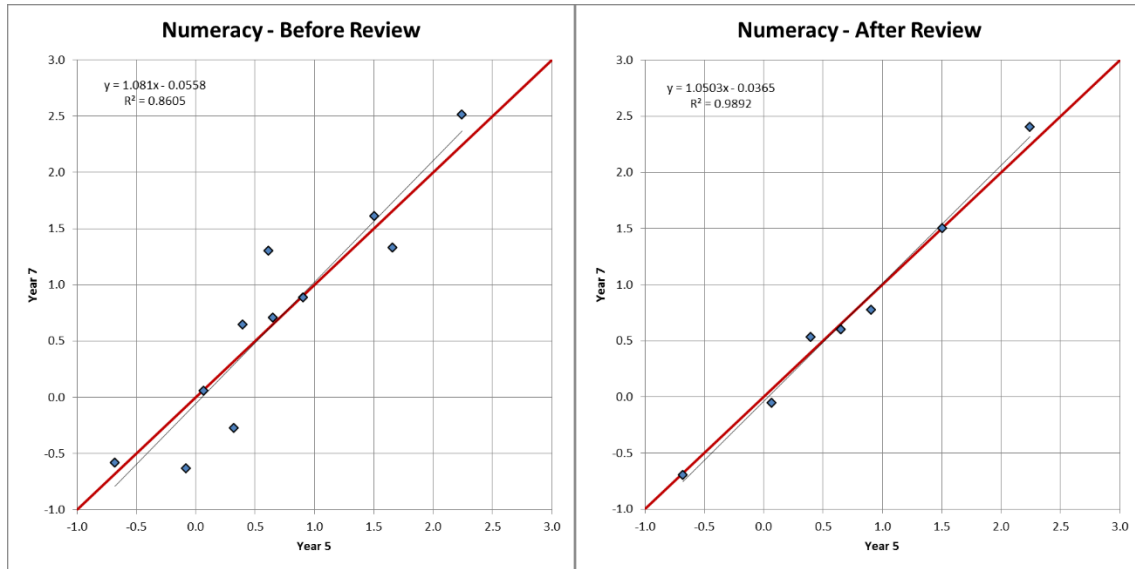


Figure 88. Scatterplot for vertical link item review for numeracy between Year 5 and Year 7 paper tests

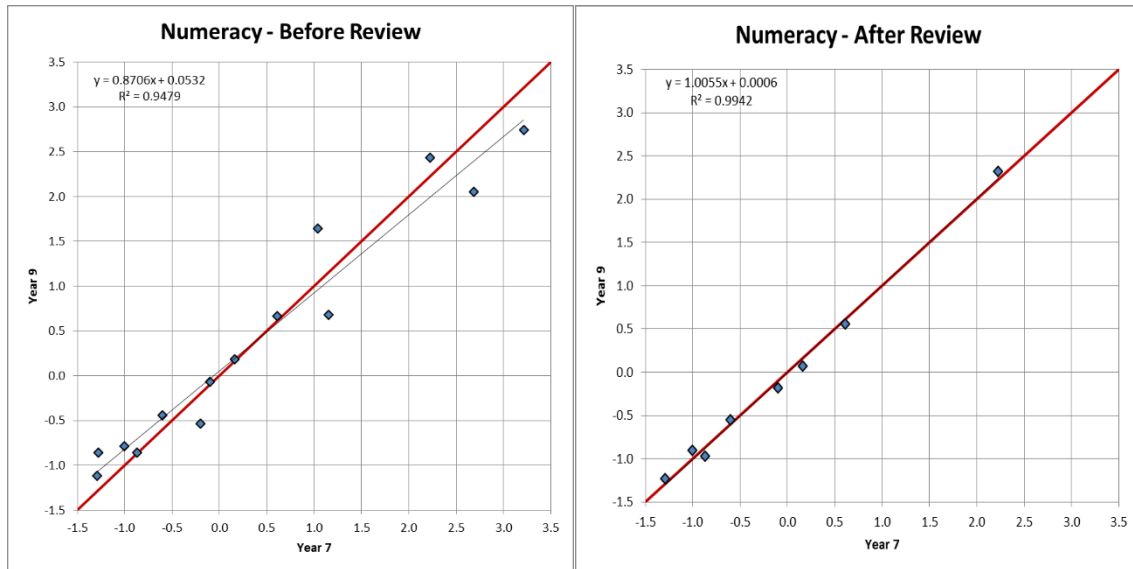


Figure 89. Scatterplot for vertical link item review for numeracy between Year 7 and Year 9 paper tests

Vertical link item review for online tests

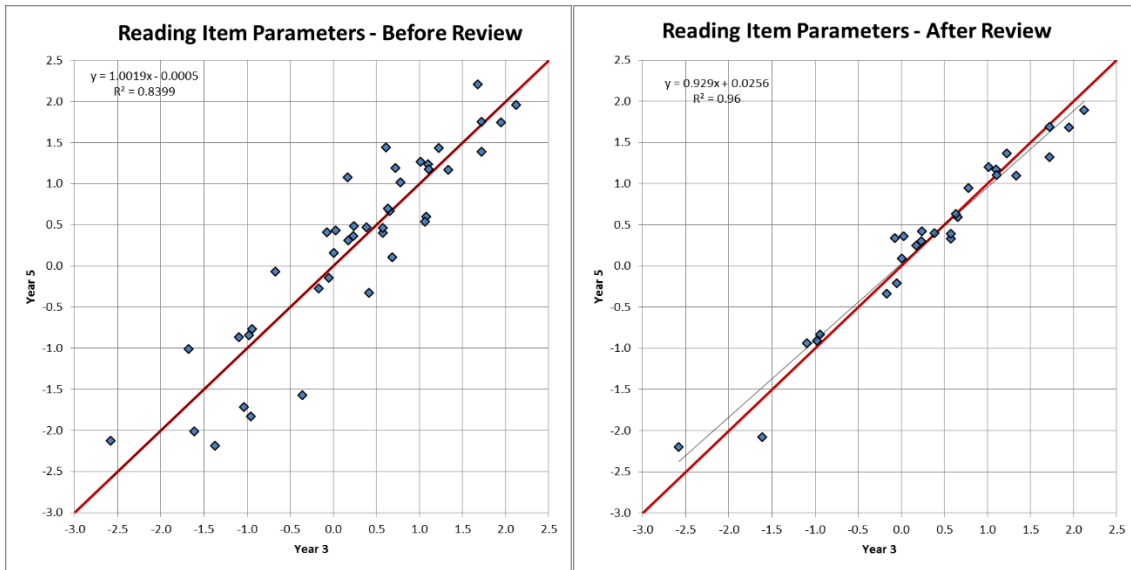


Figure 90. Scatterplot for vertical link item review for reading between Year 3 and Year 5 online tests

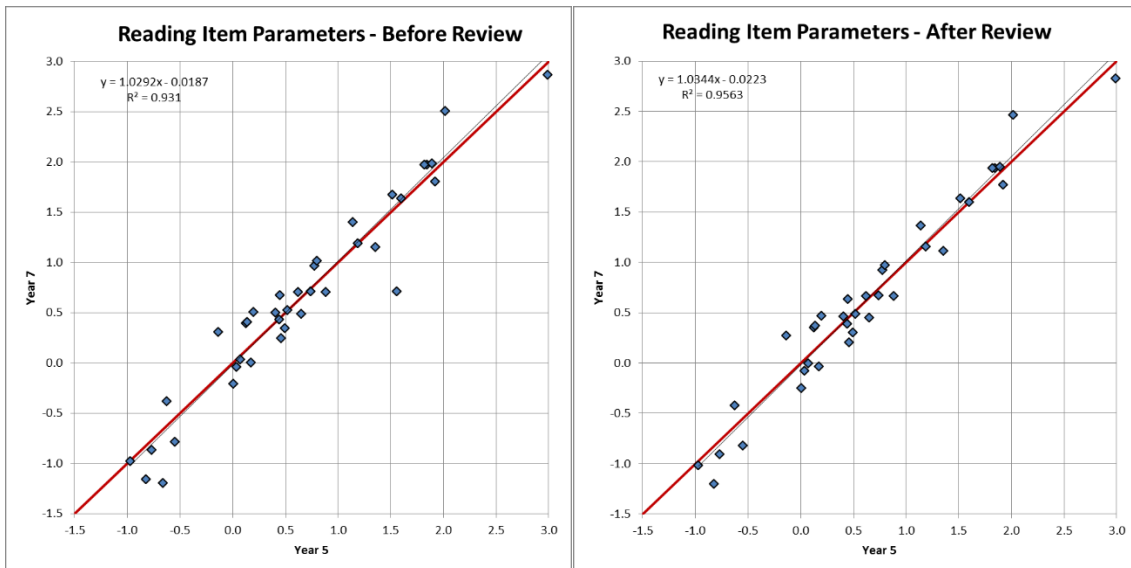


Figure 91. Scatterplot for vertical link item review for reading between Year 5 and Year 7 online tests

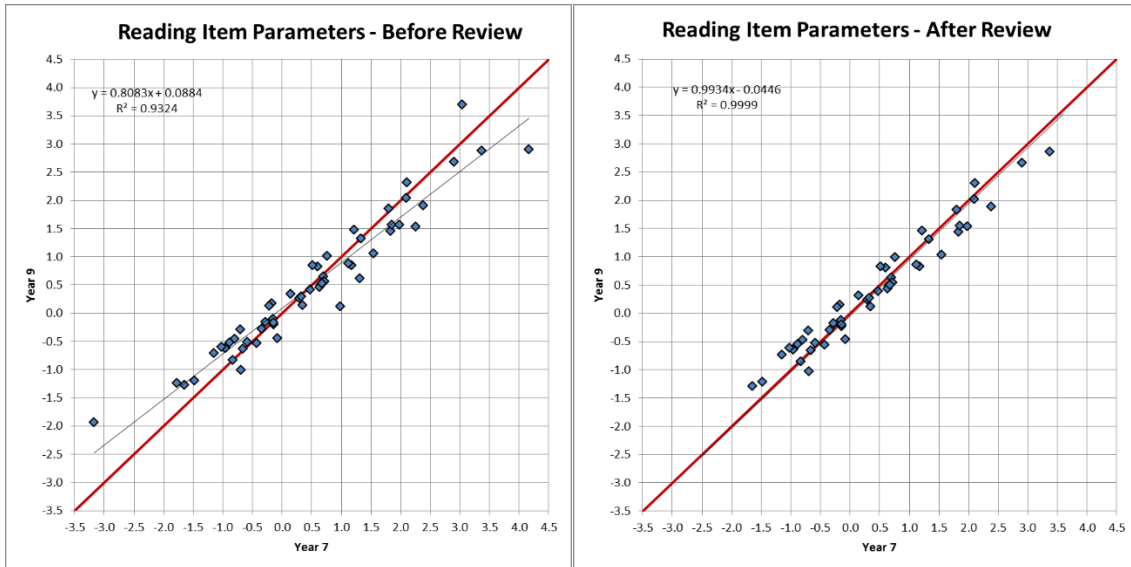


Figure 92. Scatterplot for vertical link item review for reading between Year 7 and Year 9 online tests

