Appendix E:
Modelling of student outcomes

We performed multi-level modelling to model students' learning (from year 3 to year 9) for various subgroups (Aboriginal, EAL, disadvantage) in each NAPLAN domain. The purpose of this analysis is to understand the rate of learning and how these vary by subgroup.

Similar modelling was performed on teacher judgement scores as a comparison. The results from both models are mostly consistent.

Analysis in this report is primarily based on NAPLAN due to some limitations in teacher judgement data and clear representation of the analysis.

NAPLAN

Dataset

Ignoring repeats, each student takes up to 4 NAPLAN tests for each domain, which correspond to tests taken in years 3, 5, 7 and 9.

We based our analysis on the groups of students denoted with coloured cells in the table below.

To estimate learning, we made sure that each student in the 2012, 2013 and 2015 reference cohorts had 4 data points, while students in all other reference cohorts had 3 data points.

Figure E1: Illustration of student groups with 4 data points of NAPLAN test results

Reference cohort	2010	2011	2012	2013	2014	2015	2016	2017
Year 3	2010	2011	2012	2013	2014	2015	2016	2017
Year 5	2012	2013	2014	2015	2016	2017	2018	2019
Year 7	2014	2015	2016	2017	2018	2019	2020	2021
Year 9	2016	2017	2018	2019	2020	2021	2022	2023

Source: VAGO.

Data quality

Students self-report their demographic information. It is natural for some demographic information to change, such as EAL status or disadvantage level.

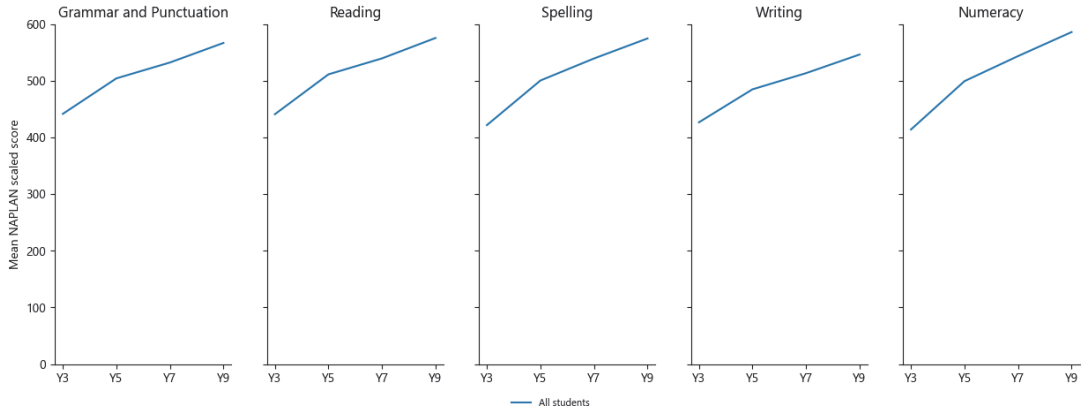
We observed a large proportion of students who were recorded as both Aboriginal and non-Aboriginal throughout their education. VAGO treated any student as Aboriginal if they were recorded as Aboriginal at any time.

Model details

We ran both random intercepts and random slopes multi-level models using the nlme package in R for NAPLAN scores where the levels are students in a particular year level and school.

Based on the following chart, NAPLAN scaled scores do not appear to grow strictly linearly. We tested various functional forms as described in the sensitivity analysis section and chose the model in Figure E2.

Figure E2: Mean NAPLAN scaled score (baseline)



Source: VAGO, using department data.

We created a CENTRED_YEAR_LEVEL variable to model 2-year improvement in NAPLAN scores. This aligns with the 2-year NAPLAN cycle. CENTRED_YEAR_LEVEL maps YEAR_LEVEL values 3, 5, 7, 9 to 0, 1, 2, 3 respectively.

The coding for the SUBGROUP categorical variable is shown in Figure E3.

Figure E3: Coding for the SUBGROUP categorical variable

The ... subgroup of level ... corresponds to ...	0	1	2
Aboriginal	Non-Aboriginal	Aboriginal	N/A
EAL	Non-EAL	EAL	N/A
Disadvantage level	Non-disadvantaged	Disadvantage level 2	Disadvantage level 1

Source: VAGO.

To explain student learning rate, we used a random slopes regression of NAPLAN score on subgroup, centred year level and an interaction term.

Limitations

Since we only considered NAPLAN data from 2012 to 2022, and NAPLAN did not run in 2020, only reference cohorts 2012, 2013 and 2015 have 4 test results per student.

We did not consider school-level effects in the model.

Sensitivity analysis

To account for the unbalanced dataset, we tested both random intercepts and random slopes models for the full dataset, as well as for the 2012, 2013 and 2015 reference cohorts only.

We also ran the models for each reference cohort individually. Results from all models were mostly consistent with each other.

We also performed sensitivity analysis for the functional form of the growth trajectory. Results suggested the linearity assumption is suitable and conclusions were consistent across the models with different functional forms.

Analysis and charts in this report are taken from the random intercepts model for the unbalanced dataset including reference cohorts 2010 to 2017. The random intercepts and random slopes model produced similar results. We chose the random intercepts model as it allows for fairer comparison with the teacher judgement analysis.

Teacher judgement

Dataset

We used semester 2 teacher judgement scores recorded between 2017 and 2022 to model the learning rate of students from year F to year 10.

We did not consider semester one scores because they were missing in 2020. We also did not consider teacher judgement scores prior to 2017 because the curriculum changed.

We joined the teacher judgement dataset with the August census dataset to include students' demographic information.

Our analysis is based on the groups of students in the coloured cells in the table below.

Figure E4: Illustration of student groups with teacher judgement scores between 2017 and 2022

Reference cohort	2012	2013	2014	2015	2016	2017	2018
Year F	2012	2013	2014	2015	2016	2017	2018
Year 1	2013	2014	2015	2016	2017	2018	2019
Year 2	2014	2015	2016	2017	2018	2019	2020
Year 3	2015	2016	2017	2018	2019	2020	2021
Year 4	2016	2017	2018	2019	2020	2021	2022
Year 5	2017	2018	2019	2020	2021	2022	
Year 6	2018	2019	2020	2021	2022		
Year 7	2019	2020	2021	2022			
Year 8	2020	2021	2022				
Year 9	2021	2022					
Year 10	2022						

Source: VAGO.

We performed the same preprocessing steps in the NAPLAN analysis to the teacher judgement analysis. We also removed part-time students and students who repeated grades from the analysis.

Limitations

Since we analysed 6 years of teacher judgement data between 2017 to 2022, we cannot observe an individual student's learning for the 11-year period from year F to year 10.

While the dataset used is mostly balanced for each reference cohort (there are 5 or 6 observations per cohort), it is unbalanced for each year level.

For example, there are 6 observations for students in years 4 and 5, but only one observation for students in year 10.

Sensitivity analysis

Since there appears to be little variation across students in teacher judgement learning trajectories, we only used random intercepts models for the 7 reference cohorts of students coloured above.

To account for the unbalanced dataset, we also ran the models for each reference cohort individually. Results from models were mostly consistent with each other.