Te Poutāhū
Curriculum Centre

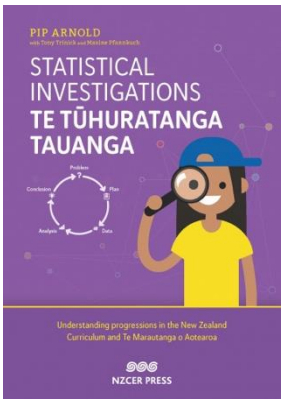
Statistical investigations

Teacher resource book

Te Kāwanatanga o Aotearoa
New Zealand Government

Te Mahau

Te Tāhuhu o te Mātauranga
Ministry of Education



Note: In 2023, every secondary school in Aotearoa New Zealand received a copy of the NZCER publication *Statistical Investigations*, by Dr Pip Arnold.

Throughout this resource you will see page numbers that correspond to information from the Statistical Investigations book.

Contents

Curriculum progressions for statistical investigations

Kaupapa | Problem

Situations for statistical investigations (Arnold, 2022, p. 34)

In Years 9 to 11, students are exploring all situations for observational studies. They are moving towards exploring populations using samples for summary and comparison situations. They are starting to explore paired comparisons for experiments. For observational studies it is expected that data sets students work with are multivariate. This provides opportunities to explore different investigative questions and to 'wonder' further about what the data is showing.

For experiments, it is expected that students will collect only the data they need to answer their investigative question. They may wonder what different conditions or treatments could show in an experiment. This could lead to further investigations. In Year 11, students are identifying the situation (summary, comparison, relationship, time series) for observational studies to solve the problem.

Investigative questions (Arnold, 2022, p. 43)

Students identify a broad area to explore using the statistical enquiry cycle. They:

- identify the population of interest and pose investigative questions about the population of interest (summary and comparison situations)
- use samples they have collected or that have been provided to answer their investigative questions
- pose investigative questions about relationship and time series situations.
- Students make predictions or assertions about what they expect the data to show. This is also reflected in the investigative question posed. For example, in a comparison situation they would write the investigative question signalling their expectation of the bigger/longer/larger group for the variable.

At this level, students are posing their own investigative questions. For summary and comparison situations they are posing investigative questions about populations. For relationship and time series situations they are posing questions about the entire data set. They are interrogating and correcting their own and others' investigative questions using the criteria for what makes a good investigative question.

Thinking about what the data might show (Arnold, 2022, p. 56)

At these year levels students are moving towards posing investigative questions about populations for summary and comparison situations. In both situations, they should be able to sketch the shape of the distribution for numerical data based on previous experiences with similar variables. When posing comparison investigative questions, they should pose their question to match their hypothesis of who is taller, heavier, longer, faster etc. For example, students might pose the following investigative question:

"I wonder if Aotearoa New Zealand Year 10 students tend to have a longer sleep time than Aotearoa New Zealand Year 13 students."

This investigative question predicts that Year 10 students will have a longer sleep time than Year 13 students. If they believed it to be the other way around, then the investigative question would be:

“I wonder if Aotearoa New Zealand Year 13 students tend to have a longer sleep time than Aotearoa New Zealand Year 10 students.”

Whakamahere | Plan

Methods of data collection (Arnold, 2022, p. 87)

In Year 11, students are planning and conducting observational studies and experiments using the statistical enquiry cycle.

Ethics (Arnold, 2022, p. 98)

From Year 9, when students are planning a statistical investigation ethical practices should be included as part of their planning processes. This may be through a set of interrogative questions that act as checks and balances. They may also use a checklist to confirm that different ethical practices have been considered.

“Who” to measure (Arnold, 2022, p. 105)

Students can identify the population of interest. They are also selecting and using appropriate sampling methods. For example, they use simple random techniques to select a sample to answer their investigative question (summary and comparison situations). They understand that a sample is used to answer investigative questions about a population. They can select a sample for primary data collection and can select a sample from a secondary data set. In paired comparison experiments, students are randomly allocating the order of treatment to experimental units in paired comparison experiments. They are working to balance variables that cannot be controlled.

What to measure (Arnold, 2022, p. 121)

Students are selecting, using, and justifying variables and the measures that they will use to answer the investigative question.