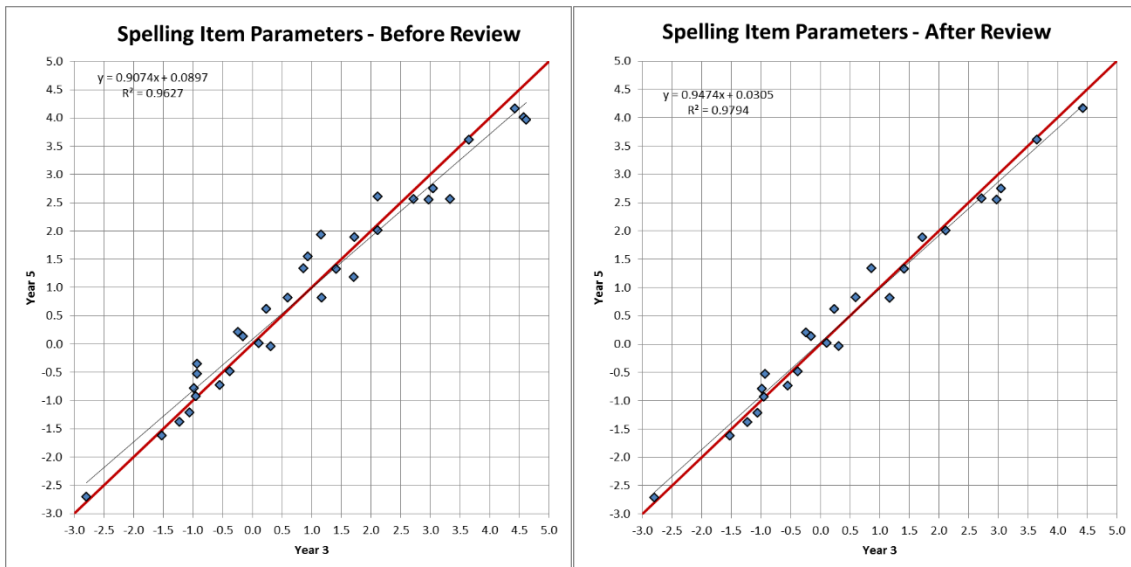


Figure 93. Scatterplot for vertical link item review for spelling between Year 3 and Year 5 online tests

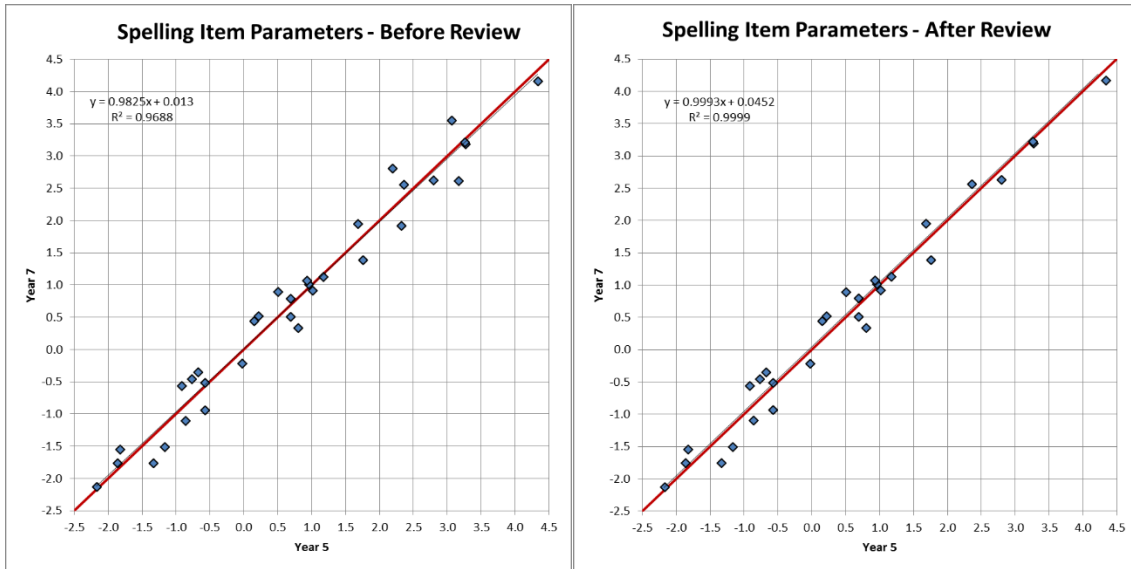


Figure 94. Scatterplot for vertical link item review for spelling between Year 5 and Year 7 online tests

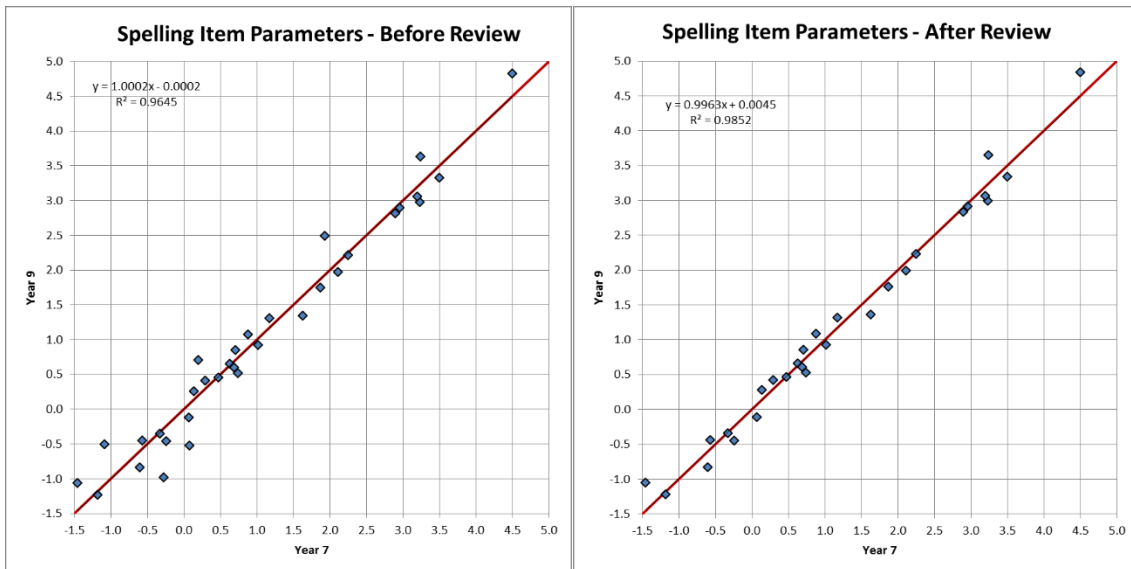


Figure 95. Scatterplot for vertical link item review for spelling between Year 7 and Year 9 online tests

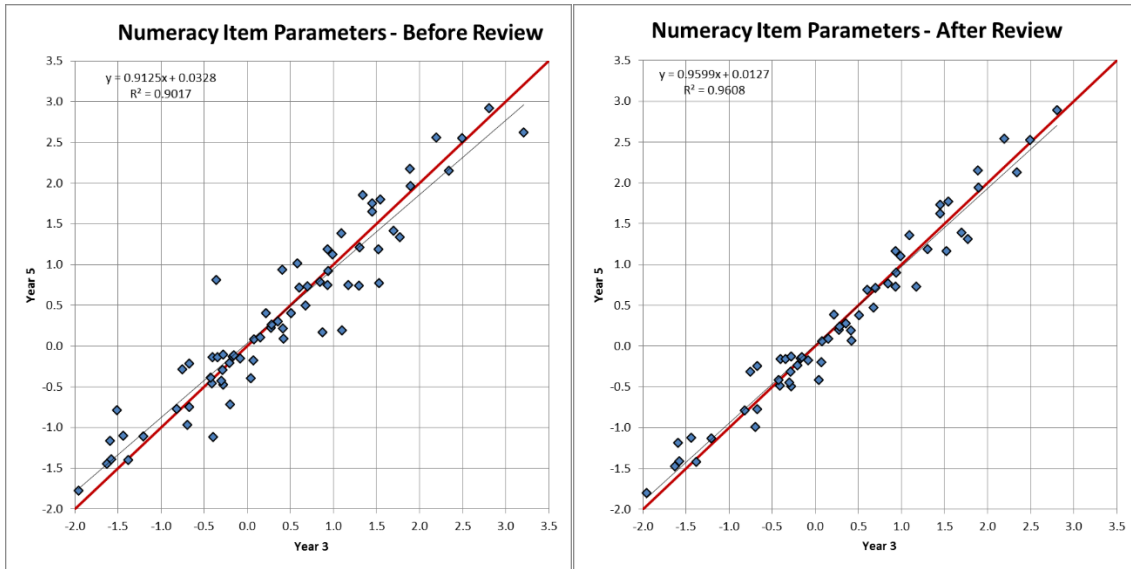


Figure 96. Scatterplot for vertical link item review for numeracy between Year 3 and Year 5 online tests

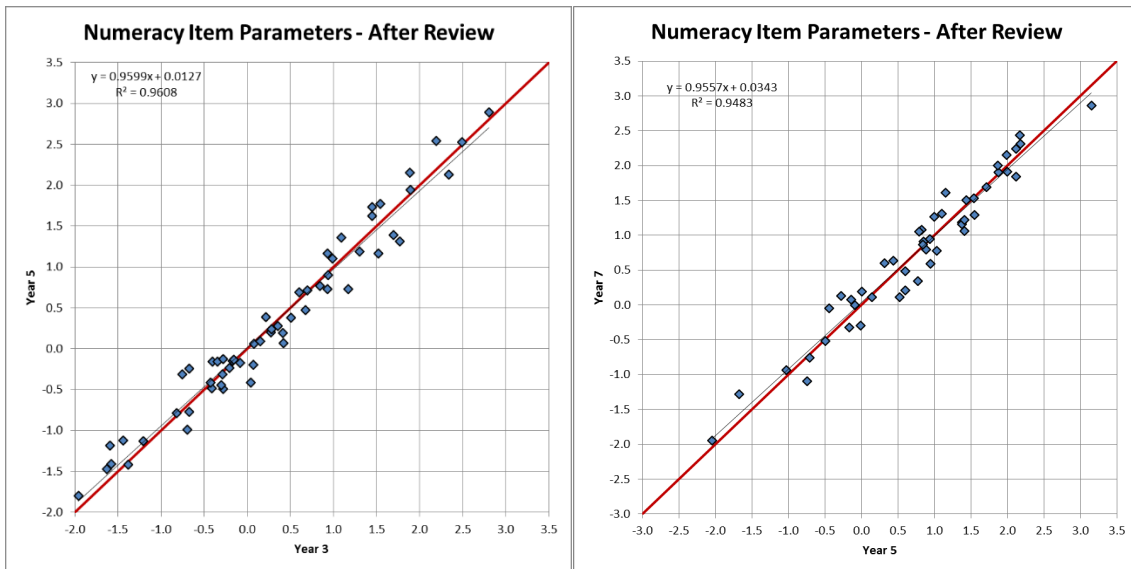


Figure 97. Scatterplot for vertical link item review for numeracy between Year 5 and Year 7 online tests

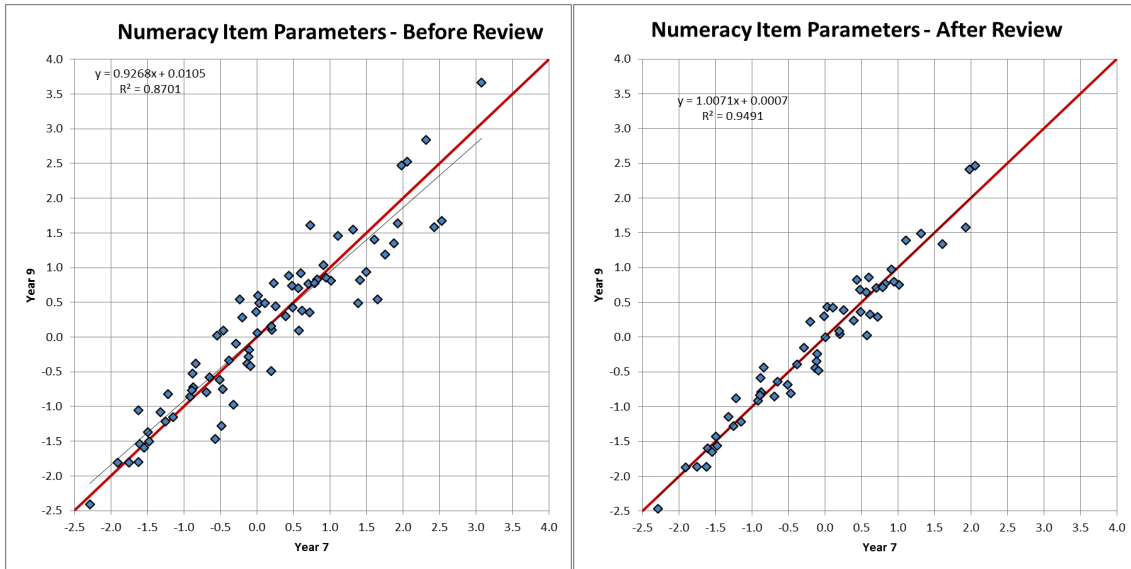


Figure 98. Scatterplot for vertical link item review for numeracy between Year 7 and Year 9 online tests

