Predicting Students' Future NCEA Attainment Using Census and Administrative Data

Analytical Paper 18/01

March 2018

THE TREASURY
Kaitohutohu Kaupapa Rawa

New Zealand Government

DISCLAIMER

The views, opinions, findings and conclusions or recommendations expressed in this report are strictly those of the authors. They do not necessarily reflect the views of the New Zealand Treasury, Statistics New Zealand or the New Zealand Government. The New Zealand Treasury and the New Zealand Government take no responsibility for any errors or omissions in, or for the correctness of, the information contained in this Analytical Paper.

The results in this report are not official statistics – they have been created for research purposes from the Integrated Data Infrastructure (IDI) managed by Statistics New Zealand.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation, and the results in this paper have been confidentialised to protect these groups from identification.

Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form or provided to Inland Revenue for administrative or regulatory purposes.

Any person who has had access to the unit-record data has certified that they have been shown, have read and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes and is not related to the data's ability to support Inland Revenue's core operational requirements.

ANALYTICAL PAPER 18/01

Predicting Students' Future NCEA Achievement Using Census and Administrative Data

MONTH/YEAR

March 2018

AUTHOR

Sarah Crichton, Sylvia Dixon and Christopher Ball

ISBN (ONLINE)

978-0-947519-68-1

URL

Treasury website at March 2018:
http://www.treasury.govt.nz/publications/research-policy/ap/2018/18-01

NZ TREASURY

New Zealand Treasury
PO Box 3724
Wellington 6008
NEW ZEALAND

Email information@treasury.govt.nz

Telephone 64-4-472 2733

Website www.treasury.govt.nz

PURPOSE OF THE ANALYTICAL PAPERS SERIES

The Treasury's aim in publishing the Analytical Papers series is to make this analysis available to a wider audience and to inform and encourage public debate, with the ultimate aim of informing our policy advice.

Analytical Papers are commissioned as part of the Treasury's core function in developing and providing advice to Ministers. They include work undertaken by staff at the Treasury or other government departments, as well as work undertaken for the Treasury by external researchers or consultants.

Analytical Papers do not themselves represent policy advice.

Contents

List of figures

List of tables

Introduction and summary

Introduction

The research outlined in this paper was undertaken in 2016 as background work to inform an official review of the decile-based allocation of operational school funding.

The current decile-based component of school funding represents about 3% of the total funding given to state primary and secondary schools. Its purpose is to increase the funding of schools that are located in low socio-economic areas, where students tend to have higher learning needs, and parents are less able to provide financial or other support for the school's activities.

In 2016, the Government put forward a proposal to shift to a 'student risk' based approach for allocating this additional funding in future. The proposal was to allocate the funding to schools according to their students' average predicted risk of not completing NCEA level 2, rather than the socio-economic status of the local population surrounding the school. This would require the use of data on individual students, rather than population-level data used in the decile funding formula.

The purpose of the research outlined in this paper was to explore and evaluate alternative methods for predicting students' future NCEA achievement.

The paper considers the following issues:

1. The choice of data sources and variables to be used in predicting students' risk of not completing NCEA level 2.
2. The predictive accuracy of models that only use administratively sourced variables that are currently available in the Integrated Data Infrastructure (IDI) compared with models drawing on other data sources such as the Census.
3. The distribution of students with different levels of predicted risk across schools and deciles.
4. How schools would be 're-ranked' if student-based risk measures were used in place of the current meshblock-based measures that underpin school deciles.
5. What further work is required to improve our capacity to predict students' risk of not completing NCEA level 2.

There are three parts to the paper.

Part 1 focuses on the first two issues. It uses both Census and administrative data to estimate and assess a range of different models of NCEA level 2 achievement, estimated at age 14. To do this, it uses a sample of youth who were born in 1998 and have linked data available from both administrative sources and the 2013 Census (N=approximately 41,000).[1] The models are estimated using Census data as at March 2013 and

[1] The 1998 birth cohort is the first cohort for whom information on birth parents is available for all New Zealand-born children. The birth parent IDs are used to derive information from the administrative data on parental characteristics including mother's age at first birth, parental income and offending history.

administrative data over the period from birth to the 14th birthday. The predictive accuracy of models using administrative data, Census data or a combination of both are compared.

Part 2 focuses on issues 3 and 4. It uses administrative data only for a larger and more inclusive sample of youth who were born in 1998 (N=approximately 51,000,) to estimate a series of models of NCEA level 2 achievement at each year of age (from 5 to 17), using data covering the period from birth to the most recent birthday. The coefficients obtained from these models are then used to derive predictions of the likelihood of completing NCEA level 2 for all students who were enrolled at state-funded schools in mid-2015. These predictions are in turn used to investigate the profile of high-risk students in 2015, their distribution across schools and the effects of substituting student-risk measures to define school deciles on the relative ranking of individual schools.

Part 3 discusses the further work that should be undertaken to refine the modelling if there is a decision to proceed further.

Two Excel workbooks accompany the paper. The first, labelled Part 1 and Appendix 1, contains the tables and figures set out in part 1 of this paper and some additional results. The second, labelled Part 2 and Appendix 2, contains the tables and figures given in part 2 and some additional results.

Key findings from part 1: Variable choice and model predictive power

Both Census and administrative data on children and their parents were used to predict NCEA level 2 attainment, based on characteristics recorded either at the time of the Census or in the years leading up to the child's 14th birthday. A large number of characteristics were included in the modelling exercise.

We found that a relatively small number of variables, around 10, were important in the model, while another 15 or so variables were included in the model but added very little in terms of predictive power.

The most important variables in the combined Census and administrative data model were:

- proportion of time supported by a benefit as a child
- mother's qualifications (as recorded in the Census)
- gender
- having a CYF care and protection notification
- mother's smoking status (Census)
- mother's age at first birth
- father's qualification (Census)
- being of Asian ethnicity
- having a CYF youth justice history
- number of addresses in the last 5 years
- being of Pacific ethnicity

- father's offending and sentence history.

A predictive model that included only administrative data had only slightly less predictive power than a model based on both Census and administrative data. The most important variables in the administrative data model were:

- proportion of time supported by a benefit as a child
- having a CYF care and protection notification
- gender
- mother's age at first birth
- father's offending and sentence history
- being of Asian ethnicity
- having a youth justice history
- having a mother who had no qualification (as recorded in benefit and tertiary data)
- proportion of time spent overseas as a child
- mother's average earned income over the past 5 years
- being of Māori ethnicity
- mother's average earned income over the past 5 years
- number of addresses recorded over the past 5 years
- migrant category/New Zealand born.

Both the Census-based and the administrative data models were significantly better at predicting a child's future NCEA level 2 attainment than the meshblock-level measures that are currently used to allocate decile funding.

To illustrate the predictive accuracy obtained, when we used the administrative data-based model and split the sample at the 15% / 85% threshold (mirroring the actual NCEA level 2 attainment rate for this sample), 82.7% of all students were correctly classified as achievers or non-achievers, 41% of those who didn't achieve NCEA level 2 were correctly identified (sensitivity), and 89% of those who did achieve the qualification were correctly identified (specificity).

There was no material difference between the predictive accuracy of the models that only used administrative data and those that used a larger set of variables that we took from the Census, including mother's highest qualification, mother's smoking status and father's highest qualification. Using the latter more comprehensive models, 83.1% of all students were correctly classified, 43% of those who didn't achieve the qualification were correctly identified (sensitivity) and 89% of those who did achieve it were correctly identified (specificity).

If we moved to a 25% / 75% threshold, there was no difference in the overall percentage who were correctly classified (77.4%), while 59% of those who didn't achieve the qualification were correctly identified (sensitivity), and 80% of those who did achieve it were correctly identified (specificity).

Key findings from part 2: Applying the risk estimation methods to the current student population

The purpose of part 2 was to explore the way in which data currently held in the IDI could be used to obtain predicted risk scores for all children in the current school population and to derive school-level measures of average student risk.

Most predictive variables

A more inclusive sample of youth born in 1998 was used in part 2, because we no longer had to exclude youth who did not have a linked Census record or those whose Census night caregivers did not link to the IDI spine. This enlarged the model estimation sample by 16% and led to some minor changes to the set of variables that were found to be most important for predicting NCEA level 2 attainment (in a model based solely on administrative data).

The main change was that being of Māori ethnicity was among the top six predictors at age 14. Other variables that were found to be important in part 1 (proportion of time on benefit as a child, CYF care and protection notifications, gender and mother's age at first birth) continued to be selected on the basis of highest predictive power.

The relative importance of different variables changes gradually at different years of age. Parental variables become less important at older ages, while measures of the child's care and protection history and youth justice history become more important.

Distribution of predicted risk scores

When the risk models are applied to the entire population of students in publicly funded schools in mid-2015, we find substantial gender and ethnic variations in the level of predicted risk, with male students having higher average risk scores than female, Māori and Pacific students having higher risk scores than European, and Asian students having lower risk scores than European. This reflects the demographic variation in actual NCEA level 2 achievement rates.

The proportion of students who are classified as 'high risk' (compared with all other students) falls rapidly at ages 16–19, due to higher-risk students leaving school. This means that funding models that are based on student risk will tend to allocate relatively less funding to upper-secondary school students and to secondary schools, unless risk pools are segmented, for example, by year of age or broad level of schooling.

School-level predicted risk measures and how they differ from the current deciles

Student-level measures of risk can be averaged over the students in each school to get school-level measures.

Within each of the current school deciles, there is considerable variation across schools in the mean level of the resulting risk measures.

Our estimates suggest that about two-thirds of schools would move to a different decile if they were ranked by the proportion of their students who are in the top quartile of the student risk distribution (or any other measure of their students' relative risk). Using one

version of this simulation, 32% stayed in the same position, 41% moved up or down one decile, 19% moved up or down two deciles and 10% moved three or more deciles.

Schools with a relatively high proportion of Māori students (at least 25%) were more likely to move downwards (to a lower decile) than schools with fewer Māori students. This reflects the role of Māori ethnicity in predicting a higher risk of non-attainment in our models. In contrast, schools with a relatively high proportion of Pacific students (at least 25%) were far more likely to move upwards than other types of school. This may be due to the fact that these schools tend to be located in low socio-economic areas, but their students have attainment rates that are better on average than those of other students in low decile schools.

If we use a single risk distribution covering students at all ages, ranking schools by measures of their students' predicted risk will lead to many primary schools moving to a lower decile and many secondary schools moving to a higher decile. This pattern of change could be avoided by segmenting the risk pools so that primary-aged children are only compared with other primary-aged children, and the same is true for secondary-aged children. There are a number of different ways in which schools could be segmented for funding purposes.

The attainment risk models were re-estimated with the gender and ethnicity variables excluded, and these alternative models were also used to predict each student's risk and analyse the distribution of risk across schools. A slightly larger proportion of schools remained in the same decile as currently (34% rather than 32%). There was also less movement of schools with relatively large proportions of Māori or Pacific students down or up the decile distribution, but the simulated movements were still substantial. This reflects the fact that other explanatory variables in the risk prediction models will tend to pick up (some of) the effects of gender and ethnicity even if gender and ethnicity are left out of the model.

Key recommendation from part 3: Further work needed to refine the modelling

Further work should be done in the following areas:

- Making the model estimation sample more comprehensive by obtaining achievement data for the students who were at schools offering non-NCEA qualifications.
- Improving the coverage and accuracy of the existing explanatory variables.
- When qualification attainment data for 2016 is available next year, updating the dependent variable to represent a final measure of the NCEA level 2 attainment (by age 18) for the 1998 birth cohort.

Increasing the coverage of the estimation sample

Students who were at schools offering non-NCEA qualifications were excluded from the model estimation sample because we don't have data on their international qualification attainment. If the relevant data can be put in the IDI, those students should be brought back in. Because they tend to be high-achieving students, adding them to the estimation sample probably won't have a large effect on the model results, however.

Improving the coverage and accuracy of the existing explanatory variables

Our administrative data models used the identities of children's birth parents to construct the parent variables, such as the birth parent's earned income and offending history. The 1998 birth cohort is the first for which information on birth parents is available for all New Zealand-born children.

However, the school-aged population includes many children who were not born in New Zealand. In addition, at any point in time, many children will be living with caregivers who are not the parents named on their birth certificate.

We explored the latter issue using the 1998 cohort linked Census sample. Children's birth certificate parents and their Census night caregivers were compared.

- For 43% of children, both Census night caregivers were the same as the birth parents listed on the birth record.
- For another 35% of children, one Census night caregiver was the same as one of the parents listed on the birth record.
- For 22% of children, there was no Census night caregiver who matched the birth certificate parents. This includes about 17% of children who had no New Zealand birth record.

When caregiver information contained in three sources – birth registrations, immigration applications and benefit data – was combined and the results compared to the Census night caregivers, the results were somewhat better:

- For 68% of children, both Census night caregivers were correctly identified using the administrative sources.
- For 24% of children, one Census night caregiver was correctly identified.
- For 8% of children, no Census night caregiver(s) was correctly identified.

We conclude that information on the identities of current caregivers that is contained in the migration and benefit data should be included in future work.

While including this information seems important in principle, the work we've done so far suggests it will raise the predictive power of the models only slightly. The impact is likely to be quite limited because of the range of information that is already included about migrant children (such as migrant category, source country and time in New Zealand) and children supported by a benefit (such as the duration and type of benefit). In addition, we have already found that adding information collected in the Census on current caregivers' qualifications and occupation did not materially improve the predictive power of the attainment model at age 14, compared with using administrative data for birth parents only.

Updating the dependent variable used in the modelling

Our qualification achievement data currently ends in 2015, the year when youth in the 1998 birth cohort turned 17. We used a modified dependent variable to account for the

fact that some children will have returned to school in 2016 and gained NCEA level 2. Our dependent variable is a measure of not obtaining NCEA level 2 by 2015 *and* not returning to school in 2016.

When attainment data becomes available for 2016, the dependent variable should be updated to bring in this additional information.

The value of collecting additional data from parents

The work summarised in this paper has shown that students' risk of not completing NCEA level 2 can be predicted with moderate accuracy using the administrative data sources that are currently available in the IDI. Comparing models, we have found that their predictive power is only slightly altered by adding or subtracting different measures of child characteristics and parent characteristics. This suggests that improving the quality of these existing measures by, for example, requiring schools to record the names and birth dates of students' current caregivers so that the correct parental IDs are used, would probably not lead to much improvement in the predictive accuracy of the models.

The predictive accuracy is more likely to be significantly improved if we were to bring in different types of data that are more strongly correlated with future achievement, such as the child's own achievement results at younger ages.

1 Models using Census and administrative data

1.1 Introduction

The purpose of this part of the paper is to investigate:

- the choice of data sources and variables to be used in predicting students' risk of not completing NCEA level 2
- the predictive accuracy of models that only use variables that are currently available in the Integrated Data Infrastructure (IDI), compared with models drawing on other data sources.

The analysis described in this section uses both Census and administrative data for a sample of youth who were born in 1998 and have linked data from both administrative sources and the 2013 Census (N=approximately 41,000). Different models of NCEA level 2 achievement are estimated using Census data as at March 2013 and administrative data over the period from birth to their 14th birthday. The predictive accuracy of the models using either administrative data, Census data or a combination of both is compared.

For data availability reasons, the dependent variable that was used in the modelling is slightly more complex than just 'failure to complete NCEA level 2'. We identify members of the 1998 birth cohort in the 2013 Census dataset and then find those who (a) did not achieve NCEA level 2 by the end of 2015 (the year they turned 17) and (b) were not enrolled in the first quarter of 2016. This represents around 15% of the modelling cohort. This outcome measure is used as a proxy for 'not attaining NCEA level 2 by the 18th birthday'. For the rest of this analysis, we will use 'not attaining NCEA level 2' to refer to this outcome.

1.2 Study population

The starting population comprises youth in the 2016 Census usually resident population who were born in 1998. On Census night, this population comprised 55,494 individuals who were aged between 14 years and 2 months and 15 years and 2 months. A number of restrictions were applied to get the final study population. These are described in Table 1.1.

Table 1.1. Criteria used to select the study population

Restriction	Motivation	Number and percentage lost at each step
Linked to the IDI spine	Necessary for data linking to data sources in IDI	2,322 (4.2%)
Have an IDI spine birth year of 1998 in addition to the Census birth year of 1998	Eliminates potential linkage error between Census and the spine	219 (0.4)
Be alive at the end of the month Census was held	Eliminates potential linkage error between Census and the spine	18 (0.0%)
Be either in a sole-parent (Census) family with the parent linked to the spine or in a two-parent (Census) family with both parents linked to the spine	Ensures that the administrative data for Census night caregivers can be derived	5,154 (9.3%)
Have a school enrolment record in at least one year	Ensures that the student's school achievement is known	1,494 (2.7%)
Not an international fee-paying student, on a 28-day waiver or on an extended 28-day waiver (on Census night)	Ensures the student is a domestic rather than international student	195 (0.4%)
Not attending a private school on Census night	Limits the analysis to those who are most affected by the current decile funding model	2,247 (4.0%)
Not enrolled in a school that offers non-NCEA qualifications	Ensures qualifications gained at school are recorded	2,049 (3.7%)
Have a geocoded usual residence address from Census night	Ensures that the meshblock Census measures used in the current decile funding model are known	90 (0.2%)
Not overseas for more than 6 months during the year they were aged 15 or 16	Ensures that we observe their outcomes (i.e. exclude those who may have gained qualifications while overseas)	855 (1.5%)
Miscellaneous data quality issues		114 (0.2%)

Additionally, 114 records (0.2%) were removed because of various data quality concerns. The final analysis population has 40,737 observations, noting that there is substantial overlap in the people removed by each of the criteria above. This is 73.4% of the base population or 76.6% of the 1998-born Census population that was linked to the IDI.