Raraunga | Data

Collecting/considering data (Arnold, 2022, p. 168)

In Year 11, students are deciding on their data collection strategy to answer their investigative question. This includes deciding if they need to take a sample or not, and whether they should collect primary data or source secondary data. They should be able to critique data sources as to their fitness for purpose.

Data management (Arnold, 2022, p. 183)

In Year 11, students should be efficiently managing their data from source to analysis. They can sort and recategorise their data to answer their investigative questions. They recognise that the metadata or data dictionary gives them important information. They can create these for their own datasets and know how to find them for secondary datasets.

Cleaning data (Arnold, 2022, p. 193)

Students are introduced to formally cleaning data from Year 9. Year 9 to 11 students are identifying data that might need cleaning by initially graphing the data and exploring unusual responses. When a data point has been identified as incorrect, students are able to make simple edits. They can decide whether to remove the data point (pairwise deletion) or the whole case, ie the individual or row (listwise deletion).

Tātari | Analysis

Data visualisations (Arnold, 2022, p. 285)

Students in Year 11 are selecting appropriate data visualisations to explore the data and communicate relevant detail and overall distributions. They are exploring summary, comparison, relationship, and time series situations. Students in Year 11 link multiple visualisations and can make connections between them.

Describing distributions (Arnold, 2022, p. 338)

Students in Year 11 are using multiple data visualisations to create a rich description of the data to answer their investigative questions. Context is interwoven into the description using variables and populations/groups of interest. Values and units are used for numerical data. Interpretations and explanations in context are evident in the descriptions.

Reasoning with data (Arnold, 2022, p. 358)

In Year 11, students are expanding their ability to ‘make the call’ in comparison situations by taking sample size into account. They are becoming familiar with the decision guides in Year 11 for different sample sizes. In Year 11, students can start to use the least squares regression line to find a predicted value for a given x value for relationship situations. They should be starting to think about the prediction interval in a visual way, obtaining the prediction interval from the scatter plot.

Whakatau | Conclusion

Answering the investigative question (Arnold, 2022, p. 388)

In Year 11, students can provide a justified conclusion that answers the investigative question. For sample-to-population inference situations (summary and comparison), they can generalise beyond the sample to the population, provide evidence, and account for uncertainty. They can provide explanations for observed patterns in the data. Students critically evaluate their investigation.

Communicating findings and writing reports (Arnold, 2022, p. 402)

Students can structure their evidence and findings into a coherent whole that communicates the entire statistical enquiry to a non-specialist. Students have answered their investigative questions with appropriate evidence.

CODAP (Common Online Data Analysis Platform)

Getting started with CODAP

Read CODAP for use with students for a guide on how to use CODAP and start analysing and developing stories about data. It provides guided video tutorials on how to get started using CODAP.

Other tools exist to analyse data that are well established in Aotearoa New Zealand such as iNZight and NZGrapher. This resource does not advocate one software tool over another, but in this instance, it will focus on CODAP.

CODAP first play is a random sample from the *2023 CensusAtSchool* questionnaire from all over Aotearoa New Zealand. See if you can populate the 'pre-made' axes and start developing some stories from the data.

1. Where is the data from? Where would you go to find out more about the variables?
2. Make graphs to answer the investigative questions.
    - Click on Text icon to get a text box, write about what you notice.
    - Change the colour of the dots (use the paintbrush menu).
    - Add a legend to the *hair type* (choose a categorical variable to use, drag and drop into the middle of the graph).
    - Add counts (use the ruler menu).
    - Convert to a bar graph (use the configuration menu).
    - Make a graph bigger or smaller.
    - Change the scale on one of the axes.

- Rearrange categorical variables, for example hair types.
- Add a movable line to the *right foot lengths* (ruler menu) and add counts and percentages. Move the line to find the median, then plot the median to check (ruler menu)
- Add another movable line to the *right foot lengths* (ruler menu). Move the two lines to find the middle 50% of values, and add a box plot to check (ruler menu).

Using CODAP for statistics and probability — a very quick start guide. This is a resource from Dr Pip Arnold from the Auckland Maths Association from 2019. It provides ideas on how to get the best of your use of CODAP.