The numbers of vertical links used and retained for each adjacent pair of year levels are shown in Table 78. Appendix L presents the 2019 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 78. Vertical link review summary

Test mode	Domain	Year 3/5		Year 5/7		Year 7/9	
		No. links retained	Total No. of links	No. links retained	Total No. of links	No. links retained	Total No. of links
Paper	Reading	8	13	10	12	10	13
	Spelling	7	8	6	7	6	8
	Grammar and punctuation	7	7	4	5	5	7
	Numeracy	14	15	7	11	8	14
Online	Reading	28	42	36	38	49	56
	Spelling	25	33	29	33	28	33
	Numeracy	58	70	48	60	56	77

The mean shifts between two adjacent year levels for each of the four domains are shown in Table 79 and mean shifts between each year level and Year 5 are shown in Table 80.

Table 79. Vertical shift constants between adjacent year levels

Test mode	Shift	Reading	Spelling	Grammar and punctuation	Numeracy
Online	Years 3 to 5	-0.899	-1.870		-1.236
	Years 5 to 7	-0.673	-1.284		-1.236
	Years 7 to 9	-0.477	-1.200		-0.684
Paper	Years 3 to 5	-0.995	-1.893	-0.665	-1.273
	Years 5 to 7	-0.958	-1.181	-0.684	-1.000
	Years 7 to 9	-0.396	-0.843	-0.258	-0.577

Table 80. Vertical shift constants from each year level to Year 5

Test mode	Shift	Reading	Spelling	Grammar and punctuation	Numeracy
Online	Years 3 to 5	-0.899	-1.870	-	-1.236
	Years 5 to 5	0.000	0.000	-	0.000
	Years 7 to 5	0.673	1.284	-	1.236
	Years 9 to 5	1.150	2.484	-	1.920
Paper	Years 3 to 5	-0.995	-1.893	-0.665	-1.273
	Years 5 to 5	0.000	0.000	0.000	0.000
	Years 7 to 5	0.958	1.181	0.684	1.000
	Years 9 to 5	1.354	2.024	0.942	1.577

The final equating parameters to place the 2019 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

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 2019 tests on the historical scale for each domain. The horizontal equating tests were first used in 2009 and reused every subsequent year.

Table 75 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 shifts.

Table 81 and Figure 99 explain the HVR method using the online numeracy test as an example. First, vertical shifts are calculated from each year level to the Year 5 scale. For Year 3, this is equal to the original shift between the two adjacent year levels. For Year 5, the shift is equal to 0. For Year 7, it is equal to $-1 \times$ shift from Year 5 to Year 7. For Year 9,

this is equal to $-1 * (\text{shift from Year 5 to Year 7} + \text{shift from Year 7 to Year 9})$. See also Table 80.

The shifts in the second column are equal to the shifts presented in Table 77. 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 81: Example of comparing horizontal shifts with vertical shifts (numeracy, online test)

	2019 vertical shift to Year 5	Horizontal shift 2019 to 2008	Adjusted horizontal shift	Predicted horizontal shift
Year 3	-1.236	-1.115	-1.511	-1.091
Year 5	0.000	0.396	0.000	0.304
Year 7	1.236	1.552	1.156	1.699
Year 9	1.920	2.550	2.154	2.471

Therefore, the horizontal shifts in column two (Y) were regressed onto the vertical shifts in column one (X). A scatterplot of these shifts is presented in Figure 99. 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 2019 results onto the historical scale. Generally, the HVR shifts were very close to the horizontal shifts.

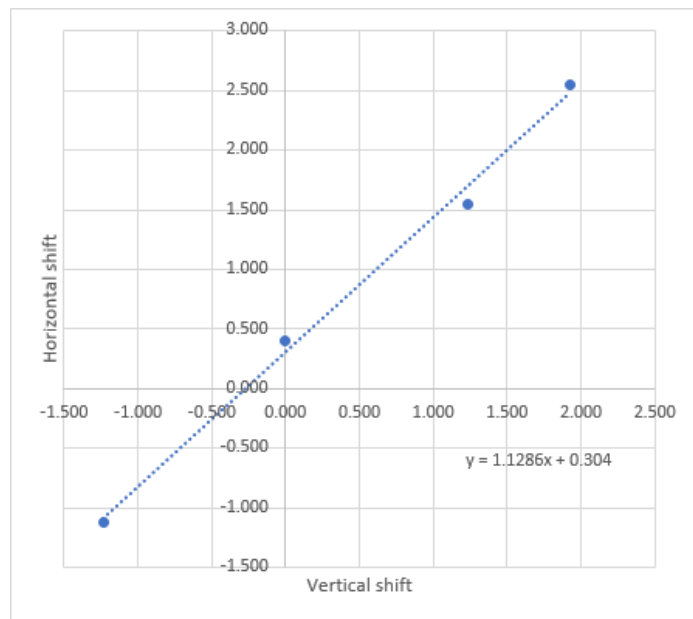


Figure 99: Example HVR-shift (numeracy, online test)

Figure 100 and Figure 101 show the plots of the positions of the four 2019 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 and for online tests, respectively. 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 other three domains, there was no particular year level that stands out as an outlier. These regression shifts were used for final equating of the 2019 grammar and punctuation paper tests to the NAPLAN historical scale. For the online tests, the similar patterns were found for reading, spelling and numeracy as the pattern in the paper tests. There was no regression equating for grammar and punctuation tests.