1.3 Information available on parents and caregivers

This section briefly describes the information that is available in the Census and the various administrative sources on the parents and caregivers that are associated with a child over time.

In the case of the Census, the characteristics of the adults identified as being in a parent role on Census night were used in this analysis. In the case of the administrative data, a number of variables for a child's birth parents (as identified in the New Zealand birth

register) were derived. We also derived some limited information on female caregivers' education, based on information recorded in the benefit and tertiary education data. The benefit data provides information on when a child was included under a parent's or caregiver's benefit.

The birth parents may not be the person or couple who actually cared for the child on Census night or at any given point in time. Many children will not be living with both birth parents at a given point in time. Some information on caregivers over time is available through the benefit data, and some information on caregivers associated with migrant children is available in the immigration application data.

First we compared Census night caregivers with birth parents and then Census night caregivers with the most recent caregivers identified through using a combination of birth parents, caregivers identified in the immigration application and benefit data for children in our linked Census sample.

Many of the administrative data-based measures for parents used in the modelling of school attainment are based on Department of Internal Affairs (DIA) birth parents only. For the 16.5% of students who were born overseas, these risk indicators will not be available. Within the remaining 83.5% of students, there are a number of students who are no longer with their DIA birth parent(s).

Figure 1.1 splits the cohort into those who have no DIA parent identified, one DIA parent identified and two DIA parents identified. The definition of the study population ensures that at least one Census night parent exists. When no DIA parent is identified, it is not possible for a match to Census night parents. When one DIA parent exists, there are three possibilities: no match, exact match or a match where the Census has an extra parent. When two DIA parents exist, there are four possibilities: no match, one match with one Census parent, one match with two Census parents and two matches. For the purposes of comparing, we are indifferent between no matches with one Census parent and no matches with two Census parents, which is why these groups have been combined.

Comparing Census night caregivers with birth parents, we find that 43.4% of children were living with both their birth parents (40.7%) or one birth parent if only one was listed on their birth certificate (2.7%). About 35% of children were living with one of their birth parents – 21% were living with one adult in a caregiver role and 14% were living with two adults in a caregiver role. About 22% of children had no Census night caregiver correctly identified, including about 17% who had no New Zealand birth record.

Figure 1.2 shows the corresponding chart using the most recent parent-child relationship recorded from DIA birth registration, immigration application or benefit spell. The coverage of this data has increased to 96.9% (from 83.6% for only DIA data). 68.3% of students have a perfect match between the most recent event and the Census, consisting of 47.2% where both parents match and 21.2% where the sole parent matches. Of those with a parent record in the combined data, 71.5% have a perfect match. Figure 1.2 also shows that 92.4% of students in the Census were residing with at least one of their New Zealand registered birth parents, which rises to 95.4% of those who have a parent relationship recorded in these data sources.

Figure 1.1. Comparing Census night caregivers with birth parents recorded in New Zealand birth register

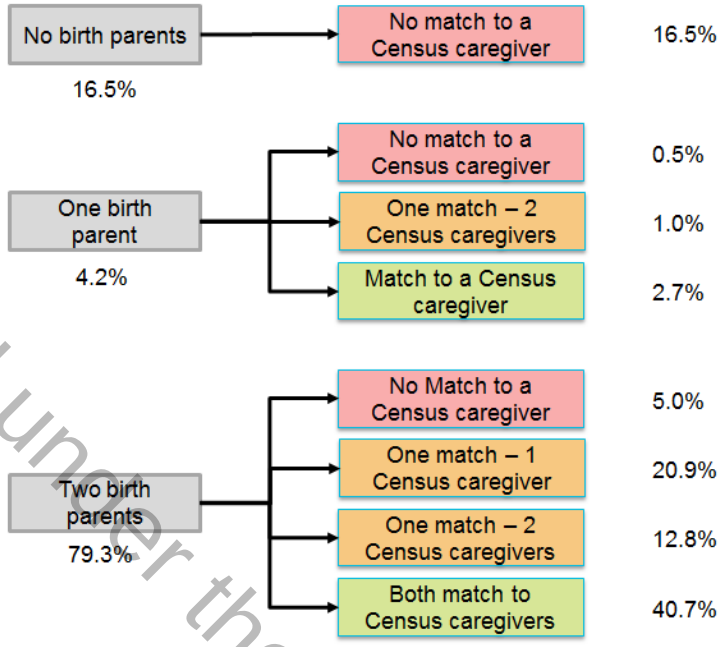
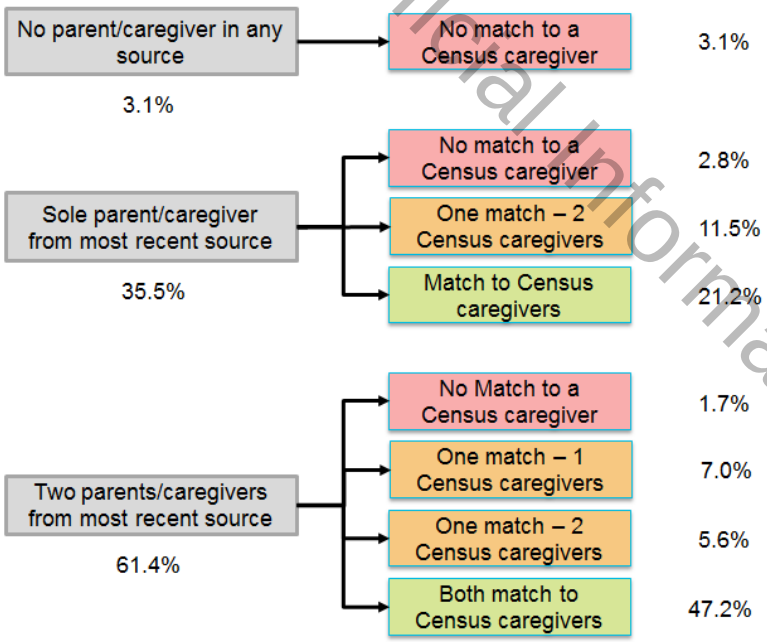


Figure 1.2. Comparing Census night caregivers with most recent caregivers identified in the births, migration and benefit spell data



Comparing Census night caregivers with the caregivers identified on the basis of the most recent information available (up to Census night) in the birth registrations, immigration applications and benefit spells (Figure 1.2), we find that, overall, 68% of children's Census night caregivers were correctly identified – 21% were living with one adult in a caregiver role and 47% were living with two adults in a caregiver role on Census night. About 24% of children had one of their Census night caregivers correctly identified – 7% were living

with one adult in a caregiver role and 17% were living with two adults in a caregiver role on Census night. About 8% of children had no Census night caregiver correctly identified.

Expanding the range of information used to determine parental risk indicators will improve the coverage of the student population while also improving the accuracy for sole parents. Before this can be implemented, further work is needed to define equivalent caregiver indicators when the caregiver changes over the period which the variable is defined.

Appendix 1 Table 10 provides a breakdown by the source of the most recent available information – New Zealand birth registration, benefit spell or immigration application approval.

It is unclear what impact incorporating information on caregivers identified in the migration and benefit data would have on the predictive power of the models we developed, and this is an area for further investigation as this work is progressed further. The impact may be fairly limited as we have included information on migrant status, including category, source country and how long they have been in New Zealand, as well as information on how long a child has been supported by a benefit and the type of benefit. We also found that including Census information on current caregiver's occupation and qualifications (for children aged 14 years) did not materially improve the predictive power of the model. It seems likely that adding information on current caregiver's earnings and corrections history (i.e. in place of or in addition to information on birth parent's earnings and corrections history) may not add much in terms of predictive power, although it should be investigated.

1.4 The characteristics of those who did not achieve NCEA level 2

In this section, we compare the characteristics of those who did not achieve NCEA level 2 by the end of the year they turned 17 and did not return to school in the following year. Appendix 1 Table 1 details all variables we included in the modelling exercise.

Three sources of data were available: Census 2013 information for the student, their family and household; Census 2013 meshblock-level information that is used in the current decile funding formula; and administrative data for children and their birth parent(s) (if the child has a New Zealand birth register record). Some limited information on mother's education, based on information recorded in the benefit system, is also available. The benefit data also provides information on children included under a parent's or caregiver's benefit.

There are five meshblock indicators derived from Census data:

- **Household income** – the percentage of households with equivalised income in the lowest 20% of households, excluding households with a member who is unemployed or households supported by a benefit.
- **Occupation** – the percentage of employed parents in occupations that are at skill levels 4 or 5 (ie, low skilled).
- **Household crowding** – the percentage of households with an equivalised crowding index greater than 1.

- **Educational qualification** – the percentage of parents with no tertiary or school qualifications.
- **Income support** – the percentage of parents who received main benefits in the previous year.

More information on how these variables are defined is available from the Ministry of Education's website (www.education.govt.nz/school/running-a-school/resourcing/operational-funding/school-decile-ratings). We obtained the 2013 meshblock indicator values from the Ministry of Education, which we rank and average across the five indices and then re-rank the average to construct a ranking of the meshblocks within the dataset. This measure is used to assess the predictive power of the current Census meshblock approach.

In the case of the Census, the analysis included characteristics of the adults identified as being in a parent role on Census night, including highest qualification, occupation and smoking status. We also derive some Census variables at the family and household level including family income, tenure, household equalised income and household crowding.

In the case of the administrative data, we derived a number of variables for children's birth parents identified in the DIA birth registration records. These derived variables include parent's offending and sentence history, parent's age at the student's birth, mother's age at first birth, number of children mother has had, income received from the benefit system, split by main benefits (tier 1), supplementary benefits (tier 2) and grants and allowances (tier 3),[2] and average earned income over the past 5 years. The administrative data also includes a very limited measure of mother's education, based on information recorded in the benefit and tertiary education data. Migrant children can be identified, including their immigration category and source country. The benefit data provides information on the periods of time the child has been supported by a parent's or caregiver's benefit.

Overall, there are a large number of variables that are correlated with NCEA attainment. Appendix 1 Table 1 shows that, for example, students who fail NCEA level 2 were more likely to have been supported by a benefit for more than 75% of the time since birth (23.1% of those who get NCEA level 2 versus 7.1% of those who don't), more likely to have been supported by a benefit at some stage since birth (55.5% versus 25.1%), more likely to have a mother with no qualifications at Census night (29.6% versus 13.1%), more likely to have a mother whose first child was born before the age of 20 (48.2% versus 24.9%) and more likely to have a CYF notification (41.9% versus 15.1%). These students were also more likely to be male (62.8% of those get NCEA level 2 versus 47.7% of those who don't) and were likely to identify themselves as Māori (17.7% versus 36.8% in administrative data).

Some of the other variables that are highly correlated with attainment include mother's smoking status, father's qualifications, father's offending and sentence history, having had a referral to the youth justice service, number of addresses recorded during the last 5 years, migrant status, country of birth and parent's average income over the last 5 years.

[2] There are three tiers to the New Zealand welfare system. The first tier consists of direct payments, the second tier consists of supplementary payments (such as Accommodation Supplement) and the third tier consists of temporary needs-based payments (such as Temporary Additional Support). More information is available from the Ministry of Social Development website.

1.5 The modelling approach

Forward selection was used to obtain the different models estimated. Forward selection is an automatic model selection algorithm that builds a model one variable at a time by adding the 'best' remaining variable above an entry threshold at each step.

The sample was randomly split into two halves – one half to estimate (or train) the model and the second to estimate the model's accuracy (validation).

To assess model performance, standard measures were used, including misclassification rates, specificity, sensitivity and area under the ROC.

The ROC curve is derived by plotting (1-specificity) versus sensitivity. The area under the ROC curve gives a summary measure of model performance across the range of predicted risk values. The area of the ROC curve ranges from 0.5 for a model with no predictive power to 1 for a model with perfect predictive power. An area of 0.5 corresponds to a ROC curve that is equal to the 45-degree line, and an area of 1 corresponds to a ROC curve that travels along the left and top border of the diagrams shown below.

Selected two-way classification tables are also shown. These involve splitting the population into those with the highest 25% of predicted risk of not achieving NCEA level 2 and the remaining 75% and comparing the proportions who actually achieve versus not. This corresponds to a single point on the ROC curve.

To ensure the model fit is not overstated, we estimate these measures on the validation sample. This means the measures of model performance based on the validation sample, such as the area under the ROC curve, will be slightly worse than for the sample used to estimate the model.

The order in which variables are entered into the model when forward selection is used can vary depending on the random sub-sample selected. One way to test the stability of the forward selection model is to repeatedly randomly select half of the dataset, record the order in which the variables enter the model and evaluate both how often the variable enters the model and, conditional on entering the model, the average entry position. This is discussed in section 0.

1.6 Modelling results

A summary of the forward selection procedure for the combined Census and administrative data model is presented in Table 1.2.

The most important variables in the model are proportion of time supported by a benefit as a child, having a CYF care and protection notification, mother's level of qualification (as recorded in the Census), gender and mother's age at first birth. Other variables that are moderately important were mother's smoking status (Census), father's highest qualification (Census) and student having a youth justice history.

Owing to random variation that arises due to splitting of the population to obtain a training and validation sample, the variable ordering may not be equal to the variable importance

shown in the following section, which tests the stability of the variables entering the model based on a forward selection of variables.

The ROC summary shows that most of the predictive power comes from the first five or six variables. The addition of the next eight or nine variables increases the predictive power of the model slowly. Adding variables beyond 14 has very little impact on the predictive power.

Table 1.2. Summary of forward selection of variables in the Census and administrative data model

Summary of forward selection					
Step	Effect entered	DF	Chi-Square	Pr > ChiSq	ROC summary
1	Proportion of time on benefit as child	5	1432.9	<.0001	0.6997
2	CYF notification	1	260.5	<.0001	0.7185
3	Mother's highest qualification (Census)	13	261.2	<.0001	0.7435
4	Gender	1	201.1	<.0001	0.7571
5	Mother's age at first birth	5	134.0	<.0001	0.7649
6	Mother's smoking status (Census)	3	103.8	<.0001	0.7709
7	Father's offending and sentence history	4	81.1	<.0001	0.7758
8	Asian ethnicity	1	53.6	<.0001	0.7788
9	Father's highest qualification (Census)	13	79.8	<.0001	0.7832
10	Youth justice referral	1	37.2	<.0001	0.7840
11	Tenure (Census)	2	39.2	<.0001	0.7864
12	Pacific ethnicity	1	26.7	<.0001	0.7872
13	Mother's labour force status	3	30.0	<.0001	0.7883
14	Father's industry of employment/not employed	19	51.8	<.0001	0.7913
15	Substantiated finding of abuse or neglect	1	13.5	0.0002	0.7919
16	Country of birth	12	35.5	0.0004	0.7940
17	Access to internet (Census)	2	12.2	0.0023	0.7946
18	Access to mobile phone (Census)	2	11.8	0.0028	0.7948
19	Household equivalised income (Census)	11	27.4	0.004	0.7960
20	Father's smoking status (Census)	3	12.3	0.0063	0.7966
21	Father's age at birth	6	17.9	0.0064	0.7973
22	Weekly rent (Census)	6	18.0	0.0062	0.7980
23	Father's average earned income in the last 5 years	15	31.0	0.0087	0.7993
24	Total family income (Census)	16	30.8	0.0143	0.8009
25	Number of addresses in the last 5 years	5	13.5	0.0191	0.8015
26	Mother of Pacific ethnicity	2	7.5	0.0232	0.8019
27	Most recent female benefit caregiver not DIA mother	1	4.7	0.0308	0.8019
28	Father of European ethnicity	2	6.6	0.0366	0.8024

The summary of the forward selection procedure for the administrative model is presented in Table 1.3. The most important variables in the administrative data-only model were the

proportion of time supported by a benefit as a child, having a CYF care and protection notification, gender, mother's age at first birth, father's offending and sentence history.

Other variables that were consistently selected into the model as moderately important include having a youth justice history, a mother who had no qualification (as recorded in benefit data), proportion overseas as a child, number of distinct household addresses over the past 5 years and mother's average income over the past 5 years.

Table 1.3. Summary of forward selection of variables in the administrative data model

Summary of forward selection					
Step	Effect entered	DF	Chi-Square	Pr> ChiSq	ROC summary
1	Proportion of time spent on benefit as child	5	1432.9	<.0001	0.6997
2	CYF notification	1	260.5	<.0001	0.7185
3	Gender	1	207	<.0001	0.7382
4	Mother's age at first birth	5	171	<.0001	0.7516
5	Father's offending and sentence history	4	103.7	<.0001	0.7588
6	Asian ethnicity	1	69.7	<.0001	0.7640
7	Youth justice referral	1	46.2	<.0001	0.7653
8	Mother's had no qualifications recorded	1	33.7	<.0001	0.7671
9	Father's average earned income during last 5 years	15	61.9	<.0001	0.7711
10	Mother's average earned income during last 5 years	15	53	<.0001	0.7740
11	Number of different addresses during last 5 years	5	33.5	<.0001	0.7755
12	Māori ethnicity	1	17.4	<.0001	0.7763
13	Mother received third tier benefits	1	10.7	0.0011	0.7771
14	Pacific ethnicity	1	10.3	0.0014	0.7771
15	Country of birth	12	34	0.0007	0.7790
16	Father's age at birth	6	19	0.0041	0.7798
17	Mother's age at child's birth	6	18.9	0.0043	0.7806
18	Substantiated finding of abuse or neglect	1	6.9	0.0088	0.7809
19	Mother's offending and sentence history	3	11.2	0.0109	0.7810
20	Last female benefit caregiver was not DIA mother	1	5.7	0.0169	0.7810
21	Proportion of time spent overseas	3	8.8	0.0318	0.7816

Comparing the variable ordering between the two models shows that the most important administrative variables in the combined model are coming in the same order, that is, the first four variables are the proportion of time supported by a benefit as a child, having a CYF care and protection notification, gender and mother's age at first birth. Other variables that consistently selected into the model as moderately important were father's offending and sentence history, student having a youth justice history, mother had no qualification (as recorded in benefit data), proportion overseas as a child, number of distinct household addresses over the past 5 years and mother's average income over the past 5 years.