Existing teacher support material

Best pedagogical practice for statistical investigation

Investigative situation	Materials available
Comparison	Sample-to-population inference — Statistics Teachers' Day 2021
Relationship	Prediction — Statistics Teachers' Day 2021
Time series	Prediction — Statistics Teachers' Day 2021
Probability experiments	Probability modelling — Statistics Teachers' Day 2021

General materials and links:

- Data Detective Poster (UPDATED!) — CensusAtSchool New Zealand
- CensusAtSchool
- Take part in CensusAtSchool New Zealand (current census teacher guides)
- Doing the census
- CODAP
- Figure.nz
- Gapminder
- Our World in Data
- Is there a problem with the problem in the PPDAC cycle?
- Awash in Data
- codap.xyz

Assessment of learning

This resource is designed to share ideas on how teachers can innovatively teach statistics and develop student's competencies around analysis and building data narratives.

Teachers can check for student readiness for assessment by utilising the Mathematics and Statistics Subject Learning Outcomes. These can be found here.

At the end of an appropriate episode of teaching and learning assessment could occur against Achievement Standard 91944 — Explore data using a statistical enquiry cycle.

The unpacking of the Standard states the following:

> **The intent of the Standard**
>
> The purpose of this Achievement Standard is to enable ākonga to show their capabilities when exploring data, following an established statistical enquiry process.
>
> Ākonga will use a statistical enquiry process to source data and carry out an investigation. For Achievement, there is greater emphasis on how data is sourced or collected, presented, and analysed. For higher levels of achievement, ākonga will provide evidence across the chosen enquiry process.
>
> *It is intended that ākonga will have the opportunity to explore the context before beginning their independent work. This could include brainstorming in groups or with the whole class, with teacher support, to gain a greater understanding of potential purposes behind an investigation.*
>
> The intent of the Achievement Standard is to complete an investigation using a statistical enquiry process. Some end points are complex or difficult to define. Ākonga should present evidence for a completed enquiry process for higher levels of achievement.

Note: The third paragraph encourages an exploration of the context and data before embarking on completion of the assessment. It is hoped that this resource can help in guiding teachers and students to make the best advantage of this exploration stage.

Comparison investigation questions: secondary data

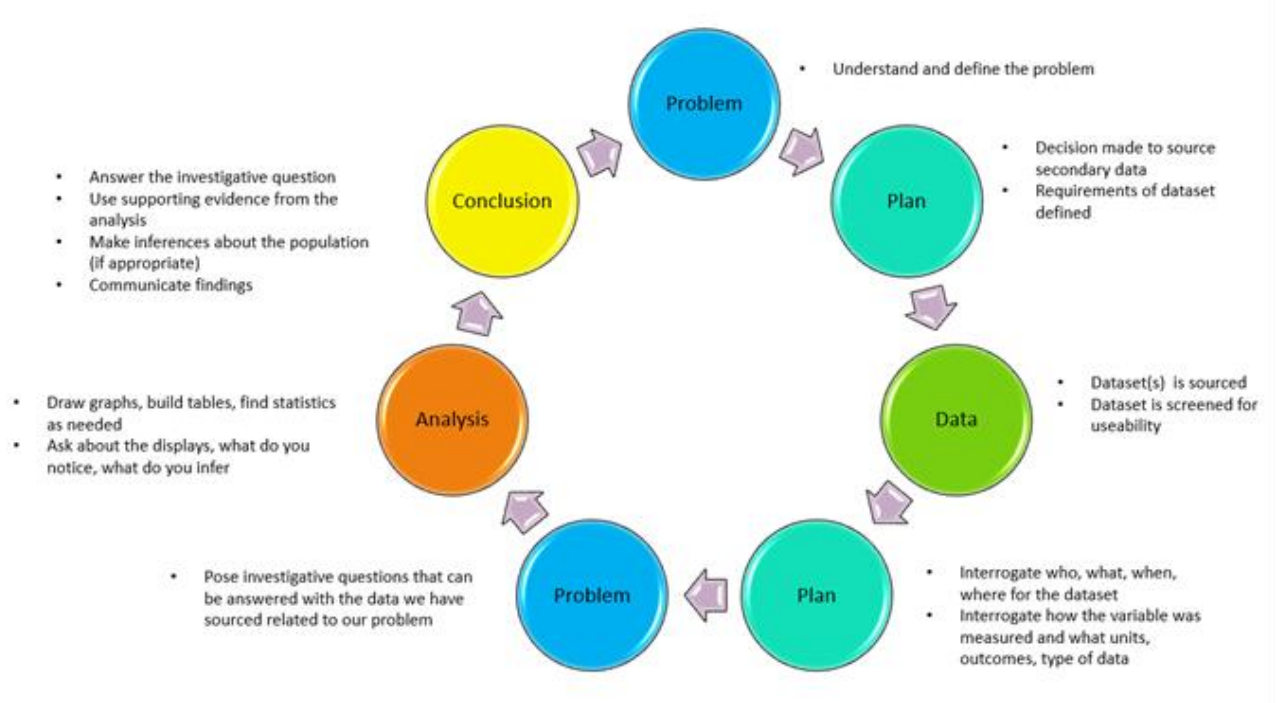
Examples of questions for a comparative question could be constructed as follows:

1. "When we did the *CensusAtSchool* questionnaire I was really interested to see the survey question about standing jumps, with and without a target. I got the sense that students in our class that had the target jumped further than those who didn't. I want to explore this for high school students in Aotearoa New Zealand.

   I wonder if Aotearoa New Zealand Year 9-13 students who had a target to aim at jumped further than Aotearoa New Zealand Year 9-13 students who didn't have a target to aim at."

2. "I notice that Year 9 students have big bags that they are carrying around and I think the amount of stuff students bring in Year 10 is less. How does this look when I explore this for Aotearoa New Zealand Year 9 and 10 students?

   I wonder if Year 9 students in Aotearoa New Zealand tend to have heavier school bags than Year 10 students."



The above image represents shuttling backwards and forwards through the statistical enquiry cycle. It starts with an overall problem, sourcing data, and then fine-tuning the investigative questions (Arnold, 2022, p. 161).

Problem

Students identify an area of interest to explore. For example, there are several opinion questions in the *CensusAtSchool* database that can be used. Not all years have opinion questions that are numerical. There are also many different variables available over the yearst pays to check if the previous censuses have things of interest.

Plan/Data/Plan

The decision is made to use secondary data, the topic has been selected, and data is downloaded from *CensusAtSchool*.

Datasets:

- Year 9-13 sample, 250 students.
- Download your own from CensusAtSchool.
- Data source — 2023 CensusAtSchool questionnaire. The questionnaire and variable list are on the right hand side.

Secondary data (Arnold, 2022, p. 79, p. 115, pp. 158-167)

The following interrogative questions provide a good starting point to understand the data, what was collected, how it was collected, and who it was collected from (p. 79).

Overall, for the dataset:

- Was the data collected using an observational study or an experiment (from Year 9)? (1. Method)

- Who was the data collected from? (2. Who)
- Who collected the data? (1. Method)
- When was the data collected? (1. Method)
- Where was the data collected? (1. Method)
- What was the purpose for collecting the data? (Initial investigator's problem/purpose)

Specific to the variable (3. What and how):

- State the variable
- What was the data-collection or survey question asked to collect the data?
- How was the variable measured?
- What are the units, if any, for the variable?
- What are the possible outcomes for the variable?
- What type of data is it? Categorical or numerical?

When we use data that has been collected by someone else, we still need to define what has been measured. Understanding what the variable is and how it was measured is an important part of identifying the variables that are available to use. If the data is provided, we use this to decide what investigative questions can be posed about the dataset. If we are sourcing the secondary data, we will know if the data will answer our investigative questions (Arnold, 2022, p. 115).

Secondary data ideas from Statistics Teachers' Day session

Interrogating secondary data is a blank slide deck that can be used with classes. Once opened, the link will prompt you to make a copy.

- **Alcohol and smoking**
  Background info Alcohol and smoking — CensusAtSchool NZ
- **Climate Change**
  Background info Important issues data — CensusAtSchool NZ
- **Healthy Lifestyle**
  Background info Important issues data — CensusAtSchool NZ
- **Natural resources**
  Background info Important issues data — CensusAtSchool NZ
- **Rubbish/Pollution**
  Background info Important issues data — CensusAtSchool NZ
- **Bullying**
  Background info Bullying data — CensusAtSchool NZ.

Problem

Comparative statistical investigations use two variables — one categorical (groups) and one numerical.