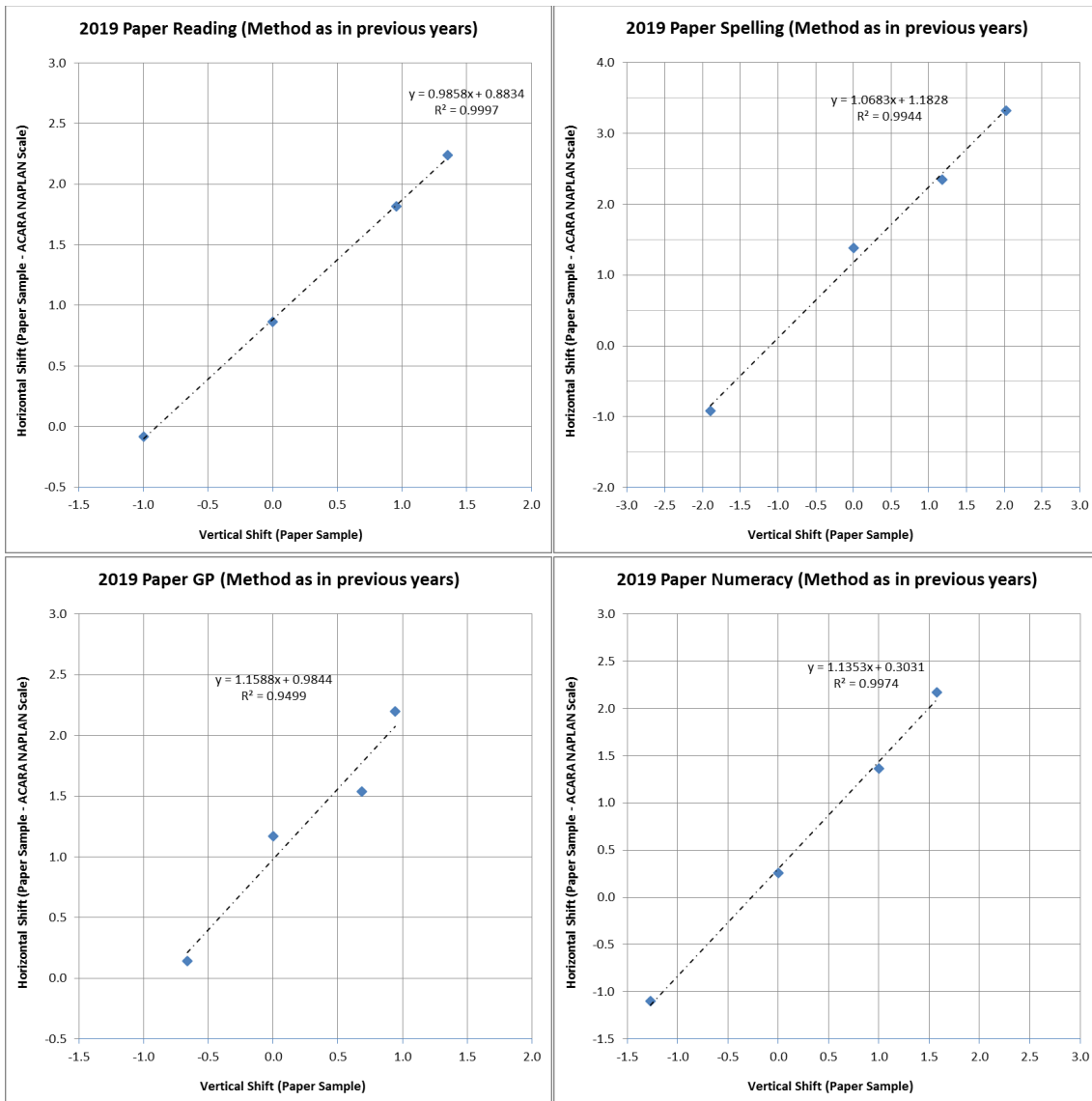


Figure 100. Comparisons of horizontal and vertical shifts of the paper tests

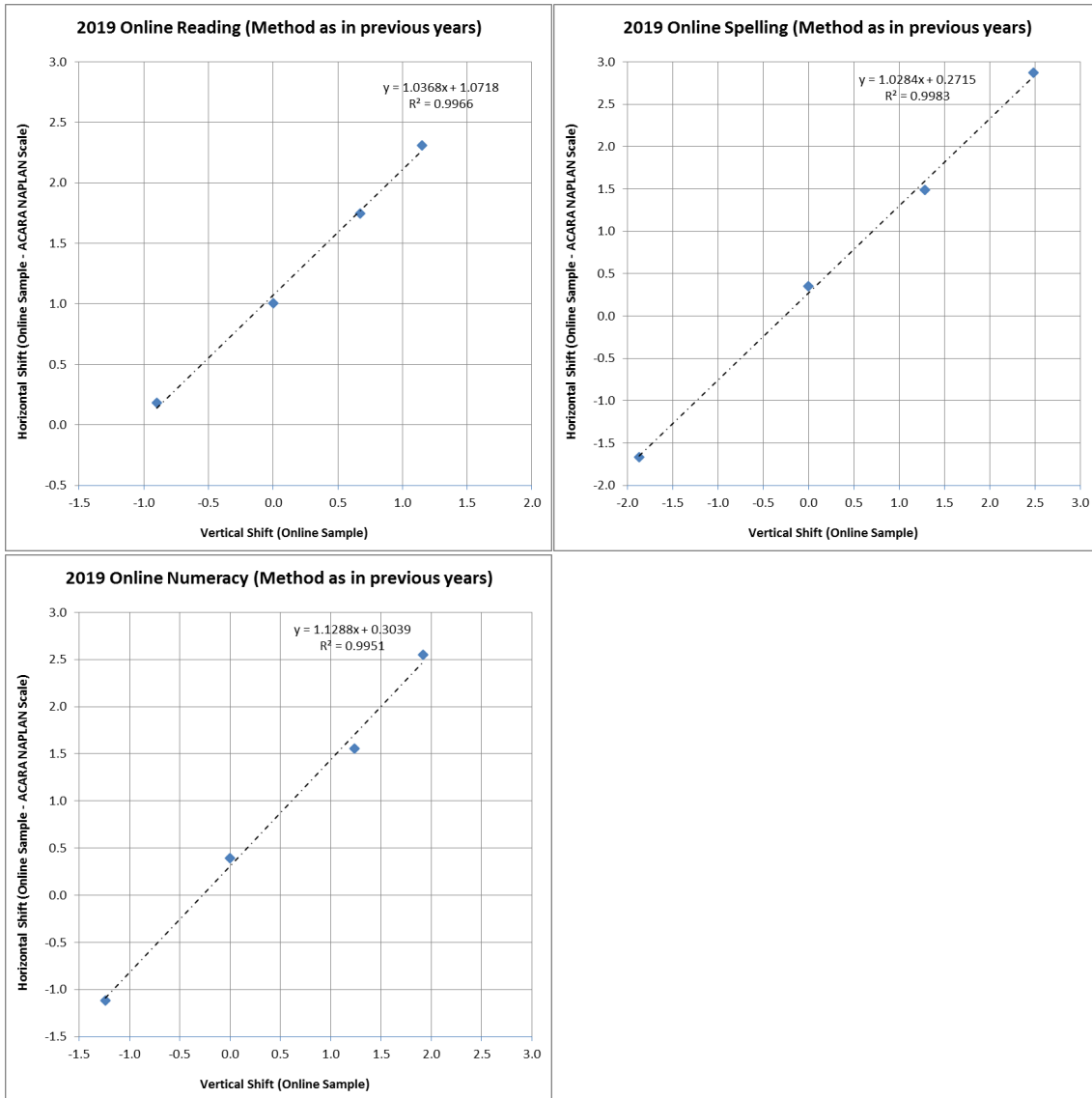


Figure 101. Comparisons of horizontal and vertical shifts of the online tests

Table 82 displays the intercepts and slopes for the regression-based combination of the vertical and horizontal equating shifts.

Table 82. Regression intercepts and slopes

Test mode	Regression coefficient	Reading	Spelling	Grammar and punctuation	Numeracy
Paper	Intercept (a)	0.883	1.183	0.984	0.303
	Slope (b)	0.986	1.068	1.159	1.135
Online	Intercept (a)	1.072	0.272	n/a	0.304
	Slope (b)	1.037	1.028	n/a	1.129

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 2019 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 2019 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.

The final, regression-based shifts for each domain by test mode that were calculated using equation 6.1 are shown in Table 83 by year level. These equating shifts were applied to paper tests and online tests in Years 3, 5, 7 and 9 to put the tests on the NAPLAN historic scale.

Table 83: Final HVR shifts applied for equating NAPLAN 2019 onto the NAPLAN historic scale

Test mode	Year level	Reading	Spelling	Grammar and punctuation*	Numeracy
Paper	3	-0.0975	-0.8390	0.2136	-1.1423
	5	0.8834	1.1828	0.9844	0.3031
	7	1.8279	2.4448	1.7765	1.4383
	9	2.2366	3.3445	2.0757	2.0940
Online	3	0.1399	-1.6518	-0.7518*	-1.0910
	5	1.0718	0.2715	0.2612*	0.3039
	7	1.7694	1.5922	0.9034*	1.6987
	9	2.3102	2.8255	1.7331*	2.4713

* The shifts of grammar and punctuation of online tests are the horizontal shifts.

Scaling factors

Applying a scale factor is sometimes necessary due to the potential impact that differences in test reliability can have on student score spread. As the equating tests measure the same construct as the NAPLAN tests, the equating test and the 2019 NAPLAN test is expected to result in the same latent distribution for the same group of students (the equating sample). In this case, the scale factor would be very close to 1. However, due to differences in test reliabilities between equating test forms and the current NAPLAN tests, the spread of scores from the equating test and the NAPLAN 2019 test for the equating sample was found to be quite different for some year levels and domains. In 2019, the scale factors were estimated using a 2-dimensional Rasch model, with items from the equating test loading onto one dimension and items from the 2019 NAPLAN test loading onto the second dimension. This concurrent analysis included students from the equating samples only. The scale factor was derived as the standard deviation (square root of the latent variance) ratio between the 2019 NAPLAN test dimension and the equating test dimension. A scale factor that was greater than 1.0 indicated that the equating test spread the students out more than the 2019 test did for that domain at the particular year level. Conversely, a scale factor that was less than 1.0 indicated that the NAPLAN 2019 test spread the students out more than the equating test for that domain at that particular year level.

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 2019 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 which provided in Table 83.

The values for *LocalMean* and *SF* are presented in Table 84 for each year level by domain. The online grammar and punctuation were equated by testlet, the online reading, spelling and numeracy were equated by year level.

Table 84: Local means and scaling factors

Domain and year	Online		Paper	
	Local mean	Scale factor	Local mean	Scale factor
N3	0.2832	1.0293	0.4303	1.0562
N5	0.2744	0.8408	0.5215	0.8721
N7	-0.0116	0.9673	0.2937	0.9312
N9	-0.2495	0.9782	0.1963	0.9692
R3	-0.1172	1.1951	0.2123	1.547
R5	0.1707	0.9642	0.3763	1.1438
R7	0.0485	0.9742	0.0461	1.0904
R9	-0.0411	1.1291	0.0928	1.2354
S3	0.3575	0.9877	-0.2354	1.0967
S5	0.4816	1.0624	-0.3606	1.1213
S7	0.5037	0.9068	-0.2685	1.0392
S9	0.2447	0.9209	-0.2185	0.9328
G3	0	1	0.2092	1.7732
G5	0	1	0.1381	1.4754
G7	0	1	0.2687	1.1206
G9	0	1	0.1404	1.1117

For online grammar and punctuation, no scale factor was applied; instead, an equipercentile equating transformation was applied, as detailed in the next section.

The same transformation was applied to the WLE ability logit scores in the score equivalence table, 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, 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 2019 data were freely calibrated and compared to the item difficulties of 2016. The year 2016 was chosen because the writing genre was persuasive in 2019 and in 2016 while the genre was narrative in 2018 and in 2017. The scatter plot between the two calendar years are shown by test mode in Figure 102. They indicate that the consistency of relative difficulties supported using the anchoring method in 2019.

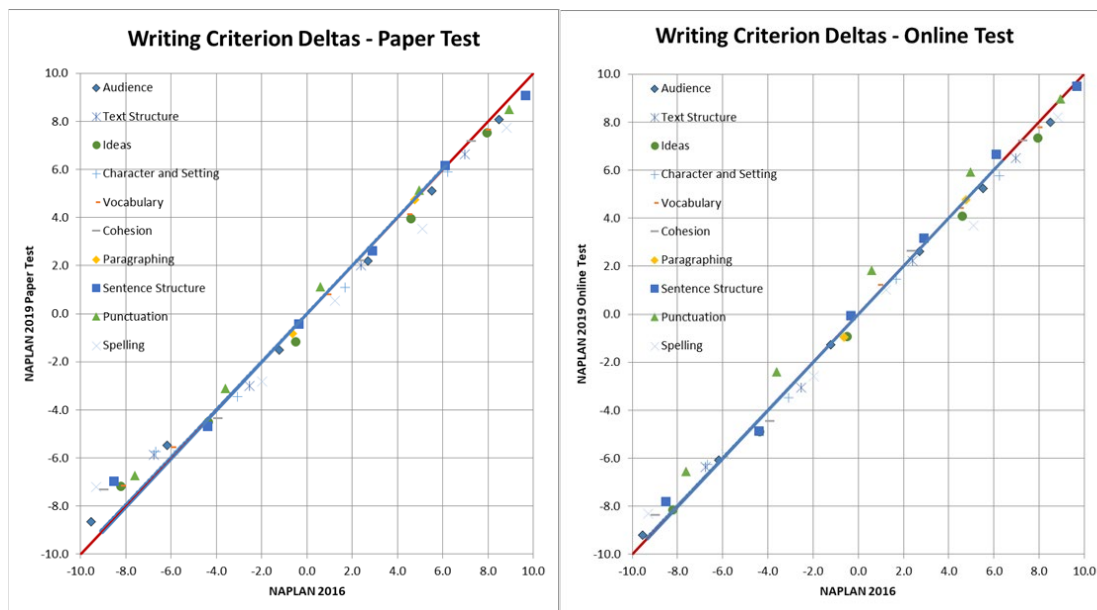


Figure 102: Scatterplot for writing criteria between 2019 and 2016 paper and online tests

In addition to comparing relative item difficulties, an equating verification study was conducted by 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 purpose of the pairwise study was to triangulate scores awarded by markers in 2016 and 2019 with a separate common scale formed from pairwise comparisons of the 2016 and 2019 scripts. This provides a common frame of reference by which to compare marking in 2016 with marking in 2019 (paper and online) as well as to compare 2019 paper marking with 2019 online marking. The study enables checks on marker consistency across time and across modes (paper vs. online), as detailed to follow.

The equating verification study comprised the following key components:

- pairwise comparisons of 2016 and 2019 scripts (both online and paper for 2019), placing the 2019 scripts on the 2016 scale
- cross-referencing pairwise locations for 2016 and 2019 with rubric-based locations (official NAPLAN difficulties) for the same scripts

- cross-referencing pairwise locations for 2019 online and 2019 paper with rubric-based locations for the same scripts
- forming a pairwise scale using 2015, 2016 and 2019 scripts.

The pairwise study design for writing in the NAPLAN 2019 assessment was similar to that used in 2018, although it included an additional pairwise comparison component cross-referencing paper and online performances.

The equating verification consisted of two pairwise comparison projects for judging:

- Project 1: 2016 paper and 2019 paper scripts were compared.
- Project 2(a): 2016 paper scripts and 2019 online scripts were compared.
- Project 2(b): 2019 online scripts and 2019 online scripts were compared.

Project 2 comparisons enabled formation of a 2019 online pairwise scale.

For the paper component of the study (project 1), there were a total of 476 scripts of which 257 were 2016 scripts and 219 were 2019 scripts. For all scripts, a rubric score was available from marking conducted in the relevant calendar year. Around 37 paper scripts from 2016 were included for each state (approximately uniform distribution of 2016 scripts per state). The score distributions for the 2016 scripts were uniform, with scores ranging from 8 to 46 for each state. There were close to equal numbers of scripts for the two tasks in the sample (223 Year 3 and 5 scripts; 253 Year 7 and 9 scripts). Thirty-six judges from around Australia made a total of 10,024 comparisons between 2016 paper and 2019 paper performances (40 judges were allocated pairs, and 4 of these did not participate).

In first component of project 2(a), 2016 scripts were compared with 2019 online performances. The same 2016 paper scripts were used as for project 1 detailed above. There were 206 online 2019 scripts, with numbers of scripts per state ranging from 24 (NSW and NT) to 40 (Vic.). The score distributions were nearly uniform for each state. Scores for the 2019 online scripts ranged from 8 to 46. Of the 206 2019 online scripts, 108 were from the Year 3 and 5 task and 98 were from the Year 7 and 9 task. Thirty-seven judges from the seven jurisdictions made 10,737 comparisons of 2016 versus 2019 scripts (two judges were allocated pairs but did not participate).