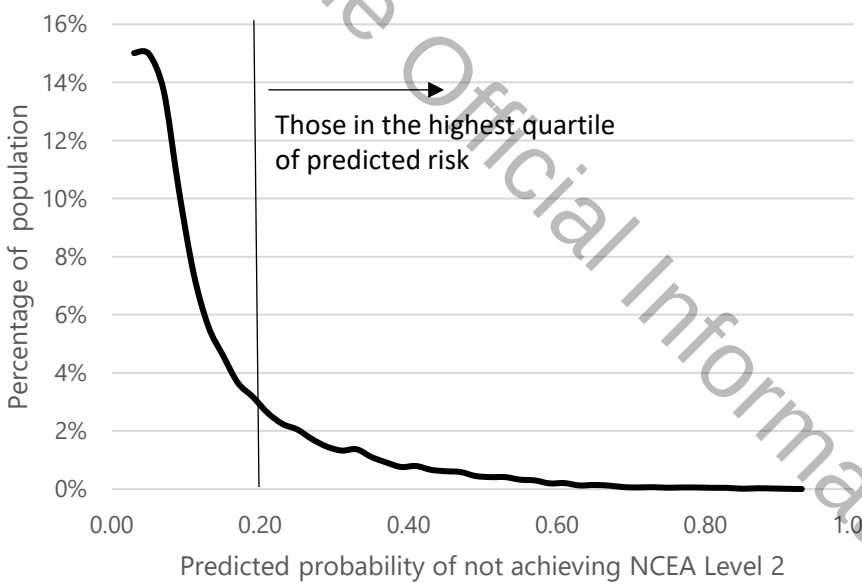
We also estimated the two models with ethnicity and gender removed, with model summaries presented in Appendix 1 Tables 2 and 3 for the combined model and

administrative data-only model respectively. These models had reduced predictive power, with ROCs around 0.02 lower in each case. The predictive power of the alternative models are examined in the next section where the ROC statistics are compared based on the *validation* samples.

As a point of comparison, we also estimated models based solely on Census data (Appendix 1 Table 5 and 6.) The most important variables in the model that included ethnicity and gender were mother's qualification, father's smoking status (which includes the effect of having no father in the household), tenure (home rented or owned), gender, mother's smoking status, mother's age, father's qualification, Asian ethnicity, mother's labour force status, Māori ethnicity and father's industry of employment (which includes the effect of having no employment).

Figure 1.3 shows the distribution of the probability of not achieving NCEA level 2 for the administrative data model. Most of the probabilities of not achieving are quite low, which follows from only 15% of the cohort not achieving NCEA level 2. If we considered those predicted by the model to be in the top quartile of predicted risk, we see this equates to having a predicted probability of not achieving above about 20%.

Figure 1.3. Distribution of predicted risk scores – administrative data model



1.6.1 Model accuracy

We compare the predictive performance of the models by comparing the ROC curves for the validation samples.

The ROC curve is derived by plotting (1-specificity) versus sensitivity. The area under the ROC curve gives a summary measure of model performance across the range of predicted risk values. The area of the ROC curve ranges from 0.5 for a model with no predictive power to 1 for a model with perfect predictive power. An area of 0.5 corresponds to a ROC curve that is equal to the 45-degree line, and an area of 1 corresponds to a ROC curve that travels along the left and top border of the diagrams shown below.

The model based on combined Census and administrative data has a ROC of 0.780, slightly higher than the administrative data model, which had a ROC of 0.770. The difference between the two models narrows to 0.10 when the performance of the models in assessed using the validation sample. This indicates that the combined model appears to be slightly 'overfitted' in that some variables are being included that are not adding predictive power to the model. These variables are likely to be later variables entering the model based on forward selection.

Figure 1.4. Comparing ROCs for the different models on the validation samples

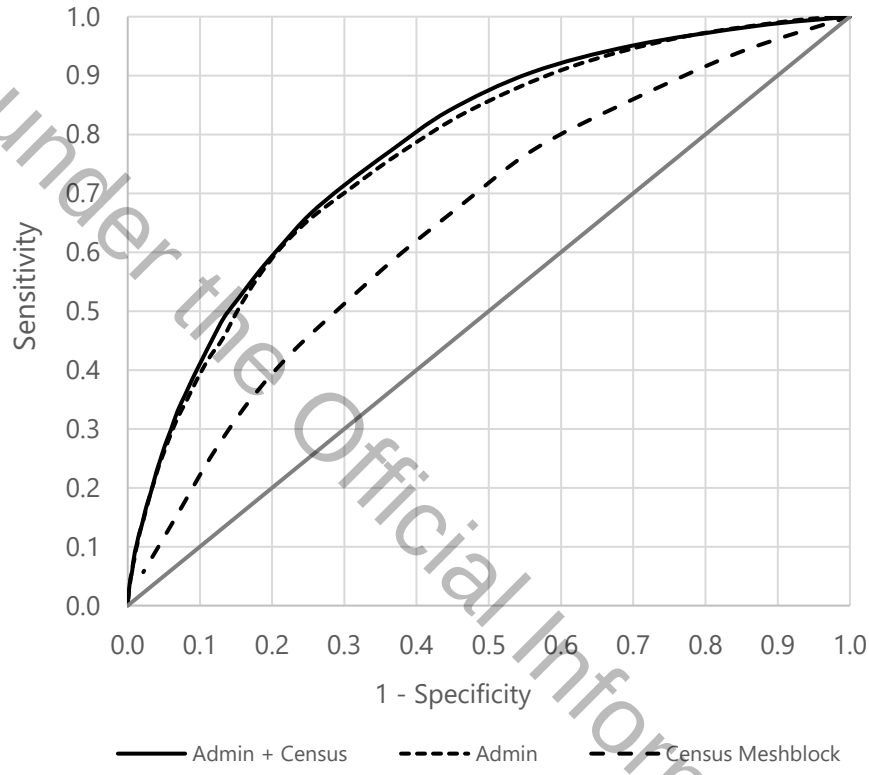


Figure 1.4 shows the ROC curves for the two models and compares this with that of a third model, which includes a single variable – the current Census meshblock index used for current decile funding to predict educational achievement. The area under the ROC curve is 0.652, which is significantly lower than either the combined or administrative data-only model (Appendix 1Table 4).

Appendix 1 Table 5 shows results for a model that includes only Census-based characteristics. The area under the ROC curve is 0.766, which is slightly lower than for the combined or administrative data-only model. Appendix 1 Table 6 shows results for a model that excludes gender and ethnicity, which gives a ROC of 0.746.

Another way to summarise the predictive power of the models is to look at two-way classification tables, which involves splitting the population into those with the highest 25% of predicted risk of not achieving NCEA level 2 and the remaining 75% and comparing the proportions who actually achieve versus not. This corresponds to a single point on the ROC curve. Appendix 1 Table 7 also includes results for the 15% versus 85% split, which aligns with the proportion who don't and do achieve NCEA level 2 in the

analysis sample. Splitting the sample at that point optimised the total proportion that are correctly classified, shaded grey in the tables below.

Table 1.4 shows the number and percentage of students in the validation sample who were correctly and incorrectly classified in the combined model. Overall, 77.4% of students were correctly classified at the 25% threshold. About 59% of those who didn't achieve NCEA level 2 were correctly identified by the model (sensitivity), and 80% of those who did achieve are correctly identified (specificity).

Table 1.4. Combined administrative and Census model

Predicted outcome administrative and Census model	Observed outcome	
	Achieved level 2	Did not achieve level 2
Highest 25% predicted risk	17.1	7.9
Lowest 75% predicted risk	69.5	5.5

Table 1.5 shows the percentage of students correctly classified based on the administrative data model. At the 25% threshold, the predictive performance of the model is the same as for the combined model.

Table 1.5. Administrative model

Predicted outcome administrative model	Observed outcome	
	Achieved level 2	Did not achieve level 2
Highest 25% predicted risk	17.1	7.9
Lowest 75% predicted risk	69.5	5.5

Table 1.6 shows the percentage of students correctly classified based on the Census meshblock model. Overall, 73.0% of students were correctly classified. About 42% of those who didn't achieve are correctly identified or classified by the model (sensitivity), and 78% of those who did achieve are correctly identified (specificity).

Table 1.6. Census meshblock model

Predicted outcome Census meshblock	Observed outcome	
	Achieved level 2	Did not achieve level 2
Highest 25% predicted risk	19.3	5.7
Lowest 75% predicted risk	67.3	7.7

Table 1.7 shows that excluding ethnicity and gender from the administrative model results in an additional 0.9% of students incorrectly classified. About 55.5% of those who didn't achieve NCEA level 2 are correctly identified (sensitivity), and 78% of those who did achieve are correctly identified (specificity). Overall, 76.5% are correctly identified at the 25% threshold.

Table 1.7. Administrative model excluding gender and ethnicity

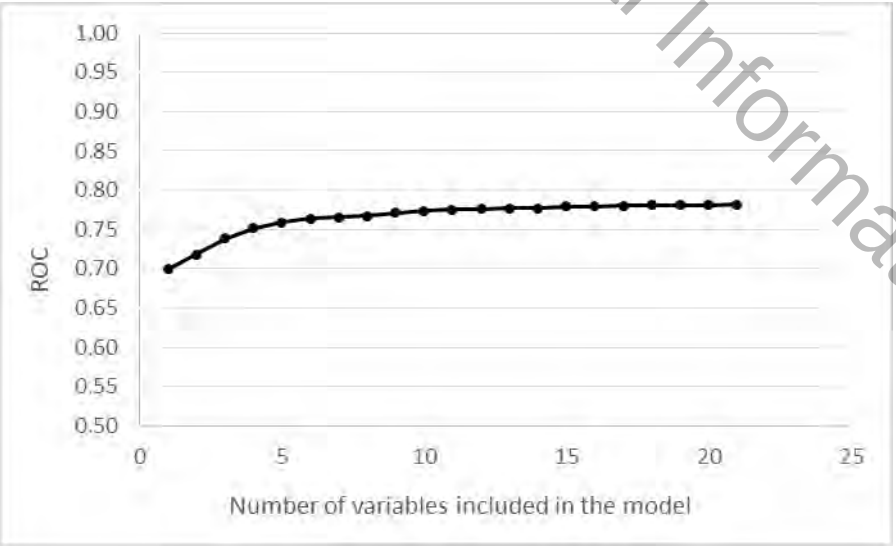
Predicted outcome Administrative model	Observed outcome	
	Achieved level 2	Did not achieve level 2
Highest 25% predicted risk	17.5	7.5
Lowest 75% predicted risk	69.0	6.0

Splitting the sample at 15% versus 85% optimises the total proportion of students that are correctly classified, sensitivity is reduced and specificity increased. For the combined model, 83.1% of students are correctly classified at the 15% threshold, 43% of those who didn't achieve NCEA level 2 were correctly identified (sensitivity) and 89% of those who did achieve NCEA level 2 were correctly identified (specificity).

For the administrative data model, 82.7% of students were correctly classified, 41% of those who didn't achieve NCEA level 2 were correctly identified (sensitivity) and 89% of those who did achieve NCEA level 2 were correctly identified (specificity).

Figure 1.5 shows how the ROC statistic (shown in Table 1.3) increases as each additional variable is added to the model via the forward selection process for the model based on administrative data only. This shows that the first 10 or so variables materially increase the ROC, but adding additional variables to the model after that has little impact on the predictive power of the model.

Figure 1.5. ROC statistic (from the training sample) by the number of variables included in the model based on forward selection of variables – administrative data model



1.6.2 Stability of the variables entered in the forward selection procedure

Statistical measures of variable importance in models can be difficult to interpret. We have used one possible measure – the order in which variables are added to the model during the forward selection process.

The order in which variables are entered into the model when forward selection is used can vary across different samples. One way to test the stability of the forward selection model is to repeatedly randomly select half of the dataset, record the order in which the variables enter the model and evaluate both how often the variable enters the model and, conditional on entering the model, the average entry position. For this purpose, 100 iterations of the forward selection model were performed for both the combined model and the administrative data-only model. We compare the variables entered and the entry order by running the model on multiple sub-samples of the data to show how stable the forward selection process is.

Results for the combined model and the administrative data model are shown in Table 1.8. In the case of the combined model, the proportion of time spent supported by a benefit since birth enters as the first variable in every run of the model.

The next three variables – mother's qualifications, child having a CYF notification and gender – consistently enter second, third and fourth in varying order. Mother's smoking status and mother's age at first birth are consistently the next most important, followed by either father's qualification or being of Asian ethnicity, youth justice referral, father's offending and sentence history, and number of addresses in the last 5 years. Maternal education available from the benefit data and country of birth are the final two variables consistently entered in the model.

Table 1.8. Summary of variables selected in the combined Census and administrative model based on 100 sub-samples

Effect entered	Rank	Percentage of time entered
Proportion of time on benefit as child	1.0	100%
Mother's qualifications (Census)	2.7	100%
Gender	3.1	100%
CYF notification	3.2	100%
Mother's smoking history (Census)	5.1	100%
Mother's age at first birth	6.4	100%
Father's qualifications/no father (Census)	7.5	100%
Asian ethnicity	8.4	100%
Youth justice referral	8.9	100%
Number of different addresses during last 5 years	11.9	100%
Pacific ethnicity	12.9	100%
Father's offending and sentence history	10.6	97%
Mother has no qualifications recorded	16.6	93%
Internet access (Census)	15.3	88%
Country of birth (Census)	17.5	87%
Substantiated finding of abuse or neglect	18.7	80%
Number of children	16.0	78%
Household equivalised income (Census)	19.9	76%

Effect entered	Rank	Percentage of time entered
Father's industry of employment/not employed	18.6	74%
Tenure (Census)	14.4	57%
Mother's smoking history (Census)	21.1	56%
Mother's average earned income during last 5 years	20.1	53%
Total family income (Census)	21.6	45%
Mother's labour force status (Census)	20.0	43%
Father's industry of employment/not employed	20.3	37%
Father's average income during last 5 years	20.7	37%
Number of years in New Zealand	21.8	37%
Weekly rent (Census)	15.6	36%
Mother's age at child's birth	22.9	34%
Father received third-tier benefits	18.6	31%
Mother received third-tier benefits	21.5	31%
Father's age at child's birth	22.5	30%
Father's occupation (Census)	19.5	26%
Father lived at different address 5 years ago (Census)	22.7	25%

Table 1.9. Summary of variables selected in administrative data-only model based on 100 sub-samples

Effect entered	Rank	Percentage of time entered
Proportion of time on benefit as child	1.0	100%
CYF notification	2.1	100%
Gender	2.9	100%
Mother's age at first birth	4.0	100%
Father's offending and sentence history	5.6	100%
Asian ethnicity	6.5	100%
Youth justice referral	7.7	100%
Mother has no qualifications recorded	8.4	100%
Proportion of time spent overseas	10.2	100%
Mother's average earned income during last 5 years	10.8	100%
Māori ethnicity	11.2	100%
Father's average earned income during last 5 years	11.8	100%
Number of different addresses during last 5 years	12.9	100%
Migrant category/New Zealand born	10.7	99%
Mother received third-tier benefit	16.7	73%
Substantiated finding of abuse or neglect	16.5	72%
Mother's age at child's birth	16.9	67%
Mother's offending and sentence history	16.2	61%
Country of birth	16.5	60%
Number of children	17.4	45%
Pacific ethnicity	18.5	40%
Father's age at the child's birth	18.2	32%
Other ethnicity	18.3	25%

In the case of the administrative data-only model, shown in Table 1.9, the proportion of time spent supported by a benefit since birth enters as the first variable in every run of the model, followed by the child having CYF notifications, then gender and mother's age at first birth. The next variables to enter the model are father's offending and sentence history, Asian ethnicity, youth justice referral and having an unqualified mother (based on administrative data).[3]

There is a further set of variables that always or nearly always enter the model, typically after the previously mentioned variables – proportion of time spent overseas, mother's average income over the previous 5 years, Māori ethnicity, father's average income over the last 5 years, the number of distinct addresses recorded in the last 5 years and migrant category. The remaining variables do not enter the forward selection model every time, and even when they do enter hardly ever enter earlier than the 14 variables that enter the model every time. The forward selection procedure is consistently highlighting the same set of variables as important.

1.7 Characteristics of children predicted to be most at risk of not achieving at school

Appendix 1 Table 8 and 9 describe the profile of children identified by the models as being at greatest risk of not achieving NCEA level 2. Appendix 1 Table 8 splits the validation sample into two groups – those in the highest quartile of predicted risk and those not in the highest quartile of predicted risk. Appendix 1 Table 9 splits the validation sample into two groups – those in the highest 15% of predicted risk and those not. The two groups are referred to as being either 'at risk' or 'not at risk'. These tables enable the 'at risk' group identified by the different models to be compared.

Nearly all of those in the highest 15% of predicted risk based on the administrative data model had spent some time supported by a benefit as a child. Only 2% had not spent any time supported by a benefit compared to 60% of children predicted to be not at highest risk. Nearly 70% of children identified by the model as being at risk had a CYF notification, 68% were male, 68% had a mother who was aged 19 years or younger when she had her first child and 59% had a father with an offending history. About 35% of those at risk had a mother with no qualifications (Census) compared to 12% of those not at risk. About 41% of those at risk had a mother who currently smoked (Census) compared to 13% of those not at risk. Even though these two Census variables were not included in the administrative data model, they are highly correlated with variables that were included.