"I wonder if the time taken to get to school by car for Year 9-13 students in Aotearoa New Zealand tends to be faster than the time taken to get to school by bus for year 9-13 students in Aotearoa New Zealand?" (Uses continuous numerical data and comparison across populations.)

CensusAtSchool as a data source

The *CensusAtSchool* databases are big enough to be representative of all students of that year level in Aotearoa New Zealand. Therefore, when posing investigative questions from the *CensusAtSchool* databases

the population of interest can be students from Aotearoa New Zealand. This can be all students or specific year groups. This means the 'who' to measure is a sample from all Aotearoa New Zealand students or specific year groups of Aotearoa New Zealand students.

Identifying the specific year of the *CensusAtSchool* is not a requirement when posing the investigative question or identifying the population of interest. It is also not necessary to identify that the data is from the *CensusAtSchool* database. This is true for both in-class learning and assessment purposes. (Arnold, 2022, p. 36)

Analysis

- Data visualisations: dot plots (Arnold, 2022, pp. 229-237), box plots (Arnold, 2022, pp. 258-262), histograms (Arnold, 2022, pp. 254-258).
- Bar graphs (categorical and discrete numerical, Arnold, 2022, pp. 237-243).
- Describing distributions: (Arnold, 2022, pp. 296-315).
- Reasoning: (Arnold, 2022, pp.344-355).

Note: Reminder image — importing pictures in CODAP. Use the draw tool to visually check the Year 11 MTC (Making the call) guide.

Conclusion

- Answer the investigative question.
- Communicate findings.
- Reflect on findings.
- New ideas, new wonderings.
- Using text boxes (Arnold, 2022, p. 400).
- Telling stories with story builder (Arnold, 2022, pp. 400-402).
- Saving and sharing documents in CODAP.

Relationships

Relationship statistical investigations use two paired variables. Either two numerical variables, or two categorical variables.

“I wonder if there is a relationship between the number of words students in our class can write with their left hand in one minute and the number of words students in our class can write with their right hand in one minute?” (Paired discrete numerical data about a group.)

- Data visualisations: scatter plots (Arnold, 2022, pp. 275-280), categorical two-way tables (Arnold, 2022, pp. 266-271).
- Describing distributions (Arnold, 2022, pp.321-330), paired categorical (Arnold, 2022, pp.330-333).
- Reasoning (making predictions) (Arnold, 2022, pp. 355-356).

“I recently went out for a day to Rainbow’s End in Auckland with my family. I’m interested in finding out more about rides in the home of the rollercoaster — the USA! I wonder if there is a relationship between the age of a rollercoaster in the USA and the maximum speed it can travel?”

What variables (assuming ***easy*** access to enable curation of appropriate data) would be good for exploring and answering this question?

Sample of US rollercoasters across years (CODAP).

Time series

Time series statistical investigations use two numerical variables, one of which is time.

An example of a question that is suitable for a time series investigation is as follows:

“What are the patterns in the monthly number of two-litre ice-cream packs sold in Aotearoa New Zealand from 2014-2020?”

- Data visualisations: Time series graphs (Arnold, 2022, pp. 280-283).
- Describing distributions (Arnold, 2022, pp. 333-336).
- Reasoning (making predictions): (Arnold, 2022, pp 356-357).

“I’m interested in finding out more about the cost-of-living crisis and especially the cost of food. In my family we love dairy products. And in my family, I love cheese more than anyone. I wonder what has been happening with the price of cheddar cheese (per 1kg) in Aotearoa New Zealand supermarkets over the last decade?”

- Retail price of mild cheddar cheese in New Zealand (figure.nz)
- Would some worldwide factors possibly have an influence on the patterns seen in cheese prices in Aotearoa New Zealand over the last decade? Food prices (Our World in Data)

Other potential sources of good Aotearoa New Zealand-based time series data are probably best supplied by the teacher. Infoshare (Stats NZ) has interesting Aotearoa New Zealand-based data, relevant to students. This data is up to date, sometimes only days old.

Infoshare data for time series is a video that demonstrates how to download very good time classic time series data sets. Although aimed at downloading data sets for NCEA Level 3, the same process is used for downloading data sets suitable for Level 1.