For the second component of project 2 (b), the 206 online 2019 scripts were compared with one another to form a 2019 online scale. The reason for these comparisons was to check that the online–online comparisons produced a comparable scale for the 2019 online scripts as the scale produced from the paper–online comparisons. These comparisons were made by the same 37 judges as for the 2016 versus 2019 online comparisons. For the 2019 online versus 2019 online component, 5,302 comparisons were made.

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

To evaluate fit to the Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959), judge outfit indices were calculated for both project 1 and project 2 after removing extreme observations (comparisons for which the standardised residuals were greater than 7). For project 1, all but four judges had good outfit indices (less than 1.2), and all but two judges had outfit values below 1.3. The highest judge outfit for project 1 was 1.454. For project 2, only three of the 37 judges had outfit values above 1.2. The highest outfit values for project 2 were 1.259, 1.386 and 1.852.

Figure 103 shows the plot of pairwise scale locations (x-axis) against locations based on the rubrics (y-axis) for 2016 and 2019 paper scripts separately (project 1). The correlation overall was $r = 0.970$ for 2016 paper scripts and $r = 0.933$ for 2019 paper scripts. As can be seen, the fitted curves were somewhat curvilinear as in previous years of NAPLAN.

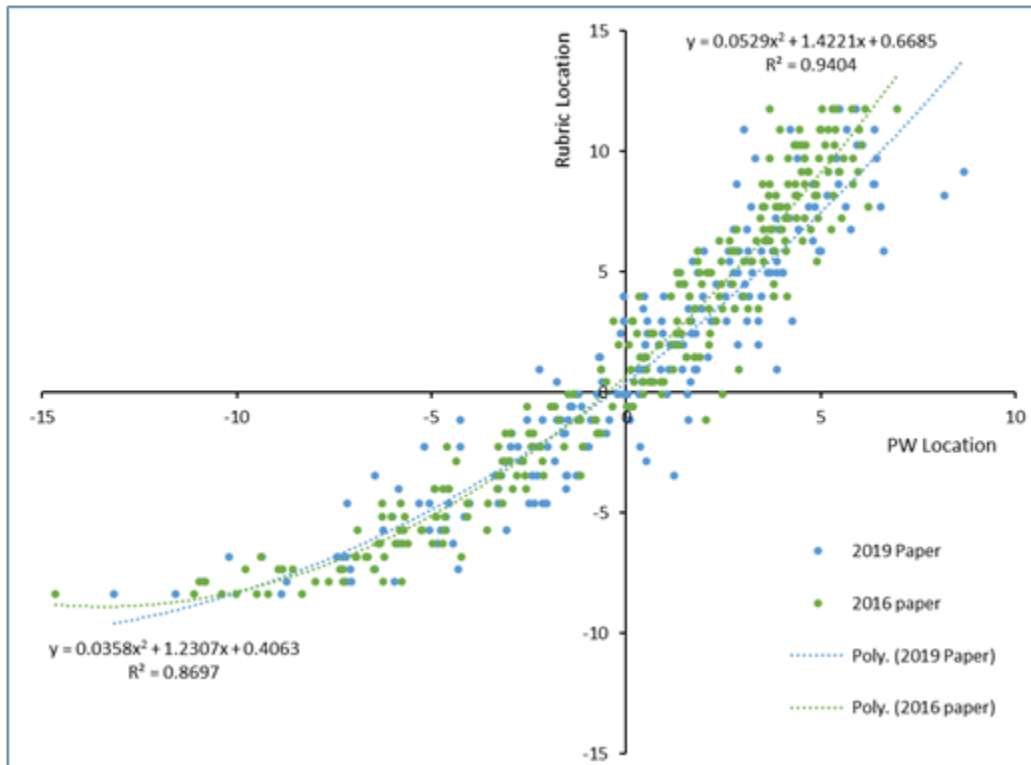


Figure 103: Scatterplot of the NAPLAN rubric and pairwise scale locations for 2016 and 2019 paper performances

The pairwise scale locations show the ordering of the scripts based on direct comparisons, whereas the NAPLAN scale locations are based on rubric marking. In the plot, 2016 paper and 2019 paper are highlighted separately. Regression lines are also shown separately for each of these years. There was a similar correspondence between the pairwise and NAPLAN scale locations for 2016 and 2019 scripts, with some departure in predicted values at the upper extreme. The correlation and nature of the relationship were relatively similar for both of these calendar years to the relationship observed in previous calendar years of NAPLAN.

Figure 104 shows pairwise scale locations and rubric locations for 2019 online and paper performances. Similar rubric locations were predicted from pairwise locations for both online and paper performances, though with some difference in the region of -5 to 0 logits on the pairwise scale. Note that 2019 paper performances and 2019 online performances were not compared directly. Rather, 2019 paper scripts were compared indirectly with 2019 online scripts using comparisons with 2016 paper performances. That is, both 2019 online and 2019 paper scripts were compared with 2016 paper scripts (projects 1 and 2(a)).

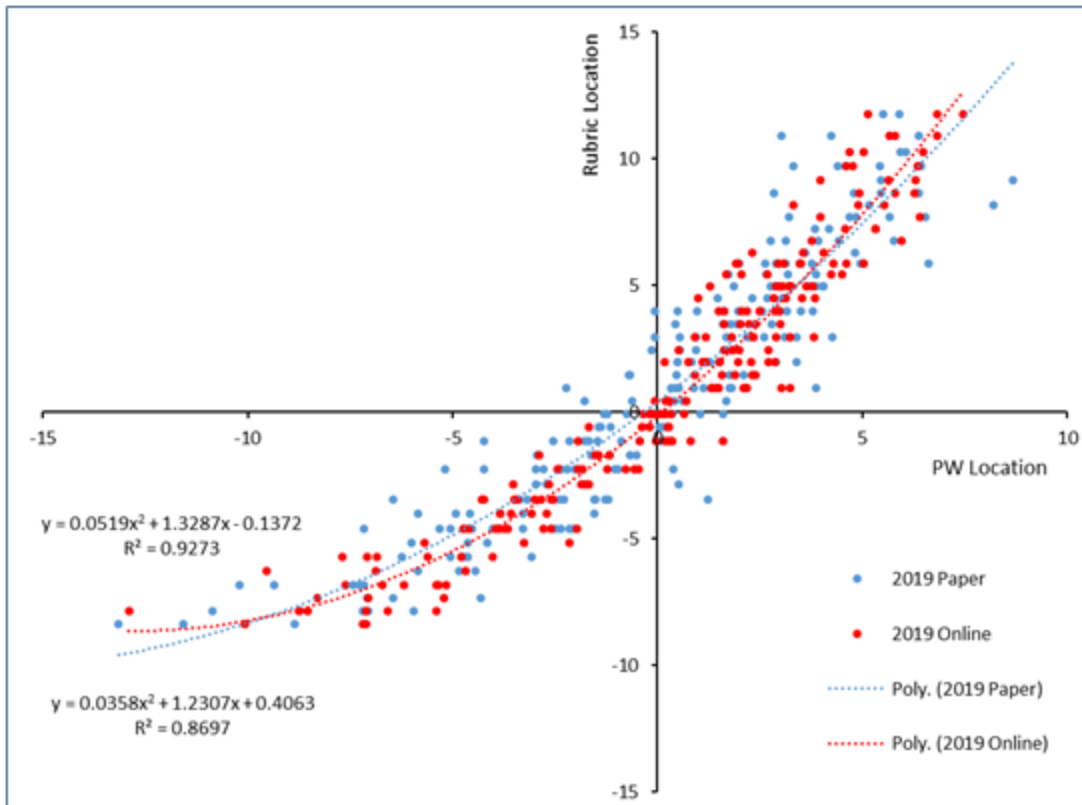


Figure 104: Scatterplot of the NAPLAN rubric and pairwise scale locations, comparing 2019 paper and online performances

Figure 105 shows 2019 online, 2019 paper and 2016 performances all together. Cross-referencing the pairwise locations with rubric locations indicated quite a high level of consistency in marking between 2016 and 2019. This evidence can be interpreted in combination with the high level of correspondence between threshold locations in 2016 and 2019. However, some differences were observed in relation to the marking of performances administered online.

As a check that 2019 paper and 2019 online scripts can be placed on the same scale, the linear relationship between scale locations for online performances was compared from two sets of comparisons: (i) 2016 paper versus 2019 comparisons; and (ii) 2019 online versus 2019 online comparisons. The first of these sets of comparisons involved indirect scaling of the 2019 online based on comparisons with the 2016 scripts (project (2a)). The second of these sets of comparisons involved direct scaling based on direct comparisons of online performances with each other (project (2b)). The data points in Figure 106 represent the 2019 online performances common to both scales based on the sets of comparisons, shown in the x- and y-axes. The linear relationship between the two scales was strong, and the fitted values corresponded closely to the identity line. Because the scale locations were effectively the same based on indirect and direct comparisons, the relationship indicated that it was justifiable to place 2019 online and 2019 paper performances on the same scale.

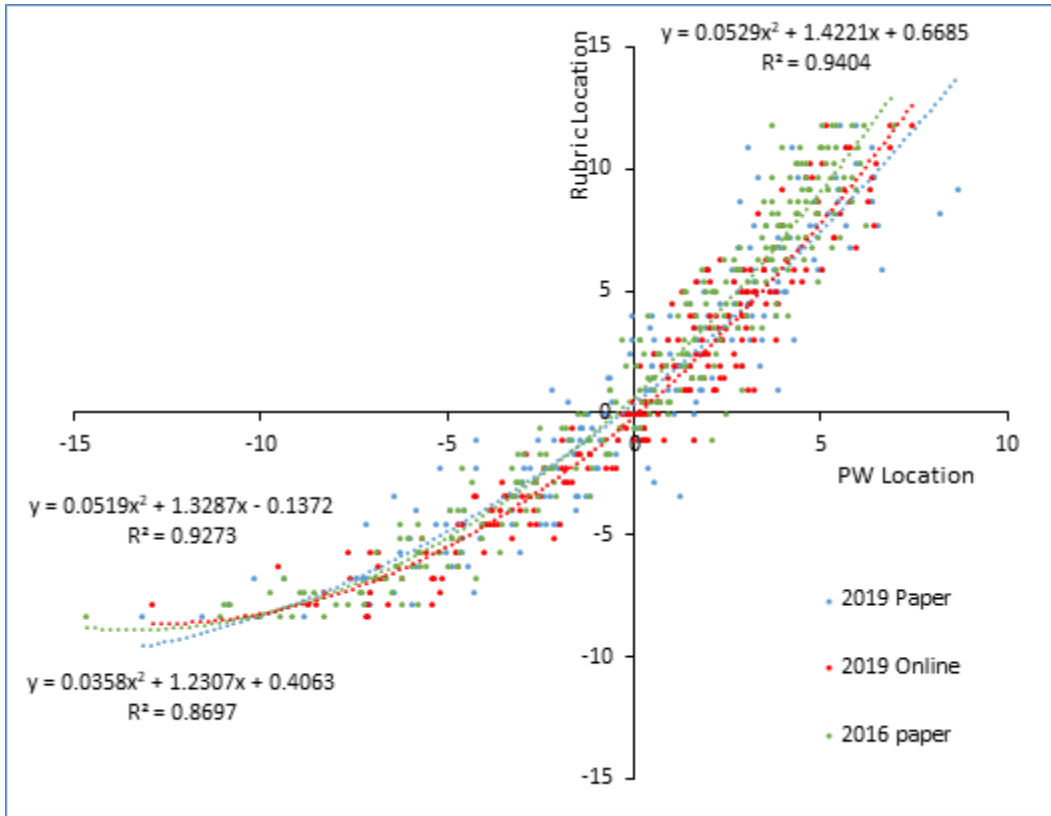


Figure 105: Scatterplot of the NAPLAN rubric and pairwise scale locations, for all 2016 and 2019 with online performances included