Those in the highest quartile (25%) of predicted risk based on the administrative data model look quite similar. For example, 94% had spent some time supported by a benefit as a child, 55% had a CYF notification, 66% were male, 62% had a mother who was aged 19 years or younger when she had her first child and 52% had a father with an offending history. About 30% had a mother with no qualifications (Census), and 35% had a mother who currently smoked (Census).

Overall, those in the highest quartile of predicted risk based on the combined Census and administrative data model looked very similar to those identified by the administrative data

[3] Administrative data, excluding the Census, only records educational attainment for those in the benefit system and those who have attained a New Zealand qualification after 2007.

model. For example, 89% had spent some time supported by a benefit as a child, 52% had a CYF notification, 65% were male, 58% had a mother who was aged 19 years or younger when she had her first child, 47% had a father with an offending history, 37% had a mother with no qualifications (Census) and 41% had a mother who currently smoked (Census). The greatest differences between the two groups of children identified in the combined model and the administrative data model are for these last two characteristics, as these are the two most important Census variables in the combined model that can't be included in the administrative model.

1.8 School-level risk profile comparisons

We briefly compare the risk profile of children at each school, as calculated using the combined model and administrative model. The two measures used are:

- the percentage of children in each school in the upper quartile of predicted risk
- the average (or mean) predicted risk of children in each school.

The figures below only show schools that had more than 20 students in our 1998 cohort sample. Outliers have been removed.

Comparing Figure 1.6 and Figure 1.7 shows that the correlation between the two models is very strong in both cases but is materially weaker for the percentage of children in the upper quartile of predicted risk compared to mean risk.

Figure 1.6. Comparing the percentage of children at each school in the highest predicted risk quartile – administrative and Census data model versus administrative data model

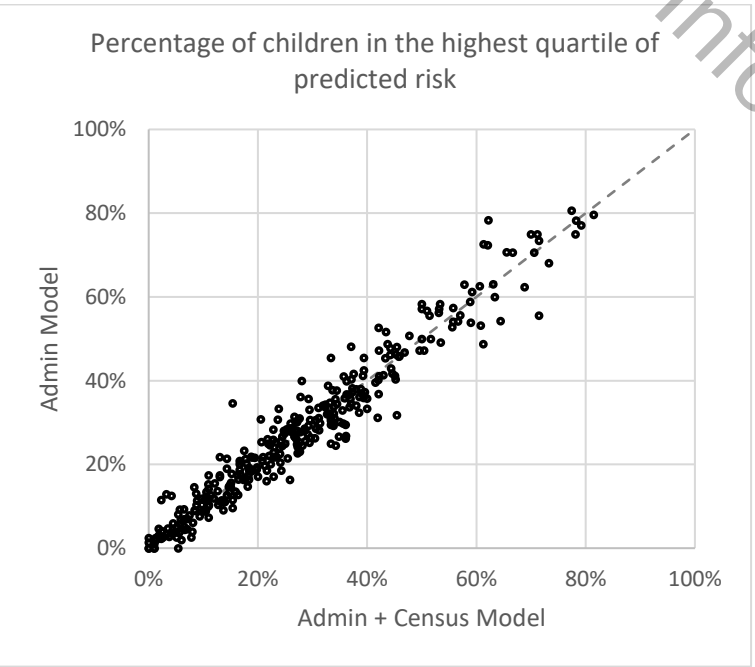


Figure 1.7. Comparing the average predicted risk of children at each school – administrative and Census data model versus administrative data model

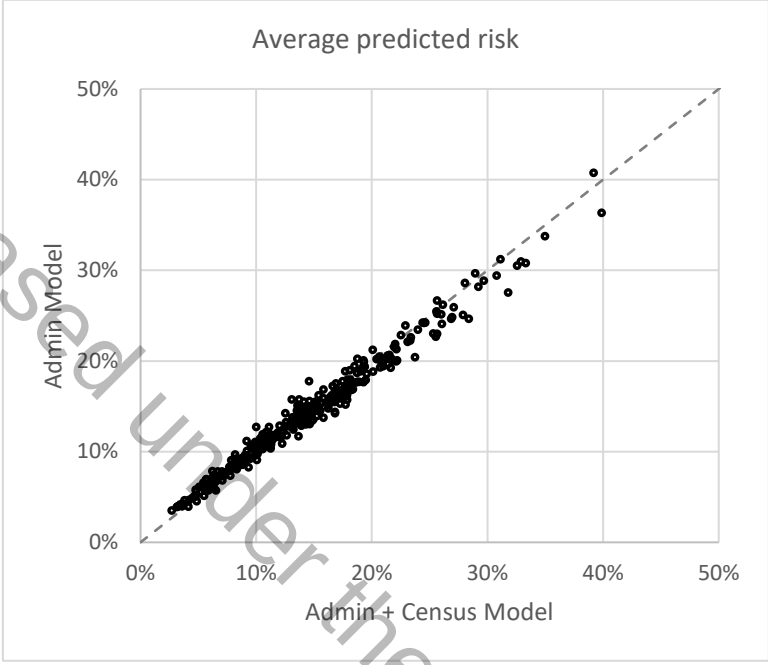
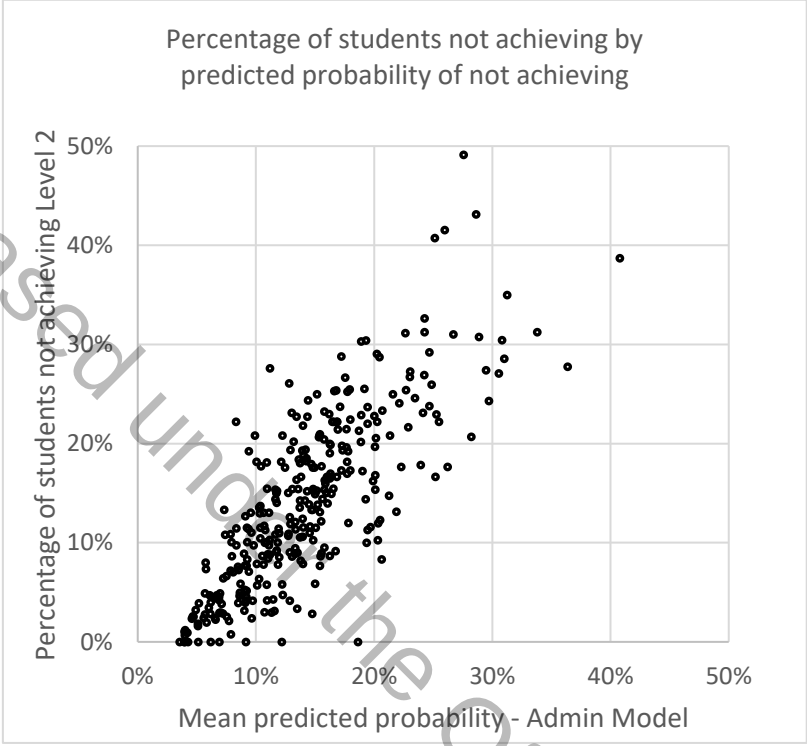


Figure 1.8 shows the relationship between the mean predicted risk score at each school and the actual proportion of children at each school who did not achieve NCEA level 2 for the administrative data model. Appendix 1 Figure 1.3 shows the relationship between the two different measures (mean risk and percentage in the upper quartile) and the actual proportion of children at the school who achieved NCEA level 2 for the combined data model and the administrative data model. The relationship between predicted and actual outcomes at the school level is stronger for mean predicted risk than for the percentage of children in the upper quartile of predicted risk. This can be expected, as an average risk score is capturing more information about *the distribution of predicted risk* at a school than a measure based on a percentage of children above a particular risk threshold.

Figure 1.8. Mean predicted probability of not achieving versus percentage of students not achieving NCEA level 2 – administrative data model



1.9 Limitations

Note that some of the variables that could have been used in the prediction models were deliberately excluded because they are (potentially) under the influence of schools – such as a student's history of achievement, their history of stand-downs and suspensions, and school characteristics. We were also asked to test the impact of excluding gender and ethnicity from the models on their predictive power.

The most significant limitation of the results in this part of the paper is the potential biases introduced by the various restrictions that had to be applied to obtain the final study population. Being a member of the final study population is likely positively associated with achieving NCEA level 2. Part 2 of this paper uses a more inclusive sample of the 1998 birth cohort and shows that variable importance can change as a result. In particular, Māori ethnicity is more important in the age 14 model developed in part 2 than it was here.

The core measure of NCEA level 2 achievement is based on incomplete data and could be updated after data on qualification attainment in 2016 is added to the IDI.

Incorporating information on non-NCEA qualification attainment into the IDI, so that children who attended a school that offers international qualifications can be included in the estimation samples, is another area for future work.

2 Predicting educational outcomes by year of age and estimating risk for a current population of students

2.1 Introduction

The purpose of this section of the paper is to investigate:

- which variables are most useful in predicting children's future NCEA achievement at different ages
- the location of students with different levels of predicted risk within schools and deciles in state-funded schools at present
- the way in which schools would be 're-ranked' if student-based risk measures were used in place of the current meshblock-based measures that underpin school deciles.

We use administrative data for a larger and more inclusive sample of youth who were born in 1998 to estimate a series of models at each year of age, using information on the child's history from birth to their most recent birthday. The coefficients obtained from these models are used to derive predictions for all domestic students who were enrolled at state-funded schools on 1 July 2015. These predictions are then used to investigate the profile of high-risk students in 2015, their distribution across schools and the effects of substituting student-risk measures to define school deciles on the relative ranking of individual schools.

Section 2.2 outlines the methods used and the model fit results. Section 2.3 analyses the results.

2.2 Modelling approach

2.2.1 Data sources

The main data source is the IDI, using the April 2016 version. Data on each school's decile and step in 2015 and the 2015 decile funding rates were taken from the Education Counts website and the file 'SchoolDecileChanges18June2015review.xls'.

2.2.2 Outcome measures

The 'risk of not achieving NCEA level 2' is the same outcome measure as that defined in part 1 of the analysis (section 1) and is the predicted risk of not gaining NCEA level 2 by the end of the year when the child turned 17 and not being enrolled at a school at the start of the following year when the child turned 18.

To construct this measure, the likelihood of being in the non-achieving, non-enrolled group is estimated for each child in the current student population using coefficients from a series of age-specific logistic regression models. The models are first estimated using data for children born in 1998 – whose attainment we can observe in the IDI using the latest available data. Measures of 'risk' are then constructed using either (a) the child's rank among students of all ages or (b) the child's rank among students of the same age.

For comparative purposes, we also identify children who have 'two or more of four factors' – whether the student has two or more of the following four indicators of disadvantage:

- Supported by a benefit for 50% or more of their life.
- Has been the subject of a CYF notification.
- Their birth father or birth mother has served a Department of Corrections custodial or community sentence.
- Their birth mother is recorded as having no qualifications in the benefit data.

These criteria were used to identify young children at risk of poor outcomes in previous work done by Treasury in 2015.

2.2.3 More details on the estimation sample used to generate model parameters

The estimation sample that provides the coefficients for the risk predictions is made up of all children who were born in 1998, excluding those who:

- had no domestic school enrolment record at ages 14–17
- were enrolled at private secondary schools while aged 15–17
- were enrolled at public secondary schools that offer international qualifications while aged 15–17
- were away from New Zealand for more than 6 months while aged 15–17
- were not living in New Zealand at the reference age
- were not enrolled at school for most or all of the year at the reference age (ie, they had left school already).

Essentially, we excluded individuals for whom the administrative data might not provide a comprehensive record of their NCEA achievement, because this would lead to biases in the dependent variable. We also excluded individuals who might be systematically different from those enrolled in publicly funded schools, such as youth who were enrolled at private schools during their upper secondary years.

The total size of the estimation sample is approximately 51,000 individuals, but the number used in each 'year of age' model is lower (due to the requirement that they be living in New Zealand at the reference year of age) and ranges from around 44,000 to 50,000.

This estimation sample is larger than the one that was used in the Census-IDI model comparison, because we did not need to drop individuals who did not have Census records linked to them in the IDI. A profile of the characteristics of the estimation sample is given in Appendix 2 (a separate Excel workbook).

2.2.4 Estimating predictive models by year of age

Separate models of NCEA level 2 attainment were estimated for each year of age from 5 to 17 years, each using data covering the period from birth up to the most recent birthday. For example, for children who were 5 years old, we use data covering the period from birth to the day before their 5th birthday for our model, while for children who were 17 years old, we use data covering the period from birth to the day before their 17th birthday.

Table 2.1. Explanatory variables included in the model at each year of age

Age of students in the estimation sample	5–11	12–13	14–16	17
Female	√	√	√	√
European ethnicity	√	√	√	√
Māori ethnicity	√	√	√	√
Pacific ethnicity	√	√	√	√
Asian ethnicity	√	√	√	√
Other ethnicity	√	√	√	√
Migrant category	√	√	√	√
Country of birth (aggregated into regions)	√	√	√	√
Years since residence approval obtained	√	√	√	√
Proportion of childhood spent overseas	√	√	√	√
Number of addresses in last 5 years		√	√	√
Mother's age at birth of first child	√	√	√	
Mother has no qualifications	√	√	√	
Mother's age at birth of reference child	√	√	√	
Mother's number of children at birth of reference child	√	√	√	
Mother was a single parent at birth of reference child	√	√	√	
Father's age at birth of reference child	√	√	√	
Father not listed on birth certificate	√	√	√	
Has female caregiver who is not the birth mother	√	√	√	
Has male caregiver who is not the birth mother	√	√	√	
Mother's proven charges and sentences served	√	√	√	
Father's proven charges and sentences served	√	√	√	
Proportion of childhood supported by benefits	√	√	√	√
Main benefit type in childhood	√	√	√	√

Age of students in the estimation sample	5–11	12–13	14–16	17
CYF care and protection notification	√	√	√	√
CYF care and protection finding	√	√	√	√
CYF care and protection placement	√	√	√	√
CYF youth justice referral			√	√
CYF youth justice placement			√	√
Mother's time in wage or salaried employment, last 5 years	√	√	√	
Mother's earned income, last 5 years	√	√	√	
Mother's first-tier benefit income, last 5 years				
Mother's second-tier benefit income, last 5 years	√	√	√	
Father's time in waged or salaried employment, last 5 years	√	√	√	
Father's earned income, last 5 years	√	√	√	
Father's first-tier benefit income, last 5 years	√	√	√	
Father's second-tier benefit income, last 5 years	√	√	√	

Data covering whole of the 17th or 18th year is not yet available for all members of the 1998 birth cohort. For this reason, we use the parameters from our model for 17-year-olds to predict the achievement of students who were 18 or 19 and still at school in mid-2015.

Only administrative data variables were included in the models. The explanatory variables were almost exactly the same as those used in part 1, except that some variables were either not available or not relevant for the models estimated for students at younger or older ages. At younger ages, information on the number of addresses lived at in the past 5 years wasn't available for the estimation sample, and the child's contact with the youth justice system wasn't relevant. At ages 17–19, parental variables weren't included in the models because the proportion of students in the current student population that have linked parents (identified via the DIA birth records) was much lower than for ages 5–16. The full list of variables that were included at each year of age is given in Table 2.1.

Variables were included one by one using forward selection, where the variable with the greatest explanatory power is selected at each step.

2.2.5 Model fit and variables that were consistently selected into the models

Table 2.2 provides model fit statistics. The ROC statistic is around 0.77 for younger ages and 0.78 for ages 13–15. The ROC statistics for the model that was estimated at 14 years of age is 0.788, which is slightly higher than the ROC statistic for the administrative data-only model in part 1 (0.770). This likely reflects differences in the composition of the samples, with the Census analysis based on a more restricted sample.

Table 2.2. Area under the ROC curve for logistic regressions on NCEA level 2 attainment

Age	Training sample	Validation sample
5	0.7741	0.7694
7	0.7791	0.7693
9	0.7779	0.7719
11	0.7849	0.7727
13	0.7875	0.7799
14	0.7830	0.7881
15	0.7935	0.7863
17	0.7414	0.7536

Note: Each model was estimated on a randomly selected 50% sample (the training sample) and then fitted to the remaining observations (the validation sample).

Table 2.3 shows the variables that were most consistently selected into the models of NCEA level 2 attainment at ages 5, 7, 9 11, 13, 14, 15 and 17. To get a set of results that are not influenced by any idiosyncratic features of the 1998 cohort sample, we took 20 random draws comprising the 50% the sample, re-estimated the model using each sub-sample and averaged the results of the 20 iterations.

The first column shows the proportion of time that the variable was selected and the second column the mean order (or rank) in which the variable entered the model in the forward selection process. Only variables that were selected at least half the time are shown.

The highest ranked variable, across all ages, is the proportion of time the child was supported by a parent's (or caregiver's) benefit since birth. Gender (ie, being male) and having Māori ethnicity are nearly always selected within the first five. The mother's age at her first birth and an indicator for the mother being unqualified are also frequently selected in the top five. Other variables that have a high probability of selection within the first 10 include whether the child has had a CYF care and protection notification, the mother's earned income in the last 5 years, the father's earned income in the last 5 years, the father's justice sector history (covering proven charges, community sentences and custodial sentences), the mother's justice sector history and having Asian ethnicity.

At 15 years of age, whether the child has had a referral to CYF youth justice services is the fifth highest ranked predictive variable. (This is not relevant at younger ages.)