Probability experiments

PPDAC stage	As applied in probability at Year 10 (end Phase 4)
Problem	Students will: • recognise and pose investigative questions for two- and three-stage chance situations • anticipate what they think will happen before they undertake probability investigations.
Plan	Students will: • state assumptions for the two- and three-stage chance situation • systematically list the sample space (all possible outcomes) of two- and three-stage chance situations, using, for example, tables and tree diagrams • construct the model where appropriate, for two- and three-stage chance situations • design probability experiments for two- and three-stage chance situations , including: – what random generating devices will be used – what number of trials will be carried out – how the results will be recorded • design probability experiments that use real data to create probability distributions for numerical variables such as bag weights or time to run 50m.
Data	Students will: • run simulations for probability experiments • systematically record data from the probability experiment using lists, tables, and digital tools as appropriate.
Analysis	Students will: • create data visualisations for the distribution of observed outcomes from probability experiments and the distribution of possible outcomes for theoretical probability situations • describe probability distributions including using simple, joint, and conditional probabilities.
Conclusion	Students will: • identify similarities and differences between their findings and those of other groups when undertaking probability experiments • identify similarities and differences between their findings from probability experiments and associated theoretical probabilities as appropriate (if the theoretical model exists) • answer investigative questions by choosing statements from their findings that best describe the chance situation • propose new theoretical models based on probability estimates from probability experiments • reflect on anticipated outcomes.

Student explanation example

Read this example of a CODAP data exploration.

- What is good about this data exploration?
- Would it meet the expectation of the Standard? If not, why not?

References and links

Arnold, P., & Pfannkuch, M. (2020) On being data detectives: Developing novice statisticians using the statistical enquiry cycle. SET 2020(1), 34-41 https://doi.org/10.18296/set.0159

Arnold, P., & Franklin, C. (2021). What Makes a Good Statistical Question? *Journal of Statistics and Data Science Education*, 1-11. https://www.tandfonline.com/doi/full/10.1080/26939169.2021.1877582

Arnold, P. (2022). *Statistical investigations | Te Tūhuratanga Tauanga.* NZCER Press.

Arnold, P. (Aug 27, 2021). *Is there a problem with the problem in the PPDAC cycle?* https://www.aucklandmaths.org.nz/is-there-a-problem-with-the-problem-in-the-ppdac-cycle/

Arnold, P., & Pfannkuch, M. (2022). Engaging novice statisticians in statistical communications. *Mathematics Education Research Journal*, 1-27. https://link.springer.com/article/10.1007/s13394-022-00442-w

Arnold, P. (2013). *Statistical Investigative Questions – An Enquiry into Posing and Answering Investigative Questions from Existing Data*, (Doctoral thesis), Retrieved from https://researchspace.auckland.ac.nz/handle/2292/21305

Pfannkuch, M. (2005). Characterizing year 11 students' evaluation of a statistical process. *Statistics Education Research Journal*, *4*(2), 5-25.

Pfannkuch, M., Arnold, P., & Wild, C. J. (2015). What I see is not quite the way it really is: Students' emergent reasoning about sampling variability. *Educational Studies in Mathematics*, *88*(3), 343-360.

Pfannkuch, M., Arnold, P., & Wild, C. J. (2011). *Statistics: It's reasoning not calculating.* (Summary research report on Building students' inferential reasoning: Levels 5 and 6) Retrieved from http://www.tlri.org.nz/tlri-research/research-completed/school-sector/building-students-inferential-reasoning-statistics

Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. (2011). Towards more accessible conceptions of statistical inference (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(2), 247–295. doi: 10.1111/j.1467-985X.2010.00678.x https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1467-985X.2010.00678.x

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–265. doi: 10.1111/j.1751-5823.1999.tb00442.x https://www.stat.auckland.ac.nz/~iase/publications/isr/99.Wild.Pfannkuch.pdf

Te Poutāhū
Curriculum Centre

We **shape** an **education** system that delivers **equitable** and **excellent outcomes**

He mea **tārai** e mātou te **mātauranga** kia **rangatira** ai, kia **mana taurite** ai ōna **huanga**

temahau.govt.nz

Te Kāwanatanga o Aotearoa
New Zealand Government

Te Mahau

Te Tāhuhu o te Mātauranga
Ministry of Education