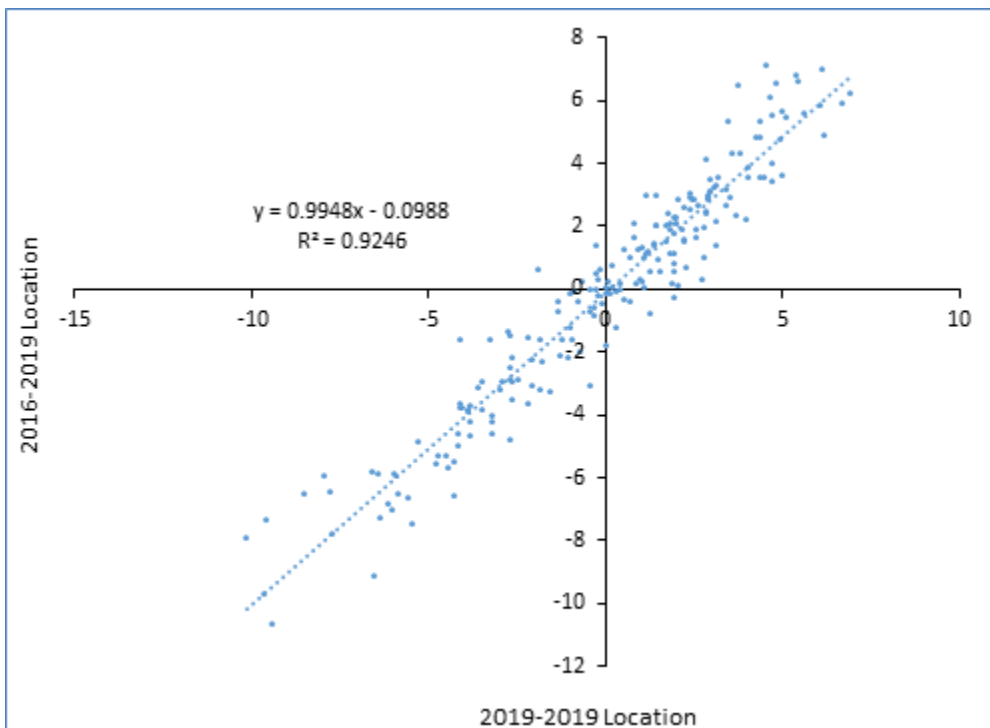


Figure 106: 2019 Online locations based on direct online-online (x-axis) and indirect online – paper (y-axis)

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 Year 5 domain mean and overall domain standard deviation calculated using the 2008 scientific sample. These are presented in Table 85.

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 85: 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

Equipercntile equating

From NAPLAN 2018, three assessment years serve as transitioning years from static paper to branched online testing. During these years, some differences in achievement distributions may be caused by differences in assessment mode or design rather than differences in student achievement. Hence, achievement distributions (mean, standard deviation, percentiles) were compared between 2019 results and three previous years, overall and by assessment mode. That is, the group of schools administering NAPLAN Online in 2019 were compared with their achievement in previous years, and the group of schools administering NAPLAN on paper in 2019 were compared with theirs.

In cases where the 2019 distributions were considered to be incomparable to previous years by NADAR members, quadratic equipercntile equating was applied to the results from one year level and one domain for the respective assessment mode. This quadratic function made the mean, standard deviation and percentiles of the online or paper group of schools equal to their respective values in 2017. Equipercntile equating was applied to the following results:

- Numeracy
 - Online: Y3, Y5, Y9
 - Paper: Y9
- Reading
 - Online: Y3, Y9
 - Paper: Y3, Y9

- Spelling
 - Online: Y5
 - Paper: Y5
- Grammar & punctuation
 - Online: Y3, Y5, Y7, Y9
 - Paper: Y3, Y5, Y7, Y9
- Writing
 - Online: None
 - Paper: None

To determine equipercetile equating parameters, seven percentiles were computed for the 2019 scale and for the 2017 scale, either for schools administering the NAPLAN tests in 2019 online or on paper. The quadratic relationship was then determined for the seven markers in the scatterplot.

For example, the 2019 Year 3 mean numeracy achievement in online schools before equipercetile equating was 403 and the standard deviation was 77. In 2017, the mean and standard deviation of the same schools was 407 and 73, respectively. A scatterplot of the 5th, 10th, 25th, 50th, 75th, 90th and 95th is presented in Figure 107. The broken line depicts the quadratic function:

$$\hat{\theta}_{17} = 100 + 0.00048 * \theta_{19}^2 + 0.56178 * \theta_{19} \quad (8)$$

After applying this transformation, the 2019 mean and standard deviation in numeracy for the Year 3 online schools were equivalent to the values of 2017 (408 and 73, respectively).

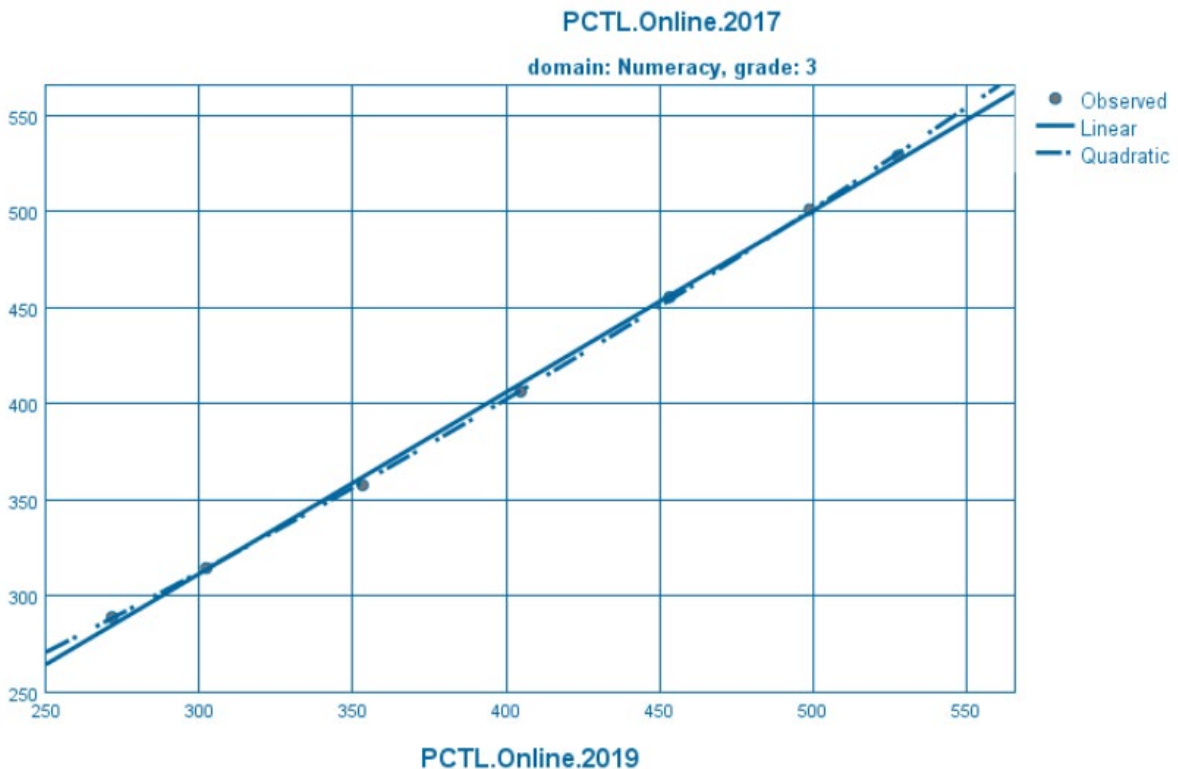


Figure 107: Scatterplot of percentiles in 2019 and 2017, Year 3 online numeracy

Equipercetile equating parameters for the relevant domains and year levels are included in Table 86. The letters refer to the parameters in the following generic formula:

$$\hat{\theta}_{17} = a + b * \theta_{19}^2 + c * \theta_{19} \quad (9)$$

Table 86: Equipercetile equating parameters

Mode	Domain	a	b	c	
Online	N3	100.49140	0.00048	0.56178	
	N5	145.33553	0.00058	0.42416	
	N9	253.30670	0.00035	0.36937	
	R3	134.01356	0.00059	0.43734	
	R9	35.61401	-0.00015	1.03585	
	S5	114.13480	0.00018	0.69006	
	G3_C	95.63552	0.00041	0.58622	
	G3_E1,2	95.63552	0.00041	0.58622	
	G3_E3	95.63552	0.00041	0.58622	
	G3_F	95.63552	0.00041	0.58622	
	G5_C	125.42115	0.00053	0.46176	
	G5_E1,2	125.42115	0.00053	0.46176	
	G5_E3	125.42115	0.00053	0.46176	
	G5_F	125.42115	0.00053	0.46176	
	G7_C	121.30717	0.00043	0.55426	
	G7_E1,2	121.30717	0.00043	0.55426	
	G7_E3	121.30717	0.00043	0.55426	
	G7_F	121.30717	0.00043	0.55426	
	G9_C	-25.55819	-0.00012	1.10480	
	G9_E1,2	-25.55819	-0.00012	1.10480	
	G9_E3	-25.55819	-0.00012	1.10480	
	G9_F	-25.55819	-0.00012	1.10480	
	Paper	N9	174.34604	0.00014	0.62647
		R3	116.59260	0.00041	0.55428
R9		-183.82113	-0.00074	1.75985	
S5		-2.52682	-0.00025	1.14646	
G3		130.49147	0.00015	0.63960	
G5		49.62513	0.00005	0.89355	
G7		10.30265	0.00005	0.92608	
G9		-32.54687	-0.00027	1.22385	

Summary of equating parameter estimates for NAPLAN 2019

In 2019, the equating procedures for the NAPLAN results were applied separately for the online and the paper tests. The combined formula for the equating procedures to place the 2019 online or paper results onto the historical scale before equipercetile equating, as described in this chapter, is:

$$\theta_{19}^* = 100 * (SF(\theta_{19} - LM) + LM + HVR - MN_{\theta_{Y5_08}}) / SD_{\theta_{All_08}} + 500 \quad (10)$$

Where θ_{19}^* is the equated 2019 achievement score, θ_{19} the original achievement score in logits, SF the scaling factor, LM the local mean, HVR the equating shift, $MN_{\theta_{Y5_08}}$ 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

$$\theta_{19}^{**} = a + b * (\theta_{19}^*)^2 + c * \theta_{19}^* \quad (11)$$

Table 87: Summary of parameters for transforming the 2019 logit scores to the NAPLAN reporting scales