Table 2.3. Variables with greatest predictive power at different ages

	Proportion of times entered	Average rank		Proportion of times entered	Average rank
Age 5			**Age 7**		
Proportion of childhood supported by benefit	1.00	1.0	Proportion of childhood supported by benefit	1.00	1.0
Gender	1.00	2.0	Gender	1.00	2.0
Māori	1.00	3.6	Māori	1.00	4.0
Mother's age at birth of first child	1.00	4.7	Mother's age at birth of first child	1.00	4.9
Mother has no qualifications	1.00	5.4	CYF care and protection notification	1.00	4.9
Asian	1.00	6.8	Father's justice sector history	1.00	6.6
Father's justice sector history	1.00	7.5	Mother has no qualifications	1.00	6.8
Father's earned income	1.00	8.8	Asian	1.00	6.9
CYF care and protection notification	1.00	8.9	Father's earned income	1.00	9.0
Mother's earned income	1.00	10.8	Mother's earned income	1.00	10.4
Mother's justice sector history	1.00	11.5	Mother's second-tier benefit income	1.00	12.0
Mother's second-tier benefit income	1.00	12.2	Mother's justice sector history	1.00	12.5
Pacific ethnicity	1.00	14.8	Pacific ethnicity	1.00	13.6
Mother's age at birth	0.90	14.1	Mother's age at birth	0.80	14.9
Mother's time in wage or salaried employment	0.85	15.2	Mother's number of children at birth of reference child	0.65	11.4
Father's first-tier benefit income	0.80	14.3	Country of birth	0.65	16.1
Main benefit type in childhood	0.75	18.4	Proportion of childhood overseas	0.55	15.2
Proportion of childhood overseas	0.70	16.8	Father's age at birth	0.55	16.7
Country of birth	0.55	15.5			
Father not listed on birth certificate	0.50	17.5			
Age 9			**Age 11**		
Proportion of childhood supported by benefit	1.00	1.0	Proportion of childhood supported by benefit	1.00	1.0
Gender	1.00	2.0	Gender	1.00	2.0
CYF care and protection notification	1.00	4.0	CYF care and protection notification	1.00	3.1
Mother's age at birth of first child	1.00	4.4	Māori	1.00	5.2
Māori	1.00	4.8	Father's justice sector history	1.00	5.5
Asian	1.00	6.5	Mother's age at birth of first child	1.00	5.6
Father's justice sector history	1.00	7.4	Asian	1.00	6.4
Mother has no qualifications	1.00	7.4	Mother has no qualifications	1.00	7.8
Father's earned income	1.00	8.4	Mother's earned income	1.00	9.1
Mother's earned income	1.00	9.4	Father's earned income	1.00	9.7
Mother's justice sector history	1.00	12.7	Mother's justice sector history	1.00	12.7
Pacific ethnicity	1.00	13.6	Pacific ethnicity	1.00	12.8
Country of birth	1.00	13.9	Country of birth	1.00	12.9
Mother's age at birth	0.90	13.1	CYF findings of abuse or neglect	0.80	15.9
Mother's second-tier benefit income	0.80	14.7	Mother's age at birth	0.70	15.0
Proportion of childhood overseas	0.55	15.6	Mother's second-tier benefit income	0.70	16.0
CYF findings of abuse or neglect	0.50	17.2	Mother's number of children at birth of reference child	0.65	12.8
			Father's time in wage or salaried employment	0.60	17.7
			Proportion of childhood overseas	0.50	16.3

	Proportion of times entered	Average rank		Proportion of times entered	Average rank
Age 13			**Age 14**		
Proportion of childhood supported by benefit	1.00	1.0	Proportion of childhood supported by benefit	1.00	1.0
CYF care and protection notification	1.00	2.4	CYF care and protection notification	1.00	2.2
Gender	1.00	2.6	Gender	1.00	2.9
Māori	1.00	5.3	Mother's age at birth of first child	1.00	5.2
Mother's age at birth of first child	1.00	5.4	Father's justice sector history	1.00	5.9
Father's justice sector history	1.00	5.9	Māori	1.00	5.9
Asian	1.00	6.4	Asian	1.00	6.4
Mother has no qualifications	1.00	7.7	Mother has no qualifications	1.00	7.7
Mother's earned income	1.00	9.5	Mother's earned income	1.00	9.3
Father's earned income	1.00	10.4	Father's earned income	1.00	10.1
Mother's justice sector history	1.00	12.2	Number of addresses in last 5 years	1.00	11.3
Country of birth	1.00	13.2	Mother's justice sector history	1.00	13.2
Number of addresses in last 5 years	1.00	13.6	CYF findings of abuse or neglect	1.00	14.7
Pacific ethnicity	1.00	13.8	Mother's age at birth	0.95	14.9
Mother's age at birth	0.90	14.9	Country of birth	0.95	15.1
CYF findings of abuse or neglect	0.90	15.2	Pacific ethnicity	0.95	15.2
Mother's second-tier benefit income	0.90	16.3	Mother's second-tier benefit income	0.70	16.6
Proportion of childhood overseas	0.70	16.3	Proportion of childhood overseas	0.55	15.9
			Mother's number of children at birth of reference child	0.50	13.3
			Father's time in wage or salaried employment	0.50	18.0
Age 15			**Age 17**		
Proportion of childhood supported by benefit	1.00	1.0	Proportion of childhood supported by benefit	1.00	1.1
CYF care and protection notification	1.00	2.0	CYF youth justice referral	1.00	2.1
Gender	1.00	3.0	CYF care and protection notification	1.00	3.3
CYF youth justice referral	1.00	5.2	Gender	1.00	3.7
Mother's age at birth of first child	1.00	6.2	Number of addresses in last 5 years	1.00	5.2
Māori	1.00	6.9	Asian	1.00	6.7
Asian	1.00	7.0	Māori	1.00	6.8
Father's justice sector history	1.00	7.5	Proportion of childhood overseas	0.95	8.5
Mother has no qualifications	1.00	8.3	Country of birth	0.95	8.6
Number of addresses in last 5 years	1.00	10.1	Main benefit type in childhood	0.70	10.9
Mother's earned income	1.00	10.3	CYF care and protection placement	0.50	10.0
Father's earned income	1.00	12.3			
Country of birth	1.00	15.0			
Pacific ethnicity	1.00	16.1			
Mother's justice sector history	0.95	15.3			
CYF findings of abuse or neglect	0.95	17.5			
Mother's age at birth	0.85	15.8			
Proportion of childhood overseas	0.75	17.7			
Mother's second-tier benefit income	0.70	16.5			
Mother's number of children at birth of reference child	0.55	14.9			
Mother a single parent at birth	0.55	20.5			

Note: The results are averages derived from 20 regressions estimated using 20 random draws from the base sample. Only variables that were selected at least half the time are shown in the table.

Comparing models estimated at different years of age, we find that parental variables become less important at older ages, while the child's care and protection history and youth justice history (if any) becomes more important. The variables that were most predictive in the administrative data model at age 14 that was developed in part 1 of this paper continue to be most predictive at age 14. The main difference is that Māori ethnicity

ranks more highly at number 6 compared with 11 previously. The higher ranking seems to be a consequence of the fact that the larger sample used here includes a greater number of Māori students who did not achieve NCEA level 2. (Some of these students were not available for the first-stage analysis because they did not have linked Census records.) Māori ethnicity is even more important at younger ages and is found at number 3, 4 or 5 in the models for ages 5–13 and at number 6 or 7 in the models for ages 14–17.

As discussed above, some variables were only available for some age groups so weren't used in all the models. Changes in data availability in future could affect the explanatory power of these variables. For example, the variable 'number of addresses in the last 5 years' is likely to become a more useful predictor in future when data becomes available over a wider span of ages, and it could push other variables further down the ranking.

2.2.6 Selecting the current student population as at mid-2015

The current student population in this analysis comprises domestic students who were enrolled at state-funded schools on 1 July 2015 – more specifically, those who were:

- enrolled at a state-funded primary, intermediate or secondary school on 1 July 2015 (excluding special education schools, teen units and the Correspondence School)
- linked to the IDI spine
- aged 5–19 on 1 July 2015.

The number of students in this sample is approximately 709,600, and the number of schools is approximately 2,370. A profile of the characteristics of the current student population is given in Appendix 2 (a separate Excel workbook).

2.2.7 Applying the model parameters to the current population of students

As described above, students in the current student population were each given a predicted risk score using the parameters estimated from the year-of-age models.

2.3 Results

This section describes the distribution of risk measures among students and schools.

2.3.1 Risk measures at student level

After calculating a risk score for each student, we also calculate:

- an indicator for being in the highest quartile of predicted risk (using the risk distribution for all students of all ages)
- an indicator for being in the highest quartile of risk, compared with students of the same age
- an indicator for having at least two of four of the selected risk factors (supported by a benefit for 50% or more of lifetime, CYF notification, birth parent has served a Department of Corrections sentence, birth mother has no qualifications).

Figure 2.1 and Figure 2.2 and Table 2.4 and Table 2.5 give summary statistics at the student level for different groups of students using the four indicators:

- Mean predicted risk score.
- Being in the highest quartile of predicted risk.
- Being in the highest quartile of predicted risk relative to students of the same age.
- Having at least two of the selected four risk factors.

Figure 2.1. Risk measures by year of age, 2015 student population

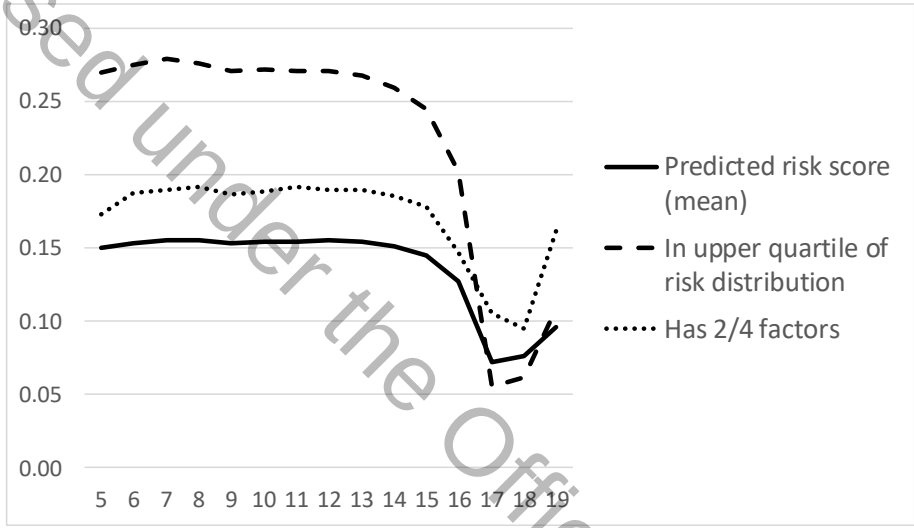


Table 2.4. Risk measures by year of age, 2015 student population

Age	Number of students	Predicted risk score (mean)	Proportion in upper quartile of risk distribution	Proportion in upper quartile of risk distribution (by age)	Has 2/4 factors
All	709,605	0.146	0.250	0.250	0.178
5	58,776	0.150	0.270	0.250	0.173
6	60,168	0.153	0.275	0.250	0.188
7	60,549	0.155	0.279	0.250	0.189
8	58,794	0.155	0.276	0.250	0.191
9	56,049	0.153	0.270	0.250	0.186
10	54,852	0.154	0.272	0.250	0.188
11	53,970	0.155	0.271	0.250	0.192
12	52,452	0.155	0.271	0.250	0.189
13	51,858	0.154	0.268	0.250	0.190
14	53,073	0.151	0.260	0.250	0.186
15	53,250	0.145	0.245	0.250	0.178
16	47,208	0.127	0.202	0.250	0.147
17	39,882	0.072	0.055	0.253	0.105
18	8,040	0.076	0.062	0.251	0.095
19	681	0.096	0.109	0.258	0.162

The breakdown by year of age shows that the proportion of students who were classified in our analysis as high risk (compared with all other students) falls rapidly at ages 17, 18 and 19. We believe this is due to higher-risk students leaving school. This means that funding models that are based on student risk will tend to allocate relatively less funding to upper-secondary school students and to secondary schools (unless the risk pool is defined differently).

There are substantial gender and ethnic variations in these measures of predicted risk, with male students showing higher average risk scores than female, Māori and Pacific students showing higher average risk scores than European, and Asian students showing lower average risk scores than European.

Figure 2.2. Risk measures by student's school decile

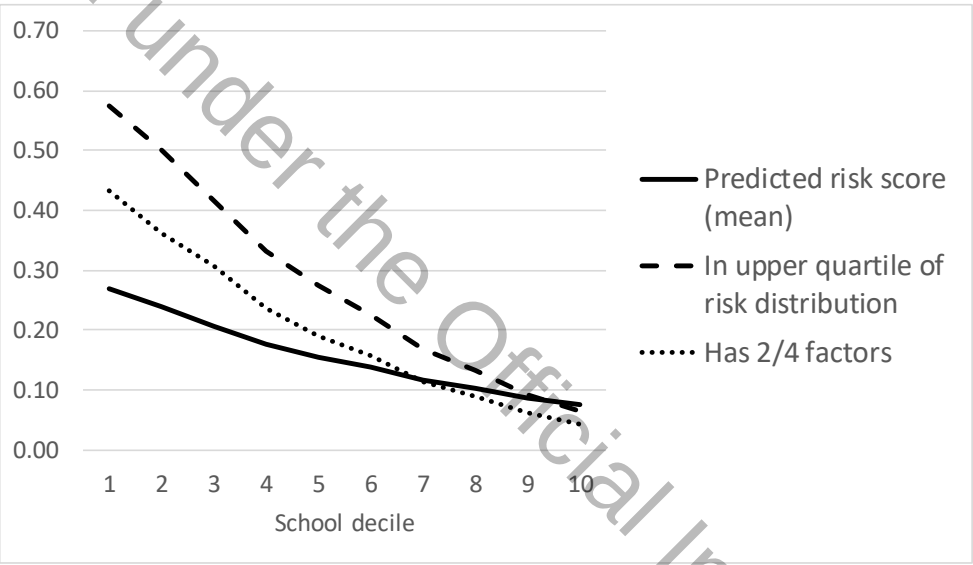


Table 2.5. Risk measures at student level, 2015 student population

	Number of students	Predicted risk score (mean)	Proportion in upper quartile of risk distribution (ages 5–19)	Proportion in upper quartile of risk distribution (by age)	Has 2/4 factors
Males	361,422	0.180	0.318	0.319	0.177
Females	348,183	0.110	0.180	0.178	0.179
Ethnic group					
European	442,458	0.118	0.173	0.172	0.119
Māori	165,582	0.280	0.601	0.601	0.411
Pacific	86,763	0.173	0.341	0.339	0.266
Asian	79,035	0.039	0.011	0.012	0.045
Primary school	430,365	0.153	0.272	0.249	0.187
Secondary school	279,240	0.134	0.216	0.252	0.164

	Number of students	Predicted risk score (mean)	Proportion in upper quartile of risk distribution (ages 5–19)	Proportion in upper quartile of risk distribution (by age)	Has 2/4 factors
School decile					
1	57,294	0.268	0.575	0.567	0.433
2	51,543	0.238	0.501	0.493	0.363
3	59,172	0.206	0.415	0.418	0.306
4	64,629	0.175	0.331	0.334	0.235
5	65,877	0.154	0.273	0.269	0.190
6	72,438	0.138	0.226	0.236	0.158
7	81,384	0.117	0.169	0.172	0.114
8	85,584	0.102	0.131	0.133	0.088
9	87,258	0.086	0.092	0.092	0.062
10	84,420	0.075	0.064	0.061	0.041
All	**709,605**	**0.146**	**0.250**	**0.250**	**0.178**

2.3.2 Risk measures at school level

The school-level risk measures are calculated by taking the average of the scores for all students who were enrolled at each school on 1 July 2015 and are included in our student dataset (see section 2.2.5). Schools with fewer than 10 students in our student dataset were dropped from the sample at this point. This reduced the number of schools from 2,373 to 2,352.

Figure 2.3 and Figure 2.4 show the distribution of two risk measures across schools: the mean risk score and the proportion of students who were in the upper quartile of risk.

Figure 2.3. Distribution of schools by their mean student risk score

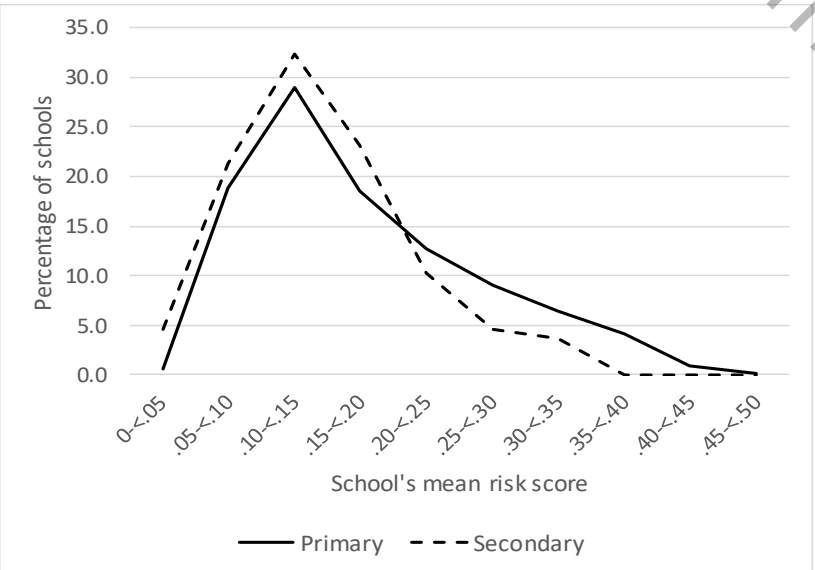
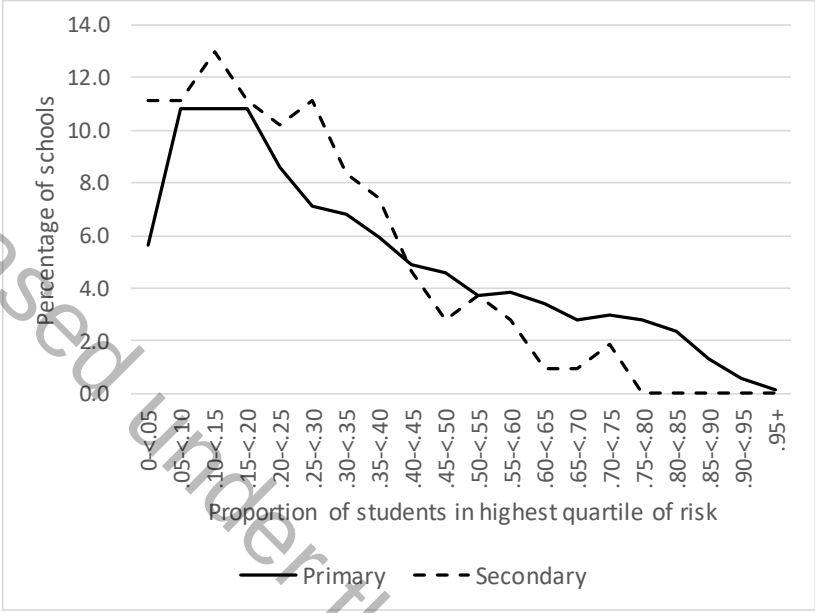


Figure 2.4. Distribution of schools by the proportion of students who are in the upper quartile of risk



Both figures show that most schools are concentrated at the lower end, and the distribution is skewed to the right (with decreasing numbers of schools at the high end). In addition, there are relatively fewer secondary than primary schools in the upper part of the risk distribution. This is partly because secondary schools are larger, leading to greater averaging over lower and higher-risk students and partly because higher-risk youth are more likely to leave the schooling system from age 16 onwards.

Note that we have included composite schools (covering years 1–13) in the primary category (because about two-thirds of their students are aged 5–12). If we took these schools out of the primary category, the results would be somewhat different.

Figure 2.5 and Table 2.6 show how these risk measures vary across schools at each decile. While in all cases there is a systematic relationship between current decile and the level of the risk measure, the slope of the relationship varies and is steepest for the measures that capture the proportion of students who were in the upper quartile of risk.

Figure 2.5. School risk measures by decile

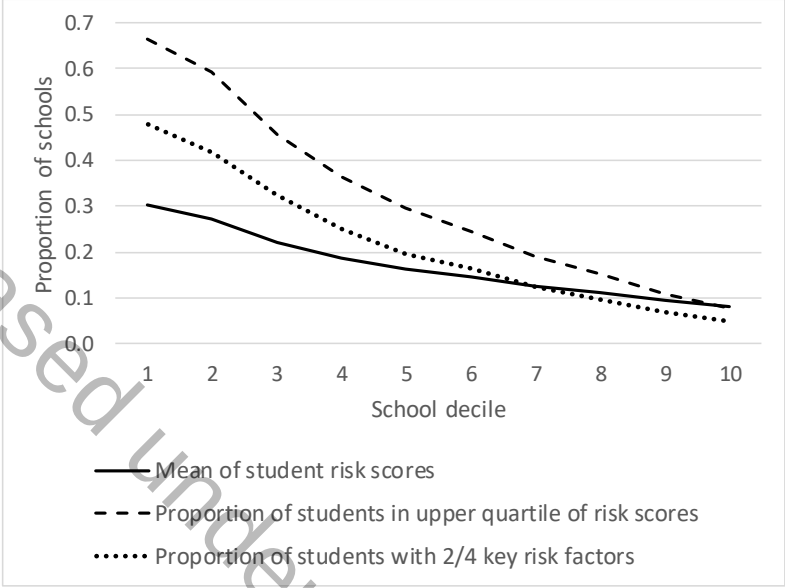


Table 2.6. School risk measures by decile

Decile	Number of schools	Mean of student risk scores	Proportion of students in upper quartile of risk scores	Proportion of students with 2/4 key risk factors
1	261	0.303	0.663	0.480
2	240	0.272	0.592	0.415
3	231	0.221	0.457	0.323
4	222	0.187	0.365	0.250
5	237	0.163	0.294	0.195
6	231	0.146	0.244	0.164
7	231	0.126	0.191	0.122
8	237	0.112	0.153	0.097
9	231	0.095	0.109	0.070
10	231	0.081	0.077	0.047

Note: The numbers of schools have been randomly rounded.

2.3.3 How individual schools would be re-ranked if deciles were based on measures of student risk

To identify the effects of using student risk measures rather than the current Census-derived indicators on the ranking of each school relative to other schools, we create several new decile classifications – one for each measure of student risk. We do this by ranking schools by their score on each risk measure and then allocating them to one of 10 bands within a new, simulated decile variable. The latter is constructed on the basis that the number of schools at each step and decile of the current system is preserved.

Table 2.7. Comparison of schools' current decile with their decile ranking using the proportion of students who are in the upper quartile of the risk measure

	Decile based on proportion of students in upper quartile of risk										
Decile	1	2	3	4	5	6	7	8	9	10	Total
1	56.3	25.3	10.3	4.6	0.0	1.1	0.0	0.0	0.0	0.0	100.0
2	36.3	33.8	13.8	7.5	3.8	1.3	1.3	0.0	0.0	0.0	100.0
3	9.1	24.7	32.5	16.9	10.4	3.9	1.3	1.3	1.3	0.0	100.0
4	0.0	13.5	24.3	23.0	16.2	13.5	6.8	2.7	0.0	1.4	100.0
5	1.3	2.5	11.4	21.5	26.6	15.2	10.1	6.3	2.5	1.3	100.0
6	0.0	1.3	2.6	17.1	23.7	19.7	14.5	9.2	9.2	2.6	100.0
7	0.0	0.0	2.6	3.8	12.8	23.1	24.4	16.7	11.5	6.4	100.0
8	0.0	0.0	1.3	0.0	5.1	12.7	29.1	22.8	16.5	13.9	100.0
9	0.0	0.0	0.0	0.0	1.3	5.2	10.4	31.2	26.0	26.0	100.0
10	0.0	0.0	0.0	0.0	1.3	1.3	2.6	9.1	35.1	49.4	100.0

Note: Each row shows the percentage of schools within a current decile that would lie in each hypothetical 'new' decile, if deciles were constructed on the basis of a school's proportion of students in the upper quartile of the predicted risk of not achieving NCEA level 2.

Table 2.8. Movement of schools if current decile system is replaced by a decile ranking based on the proportion of students who are in the upper quartile of the risk measure

	Number of schools					Percentage of schools				
Change in decile	All	>33% Māori	>25% Pacific	Prim.	Sec.	All	>33% Māori	>25% Pacific	Prim.	Sec.
Move down	879	306	12	804	78	37	41	5	40	24
No change	747	312	54	654	93	32	41	21	32	29
Move up	726	135	192	570	153	31	18	74	28	47
Total	2,352	753	261	2,025	324	100	100	100	100	100
Down 5+ deciles	12	0	0	12	0	1	0	0	1	0
Down 4 deciles	15	6	0	15	0	1	1	0	1	0
Down 3 deciles	54	15	0	51	0	2	2	0	3	0
Down 2 deciles	225	81	0	213	12	10	11	0	11	4
Down 1 decile	576	195	12	510	66	24	26	5	25	20
No change	747	309	54	651	96	32	41	21	32	30
Up 1 decile	378	102	81	303	75	16	14	31	15	23
Up 2 deciles	210	24	63	165	45	9	3	24	8	14
Up 3 deciles	96	6	30	75	21	4	1	11	4	6
Up 4 deciles	24	0	9	18	3	1	0	3	1	1
Up 5+ deciles	24	0	9	15	9	1	0	3	1	3
Total	2,352	753	261	2,025	327	100	100	100	100	100

Notes: Each column shows the number and percentage of schools that could change their decile ranking if deciles are constructed on the basis of a school's proportion of students in the upper quartile of the predicted risk of not achieving NCEA level 2. Note that the number of schools in each cell has been randomly rounded to base 3, and counts under 2 have been suppressed.

When student risk scores are used to assign schools to deciles, the majority of schools move up or down by at least one decile, compared with their current position. This is illustrated in Table 2.7 and Table 2.8, which compare the current system with one based on the proportion of students who are in the top 25% of risk scores (using the scores of all 5–19 year olds to select the top 25%).

Each row of Table 2.7 gives the percentage of schools who are classified to each decile of the (simulated) new decile system. In Table 2.8, we summarise the movements. The first column shows that 32% of all schools retain their position, 37% move down and 31% move up. While most schools only move up or down by one decile, 28% of all schools move by two or more deciles.
We also analyse the effects for different types of school:

- Those whose students are more than one-third Māori.
- Those whose students are more than one-quarter Māori.
- Those whose students are more than one-quarter Pacific ethnicities.
- Primary schools.
- Secondary schools.

Using prediction model-based risk measures, schools with a relatively high proportion of Māori students are more likely to move downwards than are other types of schools (41% did so). This reflects the role played by Māori ethnicity in predicting a higher risk of non-attainment in our models.

In contrast, schools with a relatively high proportion of Pacific students are far more likely than other schools to move upwards (75% did so). This may be due to the fact that these schools tend to be located in low socio-economic areas, but their students have better qualification attainment rates than the students of other schools in low socio-economic areas.

The effects on the ranking of primary versus secondary schools is influenced by the choice of risk measure. As discussed above, the larger size of secondary schools will lead to greater averaging across different types of students, so secondary schools are less likely to be at the extreme end of a distribution. In addition high-risk students are less likely to remain at school throughout their upper secondary years, and this affects the composition of students who are aged 16–19. Secondary schools will therefore tend to have lower-risk students, on average, if we rank all students within a single risk pool. Alternative ways of ranking student risk could be considered, such as by year of age or by level within the schooling system (for example, years 1–6, years 7–8, years 9–10, year 11, years 12–13).

Other results, using schools' mean student risk score to rank schools, are given in Appendix 2. The patterns of change are similar if we rank schools by their mean student risk score.

2.3.4 Change in school ranking using alternative prediction models that do not include gender or ethnicity

We re-estimated the above results using risk prediction models that did not include gender or ethnic group. A full set of results from these alternative models is given in Appendix 2.

Removing gender and ethnicity reduced the statistical fit of the models by around 0.02 points (see Table 2.9).

Table 2.9. Area under the ROC curve for logistic regressions of NCEA level 2 attainment, for models without gender and ethnicity

Age	Training sample	Validation sample
5	0.7544	0.7469
6	0.7528	0.7519
7	0.7607	0.7499
8	0.7569	0.7535
9	0.7598	0.7524
10	0.7596	0.7579
11	0.7665	0.7533
12	0.7665	0.7601
13	0.7707	0.7616
14	0.7674	0.7692
15	0.7777	0.7684
16	0.7798	0.7691
17	0.7163	0.7353

Table 2.10. Movement of schools if current decile system is replaced by a decile ranking when prediction models exclude ethnicity and gender

	Number of schools					Percentage				
Change in decile	All	>33% Māori	>25% Pacific	Prim.	Sec.	All	>33% Māori	>25% Pacific	Prim.	Sec.
Move down	810	276	21	753	57	34	37	8	37	17
No change	807	321	84	708	99	34	43	32	35	30
Move up	735	159	156	564	168	31	21	60	28	51
Total	2,352	753	261	2,025	327	100	100	100	100	100
Down 5+ deciles	9	0	0	9	0	0	0	0	0	0
Down 4 deciles	9	3	0	9	0	0	0	0	0	0
Down 3 deciles	63	21	0	63	0	3	3	0	3	0
Down 2 deciles	180	57	0	174	9	8	8	0	9	3
Down 1 decile	546	195	18	495	48	23	26	7	24	15
No change	807	318	84	708	99	34	42	32	35	30
Up 1 decile	426	120	90	321	105	18	16	34	16	32
Up 2 deciles	204	27	45	156	48	9	4	17	8	15
Up 3 deciles	66	6	12	57	9	3	1	5	3	3
Up 4 deciles	36	3	9	27	9	2	0	3	1	3
Up 5+ deciles	3	0	0	3	0	0	0	0	0	0
Total	2,352	753	261	2,028	327	100	100	100	100	100

Using the alternative model parameters to predict each student's risk and then analyse the distribution of risk across schools, we find that a slightly higher proportion of schools remained in the same decile as currently (34% rather than 32%), and a lower proportion moved to a lower decile (34% rather than 37%). The proportion who moved up the decile system did not change. Schools with a relatively high proportion of Māori students were somewhat less likely to move to a lower decile, and schools with a relatively high proportion of Pacific students were less likely to move to a higher decile.

3 Further work to refine the models

Further work could be done in the following areas:

- Making the model estimation sample more comprehensive by obtaining achievement data for the students who were at schools offering non-NCEA qualifications.
- Improving the coverage and accuracy of the existing explanatory variables.
- When qualification attainment data for 2016 is available next year, updating the dependent variable to represent a final measure of the NCEA level 2 attainment (by age 18) for the 1998 birth cohort.

Increasing the coverage of the estimation sample

Students who were at schools offering non-NCEA qualifications were excluded from the model estimation sample because we don't have data on their international qualification attainment. If the relevant data can be put in the IDI, those students should be brought back in. Because they tend to be high-achieving students, adding them to the estimation sample probably won't have a large effect on the model results, however.

Improving the coverage and accuracy of the existing explanatory variables

Our administrative data models used the IDs of children's birth parents to construct the parent variables, such as the birth parent's earned income and offending history. The 1998 birth cohort is the first for which information on birth parents is available for all New Zealand-born children.

However, the school-aged population includes many children who were not born in New Zealand. In addition, at any point in time, many children will be living with caregivers who are not the parents named on their birth certificate.

We explored the latter issue using the 1998 cohort linked Census sample. Children's birth certificate parents and their Census night caregivers were compared:

- For 43% of children, both Census night caregivers were the same as the birth parents listed on the birth record.
- For another 35% of children, one Census night caregiver was the same as one of the parents listed on the birth record.
- For 22% of children, there was no Census night caregiver who matched birth certificate parents. This includes about 17% of children who had no New Zealand birth record.

When caregiver information contained in three sources – birth registrations, immigration applications and benefit data – was combined and the results compared to the Census night caregivers, the results were somewhat better:

- For 68% of children, both Census night caregivers were correctly identified using the administrative sources.
- For 24% of children, one Census night caregiver was correctly identified.
- For 8% of children, no Census night caregiver(s) was correctly identified.

We conclude that information on the identities of current caregivers that is contained in the migration and benefit data should be included in future work.

While including this information seems important in principle, the work we've done so far suggests it will raise the predictive power of the models only slightly. The impact is likely to be quite limited because of the range of information that is already included about migrant children (such as migrant category, source country and time in New Zealand) and children supported by a benefit (such as the duration and type of benefit). In addition, we have already found that adding information collected in the Census on current caregivers did not materially improve the predictive power of the attainment model at the age 14, compared with using administrative data for birth parents only.

Updating the dependent variable used in the modelling

Our qualification achievement data currently ends in 2015, the year when youth in the 1998 birth cohort turned 17. We used a modified dependent variable to account for the fact that some children will have returned to school in 2016 and gained NCEA level 2. Our dependent variable is a measure of not obtaining NCEA level 2 by 2015 *and* not returning to school in 2016.

When attainment data becomes available for 2016, the dependent variable should be updated to bring in this additional information.

The value of collecting additional data from parents

The work summarised in this paper has shown that students' risk of not completing NCEA level 2 can be predicted with moderate accuracy using the administrative data sources that are currently available in the IDI. Comparing models, we have found that their predictive power is only slightly altered by adding or subtracting different measures of child characteristics and parent characteristics. This suggests that improving the quality of these existing measures by, for example, requiring schools to record the names and birth dates of students' current caregivers so that the correct caregiver IDs are used, would probably not lead to much improvement in the predictive accuracy of the models.

The predictive accuracy is more likely to be significantly improved if we were to bring in different types of data that are more strongly correlated with future achievement, such as the child's own test scores at younger ages.

1 Introduction

It has been over a year since our last meeting and now is the time to provide you with an update on our Equity Index (EI) work. You will no doubt have seen a number of reports in the news since Minister Hipkins' announcement of the decision to replace the Decile system with the EI, back in September (see links below for some examples)

https://www.rnz.co.nz/news/national/399687/some-high-decile-schools-could-get-more-funding-under-new-model-minister

https://www.stuff.co.nz/national/education/116120556/northland-principals-at-loggerheads-over-scrapping-the-decile-funding-system

https://www.newstalkzb.co.nz/on-air/heather-du-plessis-allan-drive/audio/chris-hipkins-school-deciles-to-be-scrapped-from-2021-22/

At the end of 2018 we gave the Equity Index Technical Advisory Panel (EITAP) a technical reference document which outlined the methodology of the model at that time. We garnered your critique of this methodology, summarized this feedback into themes, and held a half-day meeting in Wellington with the panel to discuss. The final contact was at the beginning of February 2019, when we sent you a 'final feedback' document, alongside a timeline for the technical work-stream.

I would like to take this opportunity to thank you once again for your contribution and for agreeing to continue your role as we develop the EI.

In the year since the EITAP met our team have been busy with some substantial changes to the EI model, many as a result of feedback from the panel. At the end of 2019, we have changed the focus of our development work. We have developed a new version of the model itself. The technical EI work programme will continue until July 2020, with most of the future changes concerning how this model will be implemented.

Below is a brief summary of the major changes we have made to the model. We are currently developing multiple technical reports, which will outline all of these changes in greater detail.

Section 2 describes the development process. In Section 3, we outline our work to determine the best dependent variable for our analysis. Because we are moving from a binary variable, in Section 4 we discuss the choice of regression model. Section 5 describes our investigations into the inclusion of other predictors, namely housing, geographic and health variables. In Section 6 we look at other improvements we have made to the modelling, such as revision and development of the code, and the evaluation of model performance. Finally, in Section 7 we work through the implications of changes in modelling on which confidentiality protections are required.