Mode	Domain & year	LM	SF	HVR	MN08	SD08	a	b	c
Online	N3	0.2832	1.0293	-1.0910	0.8102	1.6652	100.49140	0.00048	0.56178
	N5	0.2744	0.8408	0.3039	0.8102	1.6652	145.33553	0.00058	0.42416
	N7	-0.0116	0.9673	1.6987	0.8102	1.6652			
	N9	-0.2495	0.9782	2.4713	0.8102	1.6652	253.30670	0.00035	0.36937
	R3	-0.1172	1.1951	0.1399	1.1629	1.4867	134.01356	0.00059	0.43734
	R5	0.1707	0.9642	1.0718	1.1629	1.4867			
	R7	0.0485	0.9742	1.7694	1.1629	1.4867			
	R9	-0.0411	1.1291	2.3102	1.1629	1.4867	35.61401	-0.00015	1.03585
	S3	0.3575	0.9877	-1.6518	0.9406	2.6241			
S5	0.4816	1.0624	0.2715	0.9406	2.6241	114.13480	0.00018	0.69006	
S7	0.5037	0.9068	1.5922	0.9406	2.6241				
S9	0.2447	0.9209	2.8255	0.9406	2.6241				
G3_C	0	1	-1.0610	1.2529	1.3605	95.63552	0.00041	0.58622	
G3_E1	0	1	0.9351	1.2529	1.3605	95.63552	0.00041	0.58622	
G3_E3	0	1	0.5930	1.2529	1.3605	95.63552	0.00041	0.58622	
G3_F	0	1	2.7923	1.2529	1.3605	95.63552	0.00041	0.58622	
G5_C	0	1	-0.2908	1.2529	1.3605	125.42115	0.00053	0.46176	
G5_E1	0	1	1.5808	1.2529	1.3605	125.42115	0.00053	0.46176	
G5_E3	0	1	1.0901	1.2529	1.3605	125.42115	0.00053	0.46176	
G5_F	0	1	3.2508	1.2529	1.3605	125.42115	0.00053	0.46176	
G7_C	0	1	0.4683	1.2529	1.3605	121.30717	0.00043	0.55426	
G7_E1	0	1	1.9858	1.2529	1.3605	121.30717	0.00043	0.55426	
G7_E3	0	1	1.1757	1.2529	1.3605	121.30717	0.00043	0.55426	
G7_F	0	1	4.0332	1.2529	1.3605	121.30717	0.00043	0.55426	
G9_C	0	1	1.2902	1.2529	1.3605	-25.55819	-0.00012	1.10480	
G9_E1	0	1	2.4629	1.2529	1.3605	-25.55819	-0.00012	1.10480	
G9_E3	0	1	2.2391	1.2529	1.3605	-25.55819	-0.00012	1.10480	

Mode	Domain & year	LM	SF	HVR	MN08	SD08	a	b	c
	G9_F	0	1	4.3111	1.2529	1.3605	-25.55819	-0.00012	1.10480
	W3	0	1	0	1.1160	3.3679			
	W5	0	1	0	1.1160	3.3679			
	W7	0	1	0	1.1160	3.3679			
	W9	0	1	0	1.1160	3.3679			
Paper	N3	0.4303	1.0562	-1.1423	0.8102	1.6652			
	N5	0.5215	0.8721	0.3031	0.8102	1.6652			
	N7	0.2937	0.9312	1.4383	0.8102	1.6652			
	N9	0.1963	0.9692	2.0940	0.8102	1.6652	174.34604	0.00014	0.62647
	R3	0.2123	1.5470	-0.0975	1.1629	1.4867	116.59260	0.00041	0.55428
	R5	0.3763	1.1438	0.8834	1.1629	1.4867			
	R7	0.0461	1.0904	1.8279	1.1629	1.4867			
	R9	0.0928	1.2354	2.2366	1.1629	1.4867	-183.82113	-0.00074	1.75985
	S3	-0.2354	1.0967	-0.8390	0.9406	2.6241			
	S5	-0.3606	1.1213	1.1828	0.9406	2.6241	-2.52682	-0.00025	1.14646
	S7	-0.2685	1.0392	2.4448	0.9406	2.6241			
	S9	-0.2185	0.9328	3.3445	0.9406	2.6241			
	G3	0.2092	1.7732	0.2136	1.2529	1.3605	130.49147	0.00015	0.63960
	G5	0.1381	1.4754	0.9844	1.2529	1.3605	49.62513	0.00005	0.89355
	G7	0.2687	1.1206	1.7765	1.2529	1.3605	10.30265	0.00005	0.92608
	G9	0.1404	1.1117	2.0757	1.2529	1.3605	-32.54687	-0.00027	1.22385
	W3	0	1	0	1.1160	3.3679			
	W5	0	1	0	1.1160	3.3679			
	W7	0	1	0	1.1160	3.3679			
	W9	0	1	0	1.1160	3.3679			

Estimating equating errors

As with all statistics, an uncertainty is associated with equating shifts. 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. This uncertainty is expressed as the equating error and is taken into account when comparing results between assessment years (see Chapter 9).

Multiple steps were involved in the equating of reading, spelling, grammar and punctuation, and numeracy. An equating error was estimated for each step. 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 108.

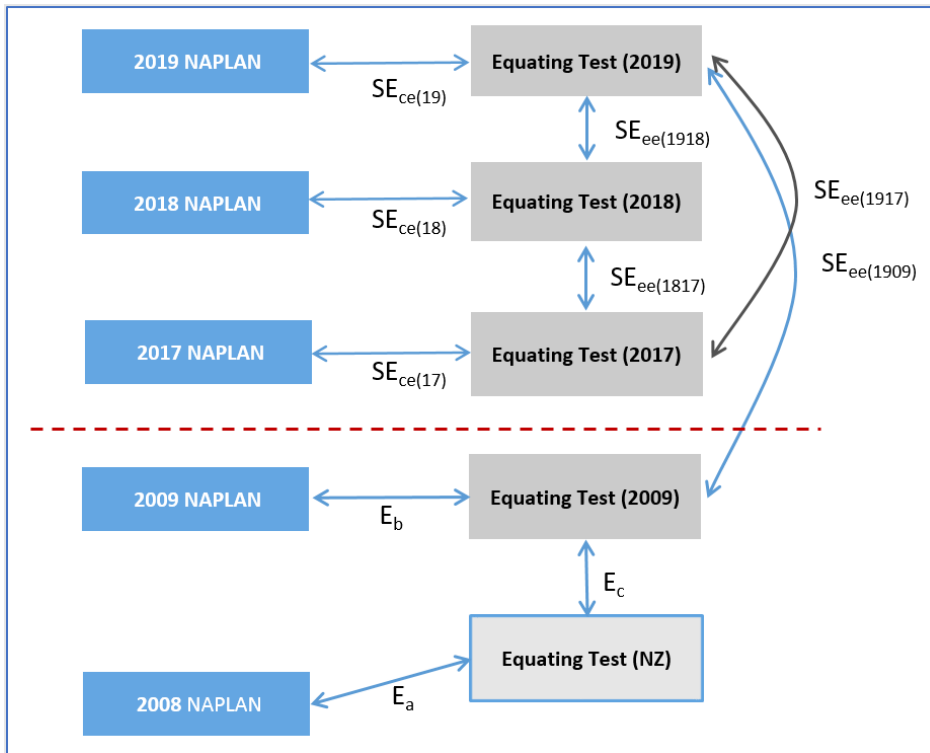


Figure 108. A schematic of the equating errors accumulated across NAPLAN administrations

For each domain and year level except writing:

- E_a is the standard error associated with equating the offshore equating test and the 2008 NAPLAN test;
- E_b is the standard error associated with equating the onshore equating test and the 2009 NAPLAN test;
- E_c is the standard error associated with equating the offshore and onshore equating tests; E_a , E_b and E_c were determined during 2009 equating process.
- $SE_{ce(19)}$ is the standard error associated with equating the NAPLAN 2019 test with the equating test (calibration to equating);
- $SE_{ce(18)}$ is the standard error associated with equating the NAPLAN 2018 test with the equating test (calibration to equating);
- $SE_{ce(17)}$ is the standard error associated with equating the NAPLAN 2017 test with the equating test (calibration to equating); and
- $SE_{ee(1918)}$ is the standard error associated with equating the 2019 and 2018 administrations of the equating test (equating to equating); and so forth.

For reporting results of NAPLAN 2019, the equating errors for equating the 2019 scale to the 2018, 2017 and 2008 scales are estimated by combining the relevant standard errors as follows:

$$SE_{2019to2018} = \sqrt{SE_{ce(19)}^2 + SE_{ce(18)}^2 + SE_{ce(1918)}^2} \quad (12)$$

$$SE_{2019to2017} = \sqrt{SE_{ce(17)}^2 + SE_{ce(19)}^2 + SE_{ce(1917)}^2} \quad (13)$$

$$SE_{2019to2008} = \sqrt{E_a^2 + E_c^2 + SE_{ee(1909)}^2 + SE_{ce(19)}^2} \quad (14)$$

In 2019, as equating procedures were carried out separately for the paper tests and for online tests, two sets of equating errors were estimated for each year level by domain. However, the final 2019 NAPLAN results were estimated based on the combined paper and online results. To simplify the estimation procedure, only one set of standard errors was applied to overall results, rather than constructing a weighted composite or applying other, simulation-based methods to combine the equating error information from the two modes. Given that both sets of equating errors were equivalent in most cases, the decision was made to apply the set of standard errors with marginally larger values in order to be conservative, with these being the standard errors determined from the paper test. Table 88 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 (4); that is, the scaling factor, the 2008 standard deviation and 100.

Table 88. Standard errors of equating

Domain	Year	Logit			Scale score		
		2019 to base	2019 to 2018	2019 to 2017	2019 to base	2019 to 2018	2019 to 2017
Reading	3	0.1019	0.0837	0.0790	6.8527	5.6320	5.3118
	5	0.0689	0.0640	0.0782	4.6333	4.3041	5.2613
	7	0.0588	0.0398	0.0394	3.9548	2.6788	2.6517
	9	0.0742	0.0563	0.0558	4.9924	3.7847	3.7533
Writing	3579	0.1270	0.1870	0.1940	3.7709	5.5524	5.7603
Spelling	3	0.1023	0.0683	0.0675	3.8985	2.6012	2.5704
	5	0.1130	0.0672	0.0709	4.3081	2.5601	2.7019
	7	0.1064	0.0564	0.0591	4.0564	2.1484	2.2539
	9	0.0914	0.0536	0.0559	3.4846	2.0425	2.1318
Grammar and punctuation	3	0.1659	0.1617	0.1292	12.1946	11.8837	9.4933
	5	0.1548	0.1039	0.1135	11.3764	7.6336	8.3412
	7	0.1073	0.0723	0.0678	7.8860	5.3166	4.9832
	9	0.0971	0.0644	0.0704	7.1371	4.7344	5.1763
Numeracy	3	0.0741	0.0693	0.0527	4.4474	4.1636	3.1637
	5	0.0651	0.0400	0.0411	3.9104	2.4015	2.4693
	7	0.0501	0.0387	0.0408	3.0081	2.3248	2.4490
	9	0.0518	0.0388	0.0337	3.1089	2.3317	2.0218

* 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 2019. This is further explained in Chapter 9.

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

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

Chapter 8: NAPLAN proficiency bands

The main feature of the Rasch model is the placement of items and students on the same scale. A student with an achievement score equal to the difficulty of an item has 50 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 89 shows the cut points between bands (lower bound) in scale scores and in logits.

Table 89: 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 90 to Table 93 for each domain.

Table 90: 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.

Proficiency band	Numeracy skills and knowledge
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 91: 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.

Proficiency band	Reading skills and knowledge
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 92: 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.

Proficiency band	Writing skills and knowledge
Band 2	Shows audience awareness by using common text elements, for example, begins writing with <i>Once upon a time</i> ; or <i>I think ... because ...</i> Uses some capital letters and full stops correctly. Correctly spells most simple words used in the writing. Some other one- and two-syllable words may also be correct.
Band 1	Writes a small amount of simple content that can be read. May name characters or a setting; or write a few content words on a topic. May write some simple sentences with correct word order but full stops and capital letters are usually missing or incorrect. Correctly spells a few simple words used in the writing.

Table 93: 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.

Proficiency band	Conventions of language skills and knowledge
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 bands for each level were regarded as achieving below the National Minimum Standard (NMS).