2 Development process

We have been working together closely with the Ministry's Education System Policy group (ESP) on further developing the EI. While our team focused on the more technical aspects regarding the construction and development of the EI model, we also had to balance this with broader considerations which came from our Policy counterparts.

This meant the development process has been iterative, with some compromises being made in certain technical decisions based on broader policy rationale including sector engagement and feedback, funding implications and public acceptability.

Each of the following sections outline and explain where policy considerations have impacted technical decisions regarding the EI model.

3 Dependent Variable

One prominent piece of feedback that we received concerned the use of NCEA Level 2 pass/fail as a dependent variable. In summary, the feedback was that:

1. It may not be the best predictor of students' achievements in the future;
2. It may not be a stable measure over time, e.g., NCEA review;
3. The use of NCEA as a binary indicator results in an unnecessary loss in information.

The Ministry's response to these concerns, with emphasis on the final concern, was to explore the development of an alternative measure of educational success.

Deciding to move to a continuous measure of NCEA was a relatively straight-forward decision, but *how* to do this was not. When coming up with a measure of NCEA, we had to consider several different aspects of the system:

1. What NCEA levels to include? 1, 2, 3, or a combination thereof?
2. How to weigh achievement quality on standards, i.e., Achieved, Merit, and Excellence?
3. Should all subjects be weighed equally? i.e., Maths and English more valuable than other subjects?
4. Equal weightings of credits by difficulty of standard? i.e., can be expressed as a pass-rate, or expected percentile.
5. Should unit standards be treated differently to achievement standards?

In deciding the dependent variable, we had to balance three requirements (in order of importance): theory/precedent, 'sensibility' and performance.[1]

<u>Theory/precedent</u>

The first and most important requirement was whether or not some sort of theory or precedent existed around the dependent variable. In general, if a theory or precedent existed, we used it. The suite we developed all had some theoretical basis; Weighted Relative Performance Index (Crampton E, Udahemuka Martine, 2008) and Expected Performance Index[2] both had literature backing them, as does the use of NCEA[3] (Treasury 2018). Score-weighted NCEA has precedent in the sector as a 'rank score'.

[1] These three criteria were used for any modelling decision we had to make, for example in the weightings of the dependent variable, choice of independent variables, and the categorisation of the independent variables (see 6.2).

[2] Documentation wasn't published publically.

[3] Since significant sector engagement had already been done surrounding the original iteration of the Equity Index (as described in Treasury, 2018), we already had the signal that NCEA was accepted within the education sector as an appropriate measure of educational success.

'Sensibility'

The second requirement was to meet a 'sensibility' criterion. To ensure public confidence in the approach, the dependent variable we use needs to be reasonably interpreted as a sensible measure of educational success. While NCEA L2 was simple to understand and aligned with the previous government's Better Public Services target for secondary education, it did not necessarily meet these requirements.

This also encompasses navigating the political environment. Sensibility was guided by ESP, as they sought feedback on the issue of measures of educational success from the Sector Reference Group (SRG) on the Equity Index, and from Ministers.

Utilising insights from the SRG, as well as broader research under taken by the ESP group, the Ministry decided that the dependent variable selection needed to be underpinned by the following criteria:

- The measure should account for both quantity and quality of NCEA standards in order to better reflect the diversity of student achievement. The original binary NCEA L2 "pass/fail" measure did not account for this.
- The measure should provide an equitable representation of students.

Performance

The third requirement was that (given requirements of theory and sensibility) the chosen variable should perform as well as possible in a modelling context. We ended up with 4 different measures that fulfilled the theory and sensibility criteria (in red, below). However there was no commonly agreed upon theory on which of these is a better measure of educational success. As a result, we chose the measure that socioeconomic variables are most able to explain variance in, or predict educational outcomes. We are aware that this approach (X calibrates Y) is not a recommended statistical approach to use, however due to a lack of a preference amongst the candidates that passed the 'sensibility criteria' (see below), we believe this approach was not unjustified.

With these three requirements in mind, we did some exploratory analyses to develop our new measure of educational success. These analyses were done in three stages:

1. Develop the suite of candidates to test. This includes reviewing literature on already existing measures. These should all pass the 'sensibility' test.
2. Statistically test best performing candidate.
3. If applicable, optimise weightings on best performing candidate.

The following sections summarise the procedure we followed for each stage.

3.1 Stage 1: Develop Suite

Following a literature review, and review of the available data on NCEA, we used or developed 10 different permutations of NCEA. These were our best interpretation of the ways you could measure NCEA. These were:

- *Sum of achievement credits, weighted by Expected Percentile[4] (EXP)*
- *Sum of achievement credits, weighted by score and EXP*
- *Sum of achievement and unit standard credits, weighted by EXP*
- *Sum of achievement and unit standard credits, weighted by score and EXP*
- *Sum of achievement and unit standard credits, unweighted*
- *Sum of achievement credits only, unweighted*
- *Sum of achievement and unit standard credits, weighted by score*
- *Sum of achievement credits, weighted by score*
- *Weighted Relative Performance Index (WRPI)*
- *Average EXP[5]*

These were examined using all the possible combinations of L1, L2, and L3 credits, namely:

- L1
- L2
- L3
- L1+L2
- L1+L3
- L2+L3
- L1+L2+L3

So overall, we had 70 (10 x 7) educational measures of success to test.

It is important to note that prior to analysis, we were aware that several of our candidates would not have passed the 'sensibility test' (in red). Education outcomes can be measured in a wide variety of ways, particularly in the context of NCEA, that it was important to consider how any candidate dependent variable could be representative of the target student population. It is important that this measure accounted for the key ways that inequity manifested in differential NCEA outcomes, namely the different types and levels of standards that students are entered in as well as their quality and-= quantity of attainment. This is why, according to the guidelines[6] provided by ESP, a suitable candidate must comply with the following rules:

- The measure should involve some measure of quality.
- The measure should consider both achievement standards and unit standards.

[4] See Appendix A – EXP and WRPI for details on how we applied EXP to unit standards, as well as details on WRPI.
[5] This measure performed so poorly that we didn't include it in the analysis. This was a 'quality-only' measure – meaning you can pass one credit with excellence and your average EXP is higher than that of another student who, for example, has passed 299 credits with excellence and one with merit, making it unsuitable for this project.
[6] Guidelines were provided as a combination of advice based on literature reviews and policy considerations.

- The measure should use a combination of NCEA level 1 and/or level 2; not level 3.

Beyond this, ESP were willing to consider the candidates that fulfilled these requirements. These requirements meant that measures like *total achievement credits* (weighted or unweighted), *WRPI* and total credits would be harder to justify, even if their performance was the best.

3.2 Stage 2: Choose the Best Performing Candidate(s)

Given the 70 candidates described in Section 3.1, we wanted to see how each candidate was able to be predicted by our chosen suite of socioeconomic predictors. Testing model performance involved running each of the candidates in a regression model (Poisson regression, constant independent variables) and measuring their goodness-of-fit. We used multiple measures of goodness-of-fit[7], and looked for candidates which consistently performed well across most of the measures.

The two measures that consistently performed top or second-best were *sum of NCEA achievement credits* and *sum of NCEA score-weighted credits,* respectively. For both these measures, NCEA level 1, 2 and 3 was usually the best NCEA combination, with NCEA level 1 and 2 being slightly below this performance.

With consideration to policy restrictions, the measure we decide to settle with was *sum of NCEA level 1 and 2 score-weighted achievement credits* – a slight compromise to the very top performer. The attached HTML files show visualisations of the goodness-of-fit, and attached is interactive versions of these plots.

3.3 Stage 3: Optimise weightings

With the decision to use *sum of NCEA level 1 and 2 score-weighted credits*, we needed to choose an optimal weighting set to use. Initially, we intended to use the well-established university rank scoring point system[8] that some NZ universities use to determine entrance into particular undergraduate qualifications (with the weightings of 2 for Unit standards, W_U, 2 for Achieved, W_A, 3 for Merit, W_M, and 4 for Excellence W_E, achievement standards); these were the weightings we used in the analysis in Stage 2. However, with the rank scoring system having no consistent guidance[9] as to how to weight unit standards differently from achievement standards (and with the inclusion of unit standards being a must from policy, to ensure coverage of all students), we decided to perform an analysis by allowing these weightings to be float or 'data-driven' and emerge from simulations. In this way, we could determine what set of weightings maximises the model performance, i.e., the resulting dependent variable would best be able to be predicted by socioeconomic variables (see our note in section 3). We would also bear in mind how much this 'optimal' set deviates from the traditional 2, 2, 3, 4[10] set (especially the last 3). Our final decision would be based on a trade-off between model performance, and theory, or established accepted weightings already used by the sector (policy-driven). We constructed different dependent variable candidates with different weightings for numbers of U, A, M, and E NCEA credits passed by students using

[7] Our chosen measures of goodness-of-fit were pseudo R^2, mean-normalized root mean squared deviation, IQR-normalized root mean squared deviation, Spearman's coefficient and to a lesser degree -Pearson's coefficient.

[8] See for example:
https://www.auckland.ac.nz/en/study/applications-and-admissions/entry-requirements/undergraduate-entry-requirements/new-zealand-secondary-school-applicants/national-certification-educational-achievement.html

[9] The inconsistency arises due to the fact that some universities accept a list of a few approved unit standards with a weighting of 2 (these lists are no necessarily the same), and some do not accept any, altogether, see for example:
https://www.otago.ac.nz/study/entrance/otago013108.html

[10] It is worth nothing that we had ($W_A = 2, W_M = 3, W_E = 4$) as *strong* preferred weights. Model performance drop would have needed to be significant for us to deviate from these. We were lucky in that we didn't have to compromise model performance in order to accommodate these priori weightings (i.e., theory aligned with the data).

$$D_c = W_U \times N_U + W_A \times N_A + W_M \times N_M + W_E \times N_E$$

where D_c is a dependent variable candidate, W_i is the assigned weighting and an integer ranging from 0 to 10 for a particular 'quality' of achievement, and N_i is the corresponding number of credits passed, for credit quality type $i = U, A, M,$ or E. The total number of candidates with all the possible first-degree linear combinations of weightings were $11^4 - 1$ (excluding the degenerate case of 0, 0, 0 and 0), or 14,640, of which 13,025 sets provide unique sets of weightings, e.g., the set of $W_U = 1$, $W_A = 2$, $W_M = 3$, and $W_E = 4$, is equivalent to the set of $W_U = 2$, $W_A = 4$, $W_M = 6$, and $W_E = 8$. We ran the regression process on all these cases and recorded their corresponding regression performance using a set of statistical measures for goodness-of-fit. Thereafter, we estimated the local maxima of different regression performance measures on this $11 \times 11 \times 11 \times 11$ grid of weighting space. The results of our analysis show that the maximum regression performance occurs around the weighting set point at $W_U = 0.45$, $W_A = 2.05$, $W_M = 3.35$, and $W_E = 4.15$.[11] We have therefore, selected the sum of NCEA level 1 and NCEA level 2 credits weighted in the following way:

- Unit Standard (U) = 0.5
- Achieved (A) = 2
- Merit (M) = 3
- Excellence (E) = 4

One major difference between our measure and the university rank score measure is that we don't differentiate between university-approved[12] subjects, and those that are not. Given that unit standards are disproportionately represented within the non-university approved subjects, this distinction could partially justify weighting unit standard achievements as one quarter of achievement standard achievements. Most universities that use rank scores (like the University of Otago) don't consider unit standards at all when generating a rank score.[13]

We performed the analysis for the 1998, 1999 and 2000 birth cohort training sets. The resulting statistically optimal weighting sets were relatively stable over time with no changes over 5%.

4 Regression Model

We replaced a logistic regression (NCEA level 2 achieved/fail) with a Poisson Linear Regression. This change was in response to wanting a continuous educational outcome measure. The Poisson Linear Regression was able to model the distribution of our dependent variable better than a standard linear regression as the former is a preferred regression method to model a count-type response. We have also evaluated the use of negative binomial regression, or generalised Poisson regression, as well as zero-inflated Poisson regression, since there are a moderate number of students with no recorded passed credits. However, we did not observe a significant improvement in the regression

[11] No condition was set that $W_U < W_A < W_M < W_E$.

[12] I.e. approved as a subject that counts towards the University Entrance qualification. University rank score measures are usually calculated only from approved subjects.

[13] This is because all the University Entrance approved subjects only have Achievement Standards associated with them (see https://www.nzqa.govt.nz/qualifications-standards/awards/university-entrance/approved-subjects/).

performance, and therefore decided to implement the least numerically intensive algorithm, i.e., the Poisson regression.

5 Inclusion of Other Independent Variables

A main feedback the panel gave us was to include new variable domains: health, housing and geographic. However, we weren't able to include any of these domains for reasons listed below. Given that the model is a dynamic one – in the future, some, or all of these domains could be added to the model.

5.1 Housing

The data available in the IDI surrounding housing were:

1. Census
2. Social Housing Data (SHD)
3. Tenancy Bond Data (TBD)

Census provides a very good insight into households, as it is defined as a household level. One of the variables we would have liked to use - 'household crowding' was used in the Decile calculation, meaning we already have precedent (and code) for its use. However, an early decision made on Equity Index was that no Census data would be used. This is for several reasons; primarily, a reliance on Census would make the project unstable as a year-year measure. Every 5 years we would observe a sudden change in a sub-set of our independent variables. The case for the use of administrative data *only* for the Equity Index was also made in Treasury, 2018.

SHD and TBD are the two administrative datasets available to us in the IDI, so they don't suffer from the same problems as Census described above. SHD is a dataset containing details about individuals waiting for social housing, and those in social housing. TBD contains information on bonds lodged for residential tenancy agreements. Both these datasets are from 2000 onwards, meaning for the 1998 birth cohort we would have no data from children ages 0 -2 (our lowest coverage data-set).

A general issue with housing-related measures was that the Equity Index model is set up at a child-biological parent level. If a child lived with their grandparents, we would not be able to retrieve the child's actual housing information– instead we could have information on a bond for a house the child wasn't even living in (like if their biological father had lodged a bond).

Due to these limitations, we prioritised the development of Health measures over SHD and TBD, but due to time constraints we were unable to develop measures for them. These data-sets may prove to be a useful addition into the Equity Index in the future.

5.2 Geographic

The geographic variables we had originally used were:

- Number of home changes
- Number of school changes (any)
- Number of structural school changes
- Number of simultaneous home and school changes

Later on we replaced 'structural school changes' with 'non-structural school changes'. Non-structural school changes have a stronger socioeconomic predictive value and so are more sensible to include – changing school because a student became year 7 (and hence moved to an intermediate school) didn't seem like a socioeconomically related event, but changing school because a student's family is transient did. We also removed 'number of school changes' since this was simply the sum of 'number of non-structural school changes', and 'number of structural school changes'.

Asides from the above, we do not have access to any useful additional geographical measures to add to the model. One possible candidate could be an urban/rural indicator for the addresses linked to the child. Statistics New Zealand has released urban/rural boundaries since 2017, but this information isn't old enough to be used for children at earlier ages, as of 2020. We also have the school isolation index with 20 years history, but school-level indicators are out of the scope of this project[14], and this measure is Census-reliant. Furthermore, geographic measures suffer from the same fundamental problem as housing measures – that we do not have a construct of the 'household'.

5.3 Health

For health, the only dataset we could use was publically funded hospitalisations, with observations from 1998. There were many other datasets we would like to have used, such as Primary Health Organisation (PHO) enrolments and the Before School Check (B4SC), but these only had coverage from 2004 for PHO and 2011 for B4SC. It would be possible to include these variables in the future, once the birth cohort exceeds the coverages in the respective datasets.

Within publically funded hospitalisations, following Baker, Zhang, & Howden-Chapman (2010) we developed several different indicators, for both parents and children. These indicators were:

- Number of Ambulatory Sensitive Care Hospitalisations (ASH) of [child, mother, father]
- Number of Housing-related Potentially Avoidable Hospitalisations (HRPAH) [child, mother, father]
- Number of hospitalisations of child due to infectious disease
- Number of hospitalisations of child due to respiratory disease
- Number of hospitalisations of [mother/father] due to infectious disease
- Number of hospitalisations of [mother/father] due to cardiovascular disease

A few potentially useful measures to incorporate in the model were rejected because ESP determined that there was a high risk of potential stigma and/or the link to socio-economic status wasn't clear. Several other measures had stricter rules surrounding their use and were consequently excluded. The measures that we considered but did not use in the model were:

- Number of home injury of child (accidental poisoning, falls, etc.)
- Number of hospitalisations of [mother/father] for mental health reason (including assault and self-harm)

[14] School level measures are out of scope since the Equity Index is supposed to be a model in which students contribute to a school level of disadvantage for funding. If a school level measure was included in this contribution, it would be double-counting in a way. A better way would be to measure the isolation of the student's household instead (which has its own problems, as mentioned in 5.1).

- Number of hospitalisations of child for mental health reason (including assault and self-harm)

We have consulted with the Ministry of Health (MoH) regarding the implementation of these additional variables. On further discussions with MoH, concerns were raised around the appropriateness in the use of current Health variables available in the IDI. Their main concern is the utilisation of data where there is uncertainty around the causal direction and the potential of using potentially spurious correlations to predict outcomes. Because of this, the inclusion of Health-related variables at the present point in time would run the risk of criticism from epidemiologists. Given the marginal benefit of including the Heath variables versus the advice we have received around the appropriateness of including these in the model, these variables have not been included.

The recommendations we received to remedy some of these problems were not viable in the later stages of the project. With this in mind, since we already had the data prepared, we did some modelling work.