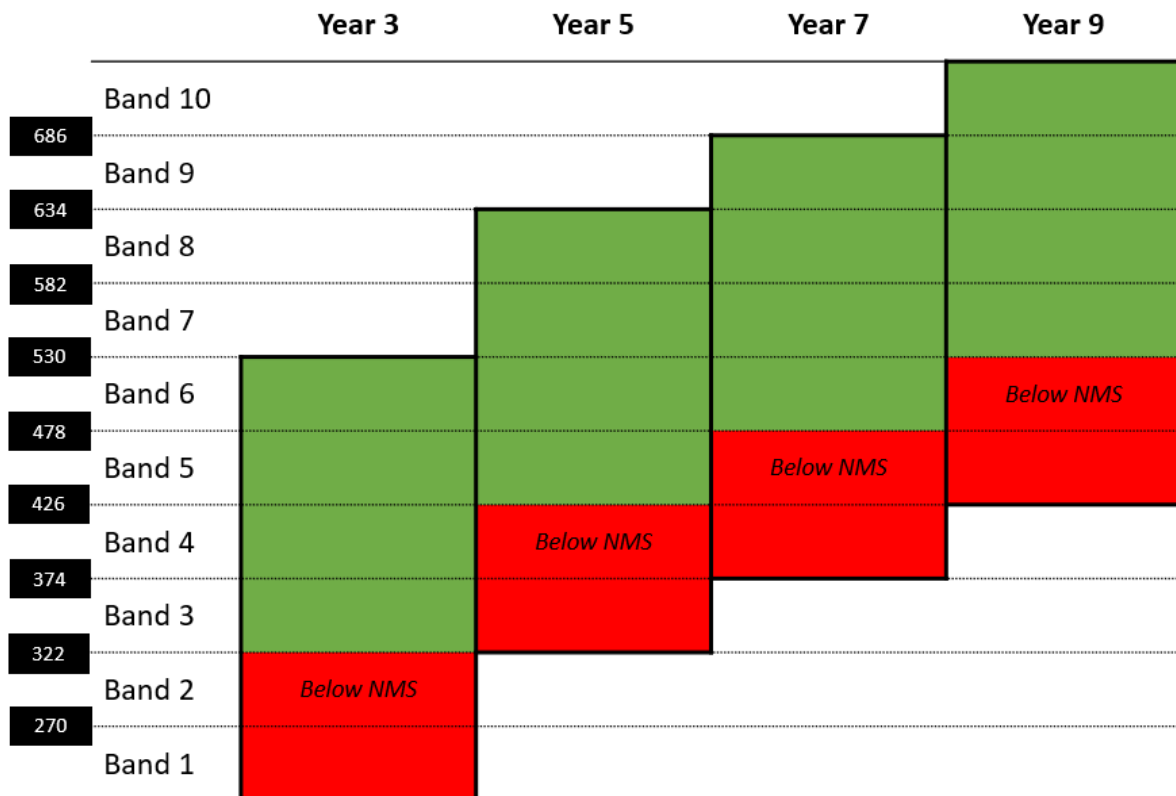


Figure 109: Schematic picture of proficiency bands by year levels

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 90). 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 90) and is ready to learn the skills and concepts described for Band 7.

Chapter 9: Reporting of national results

NAPLAN produces several reports each year. The student and school summary report (SSSR)⁵ is a preliminary report with student and school level results for school staff. The individual student report (ISR)⁶ 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 researcher. This chapter describes analysis for the national report. The last report is the website *My School*⁷ with results for individual schools and accessible for the general public.

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 of each plausible value. The final statistic was the average of the five results. Plausible values should never be averaged at the student level. Only the results should be averaged. The formal notation for this is

$$\theta = \frac{1}{M} \sum_{i=1}^M \theta_i \quad (15)$$

with θ being any type of population statistic (mean, standard deviation, percentage) and M being the number of plausible values per student and domain.

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 7. The other two types of errors are explained in this chapter.

⁵ www.nap.edu.au/docs/default-source/default-document-library/how-to-interpret-the-sssr.pdf?sfvrsn=10

⁶ www.nap.edu.au/results-and-reports/student-reports

⁷ 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 2019. 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 2019, 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 110. The sampling error is equal to the square root of the sampling variance.

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

Figure 110: Examples in SPSS and SAS for estimating sampling variance

Measurement error

Plausible value methodology enables the computation of the uncertainty in the estimate of θ due to the lack of precision of 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 perfect, the five sets of statistics would not be identical. The measurement variance is estimated as:

$$B_M = \frac{1}{M-1} \sum_{i=1}^M (\theta_i - \theta)^2 \quad (16)$$

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:

$$V = U + \left(1 + \frac{1}{M}\right) B_M \quad (17)$$

$$SE = \sqrt{V} \quad (18)$$

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 2019 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 2019 data; for example, between male and female students or between jurisdictions. The second type of comparison was between 2019 results and results from earlier assessment years. The first type of difference 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:

$$SE_{DIFF} = \sqrt{SE_A^2 + SE_B^2} \quad (19)$$

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 2019 and 2018, we need to add the equating error between these two years, $SE_{2019to2008}$, to the calculation in the following way:

$$SE_{DIFF} = \sqrt{SE_{A19}^2 + SE_{A18}^2 + SE_{2019to2008}^2} \quad (20)$$

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 2019 NAPLAN with 2018, 2017 and the base year for each domain and year level are included in Chapter 7.

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 again. 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 2019 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:

$$g = \frac{m_2 - m_1}{s_p} \quad (21)$$

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

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (22)$$

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:

$$g_{unbiased} = g_{biased} \left[1 - \frac{3}{4(n_1 + n_2 - 2)} \right] \quad (23)$$

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

$$OR = \frac{p_E q_C}{q_E p_C} \quad (24)$$

$$d_{Cox} = \frac{L(OR)}{1.65} \quad (25)$$

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.

As a result of this revised standard, the reporting of NAPLAN trends relating to geolocation and any associated subgroups were referenced against the base year in which the revision took place – that is, 2016.

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.
- 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 Paedagogiske 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?sfvrsn=2>

Appendix B: Item analysis details

Paper: <https://nap.edu.au/docs/default-source/default-document-library/appendix-b-itanal-paper-tests.pdf?sfvrsn=2>

Online: <https://nap.edu.au/docs/default-source/default-document-library/appendix-b-itanal-online-tests.pdf?sfvrsn=2>

Appendix C: Item summary tables

<https://nap.edu.au/docs/default-source/default-document-library/appendix-c-item-calibration-results.pdf?sfvrsn=2>

Appendix D: Item characteristic curves

<https://nap.edu.au/docs/default-source/default-document-library/appendix-d-icc.pdf?sfvrsn=2>

Appendix E: Item–person maps

<https://nap.edu.au/docs/default-source/default-document-library/appendix-e-variable-map.pdf?sfvrsn=2>

Appendix F: Gender DIF

<https://nap.edu.au/docs/default-source/default-document-library/appendix-f-difplot.pdf?sfvrsn=2>

Appendix G: LBOTE DIF

<https://nap.edu.au/docs/default-source/default-document-library/appendix-g-difplot.pdf?sfvrsn=2>

Appendix H: ATSI DIF

<https://nap.edu.au/docs/default-source/default-document-library/appendix-h-difplot.pdf?sfvrsn=2>

Appendix I: DIF summary tables

<https://nap.edu.au/docs/default-source/default-document-library/appendix-i-dif-summary-tables.pdf?sfvrsn=2>

Appendix J: Jurisdictional DIF for writing

<https://nap.edu.au/docs/default-source/default-document-library/appendix-j-writing-taa.pdf?sfvrsn=2>

Appendix K: Horizontal link item comparisons

https://nap.edu.au/docs/default-source/default-document-library/appendix-k_horizontal-link-item-comparisons.pdf?sfvrsn=2

Appendix L: Vertical link item comparisons

<https://nap.edu.au/docs/default-source/default-document-library/appendix-l-verticalequating.pdf?sfvrsn=2>

Appendix M: Measurement errors in individual achievement scores

It is often suggested that the online branched test gives more precise estimates of individual student achievement than the linear paper test. Estimates can be more precise in two ways: they include less bias and they include less uncertainty. The following example illustrates why individual student achievement scores include less uncertainty when they are assessed with the online branched test than with the linear paper test. NAPLAN score equivalence tables were used for this example.

Imagine a Year 5 student who has a NAPLAN score of 416 for numeracy (WLE estimate). If this student takes the online numeracy test, it will most likely follow the easiest pathway: ABC. This pathway perfectly matches their ability and this student is most likely to have a raw score of 21 out of 42 items, which is equal to 50 per cent correct. The measurement error associated with this score is 21 NAPLAN scores. Had this student taken the paper test—which is more difficult than the online ABC pathway—their expected raw score is 13 out of 42 items, which is equal to 31 per cent correct. This score is associated with a measurement error of 24 which is larger than the measurement error of 21 for the online test.

To test if this difference existed for real students, the uncertainty in numeracy achievement scores of Year 5 students was compared between students who took the online NAPLAN test and who took the paper NAPLAN test in 2018. Figure 1 shows student achievement scores on the horizontal axis and the uncertainty in achievement scores on the vertical axis. Each dot is a student with blue dots for students who were assessed online and red dots for students who were assessed on paper.

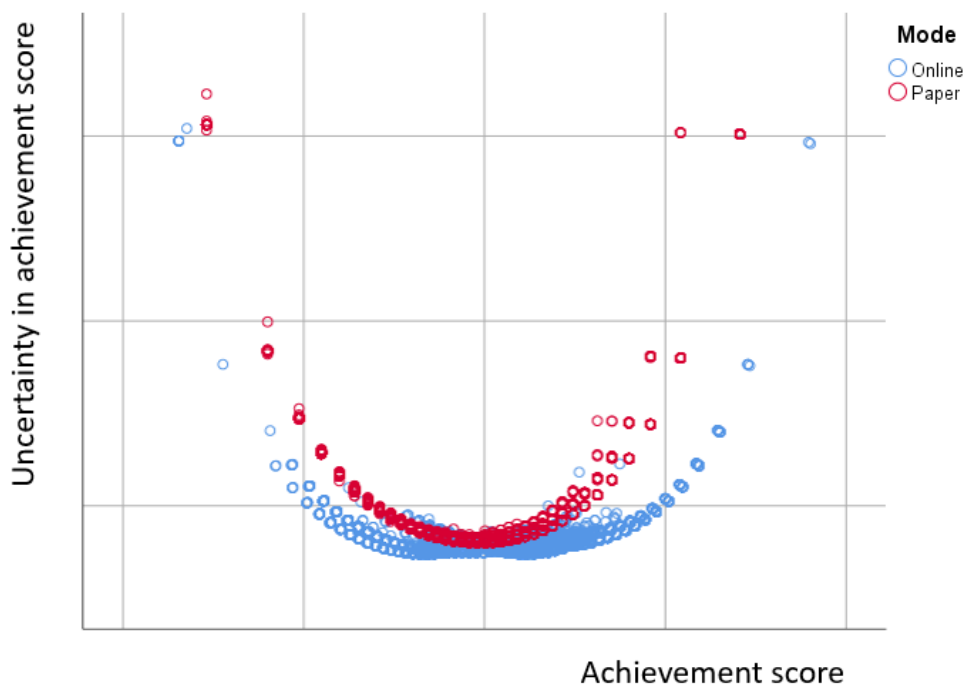


Figure 111: Scatterplot of the location of and the uncertainty in WLE achievement scores by assessment mode

As expected, the uncertainty in achievement scores was larger for students who achieved very high or very low compared to their peers. This suggests that measuring student achievement is more precise if the test is well targeted to the student's ability. Furthermore, and consistent with the fictional example above, the uncertainty in the online achievement scores (blue dots) was smaller or equal in size than in the paper achievement scores (red dots) at each achievement level. Finally, it can also be seen that the range of achievement scores was larger online than on paper, because the paper achievement estimates were affected by floor and ceiling effects (this is only true for WLE achievement scores!). This suggests that for this type of achievement scores the extreme ends of the scale include less bias when administering NAPLAN online using a branched test design.