Using health variables we ran several models under several conditions:

1. Cumulative counts (lifetime) versus within-age counts
2. With or without restrictions (overnight stay minimum + principle diagnosis only)
3. Categorising counts or not (i.e., 1-3, 4-6)

As it is presented in Table 1, overall, the health data we were able to include in the model did not improve the model performance significantly, despite having a positive contribution to the explanatory power of the model, given all other variables. As a result of both the advice from MoH and the poor modelling performance, we decided to not use health data at this stage. Given a broader set of data, and in anticipation of future engagement with MoH, we should be able to include some health information in the EI in the future.

Table 1: Health model performance. Note 1: These findings are for a health model which also includes the restricted variables, like home injury – initial results were so poor that we added these in, in an attempt to 'boost' the model. Note 2: In every situation, categorisation of counts always performed better.

Cumulative	Restrictions	Categorised	R^2 Health Stand-alone	R^2 Full-model	R^2 Full-model + Health
✓	✓	✓	0.07	0.439	0.442 (+0.003)
✓		✓	0.08	0.439	0.444 (+0.005)
	✓	✓	0.01	0.439	0.439 (+0.000)
		✓	0.08	0.439	0.444 (+0.005)

6 Other Improvements

Besides those above, we made various other improvements to the model; improving the code that generated our data, improving our interpretation and use of administrative data, improving model performance, improving data representation for modelling, and various other improvements. Overall, these changes did two things:

1. Improved our ability to make predictions.
2. Reduced the risk that something could go wrong in the future.

Below is a summary of some of the more important changes we made, as well as a performance evaluation on the effect of these changes.

6.1 Code Revision and Development

Out of any task done, we spent most time on improving the code which generated our data; this took 5 months. Code development covered every facet of the project, from data cleaning, development or refinement of business rules, improving code-efficiency, fixing bugs and improving code readability. Below is a quick summary of the code development in 2019:

- There were 133 issues or suggestions regarding the code. Appendix B: Example data cleaning issue, shows an example of a series of related issues on social welfare business rules.
- In total, without any embedded comments, the script is around 14,700 lines, with around 300 tables created in the process.
- Reduced total run-time from 9-12 hours down to 2-3 hours.
- As a result of this work, the model performance has been significantly improved.

6.2 Categorisation of Independent Variables

In an effort to improve the model performance, we wanted to represent our information in a way that lent itself best to a modelling framework. One major improvement we were able to make was to transform the continuous independent variables we had (such as income) into theory-based or data-based optimal bands. This has improved the modelling performance at least in two major ways; the dependence of educational outcome on socioeconomic variables is inherently complicated and cannot be fully explained using a simple linear relationship. Converting continuous variables to categorical, by retaining a sufficient number of levels, offers a computationally cheap and reliable way to accommodate the non-linear nature of the relationship in the straightforward framework of the first-degree linear regression, without losing too much information. In addition, such a categorisation will resolve the problem of dealing with extremes (outliers), out of the box.
We avoided a non-linear model for several reasons:

- Project can become overly complicated.
- No strong theoretical priori estimates for non-linear transformations of independent variables (i.e., do we square? Cube? Log? Why?); work required to develop these theoretical bases takes more time than we had.
- Risk of overfitting model in choice of non-linear transformations.
- Risk of non-linear transformations may not be stable over time.
- Addition of extra variables.

Categorisation of different variables was done either in a theory-driven manner (in some cases such as the age at which immigrant students received their first visa[15]), or if no theory stood out, a data-driven approach (in others such as parental income variables).

The optimised data-driven categorisation was done using the following procedure; the variable of interest is categorised using $n-1$ number of random breakpoints, resulting in n brackets (or levels). Each breakpoint must conform two rules; it must be bigger than the previous breakpoint, and the resulting category must have a pre-defined sufficient number of points within it. For example, a

[15] Our theory for this particular variable was around school transitions of students acting as key ages (entry into primary, intermediate, secondary school).

random set of 2 breakpoints for the father's income is $5000 and $40,000, resulting in 3 income brackets, i.e., $0-$5,000, $5,000-$40,000 and above $40,000 levels, provided enough number students belong to each level. The procedure is repeated 100,000 times, and the regression performance metrics are recorded. In practice, generally, all the breakpoints are in a converging path to an optimal value once they are sorted by their corresponding regression performance metrics. One hundred breakpoint sets which generate the best regression performance are selected and the median[16] values of the 1^{st} to $n-1^{th}$ breakpoint are computed and used as the optimal set. Using this procedure, we have improved the regression performance (and therefore the prediction accuracy) by at least 25%.

To evaluate whether or not this improvement is the result of overfitting, we performed the algorithm for 1998 cohort training and used the resulting breakpoints for 1999 (and 2000) cohort training set, as our cross-validation sample. In each case, the expected regression performance loss was no more than 1%. Furthermore, repeating the procedure for 1999 (and 2000) produced similar set of breakpoints for each variable.

Here we should mention that using the described procedure requires one (somewhat arbitrary) user-defined input: the number of breakpoints, $n-1$. In practice, we noticed that increasing this number beyond a certain point does not have a significant performance benefit, so we break the procedure when the performance gain is less than 0.01% (when the number of breakpoints is incremented by one).

6.3 Parental and non-parental groups

We have two distinct groups within our training sample and scoring population. One group contains a linkage to at least one parent, the second group contains no linkage to either parent. We call these models the *parental model* and *non-parental model*, respectively. We decided to treat these two groups differently, separating them into two different models. The reason we decided to separate these groups are threefold:

1. A large number of our variables are parental. For the non-parental group, this would make a majority of their variable observations as redundant, reducing model efficiency.

2. The meaning of some of our variables are distinctly different for each of these groups. This is because the types of students who have missing parental data are distinct from the rest of the population – specifically, a majority of them are immigrants. This fundamental difference between the two groups affects the meaning or nature of some of the variables we use in the model. For example, proportion of life overseas for a NZ-born student may indicate a socioeconomically advantaged life for that student, however, the same may not be the case for a refugee student. Choosing to include both groups in a unified model would contaminate the coefficients of some of our covariates.

3. Following from the previous point, since a majority of the students in the non-parental group are immigrants, we were able to include a few variables which could not be included in the parental model. These variables are *age at visa approval, migrant status,* and *region of birth (international)*. These three variables are not meaningful for the general population, but are so for this non-parental model. *Appendix C: Variable lists for parental and non-parental models* holds a full list of variables for both the parental and non-parental models.

[16] Empirically, we found the median a more appropriate measure than other candidates, e.g., mean or mode.

The parental model accounts for the majority of our population, at around 83%, for both the 1998 and the 1999 birth cohorts. This model uses a wider array of independent variables, namely parental-related variables. The combination of more independent variables, and a much larger sample means that the parental model performs better than the non-parental model (see Table 2 for a brief comparison of model coverage and performance).

The non-parental model accounts for a minority of our population, at around 17%, for both the 1999 and 1998 birth cohorts. This model uses a reduced selection of independent variables, with 'region of birth', 'migrant status', and 'age at visa approval' added as additional variables. Due to a smaller population, and reduced set of variables, this model does not perform as well as the parental-model.

Table 2: Model performance and student counts for parental vs non-parental models (1998, 1999)

1998 Birth - April19 Refresh	Parental Model	Non-Parental Model
Count	43,422	8,862
Percentage	83	17
McFadden Pseudo R^2 (Age 19)	0.42	0.24
McFadden Pseudo R^2 (Age 5)	0.34	0.17

1999 Birth - April19 Refresh	Parental Model	Non-Parental Model
Count	44,439	8,922
Percentage	83	17
McFadden Pseudo R^2 (Age 19)	0.42	0.24
McFadden Pseudo R^2 (Age 5)	0.34	0.17

7 Changes to confidentiality

With the move towards a non-binary educational outcome, the nature of the privacy rules in place had to move too.

Originally, the EI output from the IDI was a 'concentration' measure; concentration of disadvantage at a school means the *percentage of children who we have deemed to be disadvantaged*. Because of the identifying nature of our previous output (i.e. it could be reasonably inferred that a child attending a school is considered disadvantaged or not by the model), strict rules were placed on our outputs to protect these children's privacy.

These rules were:

- If a school has 5 students or fewer - supress all results.
- If a school has between 6 and 20 students - only release an estimated concentration value, based on a linear regression that predicts concentration based on mean.
- All counts used to calculate concentration should be randomly rounded to base 3 (RR3).
- If the concentration of a school is greater than 80% - supress to "80%".
- If the concentration of a school is less than 20% - supress to "20%".

All of these rules were in place to protect the privacy of the children used in this study. However, the rules did introduce some issues for operational reality. A school with a small non-supressed roll could get up to +/- 10-15% change in funding due to random chance alone, between IDI refreshes from the

same year (due to RR3 impact only). Furthermore, there were too many schools whose values were either fully supressed, or estimated using a linear regression – meaning we wouldn't be able to fund them using their EI values.

Once we had developed our new educational measure (weighted sum of NCEA L1 & L2 credits), we approached Statistics New Zealand (SNZ) to re-develop an agreed-upon confidentiality rule set. We provided evidence that the current EI is not as nearly as identifying as the previous concentration measure due to the process the model output undergoes in its transformation into the EI – the value we release from the IDI.

The steps involved in creating this current EI is summarised below:

1. Model output – estimated weighted sum of NCEA level 1 and 2 credits, at a student level.
2. Student (predicted) scores are moved to a scale from 0 to 1, and inversed (so that 1 is the lowest estimated score).
3. Outliers are removed from both ends of the distribution using sigma-clipping algorithm. Outlier values are set to the boundary values.
4. Distribution is re-normalized to $[0, 1]$.
5. Aggregate by mean to school level distribution.
6. School distribution is scaled to $[k, k + m]$ space, where k and m are arbitrarily decided upon. This is what the final EI values are.

After consultation with SNZ, we were able to move to less stringent confidentiality rules (rule 5.9 – regression models[17]). The only effect of these restrictions was that:

- Values from school with less than 5 students will be suppressed.

Although this single rule does affect fewer than 10 schools in our output, the degree to which confidentiality protections affect the quality of model outputs now is significantly smaller.

It is important to note that it is very likely that the implementation of the EI for funding purposes would involve a student roll-weighted score across 2-4 years for smaller schools. We have had preliminary discussions with SNZ about treating the counts relevant to confidentiality as the sum across these 2-3 years; this would allow us to treat the roll of a school across multiple years as one 'super-roll'[18]. If we combine rolls across multiple years, then it would be very unlikely that any school would have a supressed EI score year-on-year. Else, it would remain that a small number of schools will not be able to have their EI scores released.

[17] See https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/ for the full rule set (Microdata output guide - Fourth edition).

[18] There are some nuances here around the same student attending in each year, which need to be still resolved. A likely solution is to look at the 'unique' roll across these years, i.e., only count the same student once.

References

Baker M, Zhang J, Howden-Chapman P. Health status of Housing New Zealand applicants and tenants: Key indicators for 2004-2008. Wellington: He Kainga Oranga/ Housing and Health Research Programme, University of Otago, 2010.

Crampton E, Udahemuka Martine. SCORE! Transforming NCEA Data, The New Zealand Initiative, 2008.

Maloney, T. and Singh K. (2017) Using Validated Measures of High School Academic Achievement to Predict University Success. AUT School of Economics Working Paper Series, 2017/10 (AUT School of Economics, December 2017).

Shulruf, B., Hattie, J. and Tumen, S. (2008) Student pathways at university: patterns and predictors of completion. Studies in Higher Education, 33(3): 233-252.

Shulruf, B., Li, M., McKimm, J. and Smith, M. (2012) Breadth of knowledge vs. grades: what best predicts achievement in the first year of health sciences programmes? Journal of Educational Evaluation for Health Professions, 9(7): 1-9.

Dixon, S., (2018). Student Mobility across Schools and its Links to Underachievement. NZ Treasury Working Paper 18/01. https://treasury.govt.nz/publications/wp/wp-18-01-html

Appendix A – EXP and WRPI

Expected Percentile (EXP) was originally developed by Dr Michael Johnston at NZQA. EXP calculates a measure of relative difficulty, within an achievement standard, of achieving certain grades. Specifically, EXP calculates the mid-point between the cumulative percentage of the grade attained, and the grade below this, within an achievement standard. Table 3 shows an example calculation, taken from *Post-school choices: How well does academic achievement predict the tertiary education choices of school leavers? (MoE, 2008):*

Table 3: Results distribution of a sample achievement standard.

Result	**Frequency** A	**Percentage** B	**Cumulative Percentage** C	**Expected Percentile** $D_i = (C_{i+1} + C_i)/2$
Excellence	12	10%	100%	95%
Merit	24	20%	90%	80%
Achieved	60	50%	70%	45%
Not achieved	24	20%	20%	10%

We applied EXP in 4 different ways, with each applied for every possible combination of level 1, 2 and 3 standards:

- Sum of credits, weighted by EXP
- Sum of credits, weighted by score and EXP
- Sum of credits, weighted by EXP (unit standards included)
- Sum of credits, weighted by score and EXP (unit standards included)

In the original definition of EXP, unit standards were strictly excluded in its definition. We defined a unit standard modified EXP for the sake of this analysis to see if we were able to improve predictive ability. This modified EXP essentially calculated EXP the same as above, but with only two levels. Table 4 demonstrates of an example.

Table 4: Results distribution of a sample unit standard.

Result	**Frequency** A	**Percentage** B	**Cumulative Percentage** C	**Expected Percentile** $D_i = (C_{i+1} + C_i)/2$
Achieved	8	80%	100%	60%
Not achieved	2	20%	20%	10%

Both measures with unit standards included, performed worse than their non-unit standards counterparts. Analysis on all of these measures is explored in Section 3.2.

The second measure of standard difficulty we examined was the *Weighted Relative Performance Index (WRPI).* This was developed by The New Zealand Initiative in 2018 (see Crampton and Udahemuka 2018, for more details).

From this paper, WRPI for a student is characterised as

$$WRPI_j = \sum_{i=1}^{n} \alpha_i \ln x_{i,j},$$

where α_i represents the number of credits attained for standard i, and $x_{i,j}$ represents the relative performance on standard i by student j, and is given by:

$$x_{i,j} = \frac{\text{number of students who sat standard } i}{\text{number of students who recieved the same or better grade than student } j \text{ in standard } i}$$

WRPI is at a student level, not a student-standard level; as such, we only used one measure of WRPI – WRPI itself.

Appendix B: Example data cleaning issue

There seems to be a lot of problems with benefit business data rules. These potential problems will be added to this card in the order in which they were found. The following link provides useful information on 'new' benefit categories as of July 2013:

https://web.archive.org/web/20141001135912/http://www.workandincome.govt.nz/individuals/benefit-changes/new-benefit-categories.html

List of potential problems:

1- Benefit type code, 602: According to an online document Labour introduced the Job Search Allowance for up to thirteen weeks for people who had been made redundant after at least five years in the workforce:

http://img.scoop.co.nz/media/pdfs/0810/job_search_allowance_factsheet.pdf

As you can see this type of benefit is not exclusive to the youth population, but in the code it is considered as '1: YP Youth Payment Related'.

2- Benefit type code, 603: Youth/Young Parent Payment. In the code we read:

case when event type in ('602', '603') and event_type_2 <> 'YPP' then '1: YP Youth Payment Related'

The code 603 contains 10,368 'YPP', 16,499 'YP', and 97,801 NULL rows for event_type_2. The above code will classify all the rows with NULLs in even_type_2 as '1:YP Youth Payment Related', which doesn't seem right.

3- Benefit type code, 313: Emergency Maintenance Allowance is assistance that may be paid to SOLE PARENTS who do not qualify for any other payments. See:

https://www.workandincome.govt.nz/products/a-z-benefits/emergency-maintenance-allowance.html

In the code we read:

when event_type in ('313') or event_type_2='YPP' then '1: YPP Youth Payment Related'

For all the rows with event_type = '313' event_type_2 is NULL. That means that the mentioned code will capture all these rows and classifies them as '1: YPP Youth Payment Related'. From the above definition of code 313 it is clear that these rows should be classified as '4: Sole Parent Support Related' not '1: YPP Youth Payment Related'. This is an important blunder since '4: Sole Parent Support Related' will be translated to a variable in #panel_age.

Appendix C: Variable lists for parental and non-parental models

Variables in parental model:

- CYFS notification
- CYFS investigation
- CYFS family group conference
- CYS placement

- YJ notification
- YJ investigation
- YJ family group conference
- YJ placement

- Proportion of life on benefit
- Proportion of life overseas

- Number of home changes
- Number of school changes
- Number of structural school changes[19]
- Number of simultaneous home and school changes

- Fathers education level
- Mothers education level

- Fathers community service (any)
- Fathers proven charges (any)
- Fathers custodial sentence (any)

- Mothers community service (any)
- Mothers proven charges (any)
- Mothers custodial sentence (any)

- Fathers First-Tier Entitlement benefit amount
- Fathers Second-Tier Entitlement benefit amount
- Fathers Self-employed income amount (ignoring negative income)
- Fathers wage and salary income amount

- Mothers First-Tier Entitlement benefit amount
- Mothers Second-Tier Entitlement benefit amount
- Mothers Self-employed income amount (ignoring negative income)
- Mothers wage and salary income amount

[19] There is current work to move towards a more concise basket of home and school related indicators (namely the replacement of structural with non-structural, and removal of 'school changes'). Statistics in this document are based on the listed set, however.

- Ethnicity (prioritized: Maori > Pasifika > Asian > Other >European > Missing)

- Age of mother at birth
- Age of mother at first child
- Number of siblings at birth
- Age of father at first child

Variables in non-parental model:

- CYFS notification
- CYFS investigation
- CYFS family group conference
- CYS placement

- YJ notification
- YJ investigation
- YJ family group conference
- YJ placement

- Proportion of life on benefit
- Proportion of life overseas

- Number of home changes
- Number of school changes
- Number of structural school changes
- Number of simultaneous home and school changes

- *Region of birth (international)*
- *Migrant status*
- *Age at Visa approval*