 The scientific research proposal

This document is a teaching resource to support the Stage 6 Science Extension course[[1]](#footnote-1).

[Preface 2](#_Toc12001894)

[Acknowledgements 2](#_Toc12001895)

[Contact 2](#_Toc12001896)

[Executive summary 3](#_Toc12001897)

[Developing the Question and Hypothesis 4](#_Toc12001898)

[Inquiry and research 4](#_Toc12001899)

[Understanding the reasoning and logic behind scientific research questions 4](#_Toc12001900)

[Limits to scientific inquiry 5](#_Toc12001901)

[Types of research questions 5](#_Toc12001902)

[From ideas to research questions 5](#_Toc12001903)

[Scientific Research Proposal 10](#_Toc12001904)

[Developing the research strategy 10](#_Toc12001905)

[Writing the research plan 14](#_Toc12001906)

[Planning for data collection and analysis 15](#_Toc12001907)

[Types of quantitative data 15](#_Toc12001908)

[Collecting primary data 15](#_Toc12001909)

[Collecting secondary data 16](#_Toc12001910)

[Teaching resources and for data collection and analysis 17](#_Toc12001911)

[Processing Data for Analysis 17](#_Toc12001912)

[Detecting and removing outliers 18](#_Toc12001913)

[Teaching resource 18](#_Toc12001914)

[References 18](#_Toc12001915)

Preface

Students in the Science Extension course are expected to produce a research proposal (module 2). All scientific research projects require a proposal. The research proposal is used to plan the investigation, obtain relevant information from the literature, as well as to seek research funding. Researchers expend a considerable amount of time and effort to construct their research proposal. The proposal is usually developed over time and usually undergoes multiple rounds of drafting and editing. This document provides information on the different aspects of a research proposal, which teachers may use to illustrate the process to their students.

Acknowledgements

The Learning and Teaching Directorate at the NSW Department of Education developed this resource for science teachers. The department acknowledges the contributions of the following teachers in developing this document:

* Carl Masens, Girraween High School
* Andrew Burns, Arthur Philip High School
* Sham Nair, Learning and Teaching Directorate

Contact

For more information, contact

Sham Nair  
Science Advisor, 7-12  
NSW Department of Education  
[sham.nair@det.nsw.edu.au](mailto:sham.nair@det.nsw.edu.au)

Executive summary

Developing the research proposal is an essential aspect of scientific practice. All researchers need to construct a proposal to undertake an inquiry. This proposal then becomes the basis on which the research is funded. Research proposals are complex documents, as they consist of many facets, such as a literature review, inquiry questions, a research plan and budget. Scientists expend a lot of time and effort to prepare research proposals. In the Science Extension course, students will prepare a research proposal that contains some of the elements found in formal research proposals.

The most important aspect of the research proposal is the research question. Although there are many types of research questions, most of them fall into one of three categories: relational (establish relationships between variables), descriptive (collect information about a system) or causal (identify mechanisms or develop explanations for the behaviours of variables). Students may follow a four-step process to develop their research question:

1. Develop the initial overarching question
2. Develop inquiry questions from the initial question
3. Ground questions in theory
4. Convert the inquiry question into a research question

After the research question is develop, students should put together the research proposal. Every research proposal should:

* Define the problem (based on the four-step approach described above).
* Outline the methodology for collecting data.
* Explain how the collected data will be analysed.

The experimental methodology must outline the strategies for collecting data using designs for fair tests. While students must clearly identify the dependent and independent variables of the investigation, they must also indicate how the other variable will be controlled. Furthermore, all data analytical tools and statistical methods must be described in the proposal.

Finally, the research proposal may be written in the form of a scientific report, with the following headings:

* Abstract
* Introduction
* Materials and methods
* Results
* Conclusion
* References

Developing the Question and Hypothesis

Inquiry and research

The terms, inquiry and research, are common in science. There is a distinction between the two. Inquiry is questioning that leads to answers. The process of generating answers is an investigation. There are different forms of inquiry, such as judicial inquiries, public inquiries, market research and surveys. Research is a type of inquiry that uses formal processes that are specific to the area of research. Scientific research uses a set of rules and procedures accepted by the scientific community for conducting an inquiry. Such rules include adhering to accepted investigative processes (for example – working scientifically), repetitive observations and measurements, controlled-experiments and the peer-review process. Therefore, research refers to systematic investigations (inquiries).

Understanding the reasoning and logic behind scientific research questions

Peter Medawar described science as the art of the soluble(Medawar, 1967). In saying this, Medawar highlights science’s propensity for searching for solutions to problems, as well as its quest for explanations. All scientific research is centred on inquiry questions. These questions must be well-crafted because the success of any research endeavour depends upon them. Scientific research generally falls into one of two approaches:

* Induction: In inductive research, scientists make observations before developing explanations about some phenomena (“observation before theory”). These explanations, together with explanations from other experiments, will be used to construct a scientific theory. For example, scientists may sequence the genome of an organism before discovering the genetic elements encoded in that genome. These discoveries will then be used to construct theories about the genetic makeup of the organism or species.
* Hypothetico-Deduction: Deductive experiments are based on existing theories (“theory before observations”). Those theories are used to develop predictions (hypotheses). In deductive research, hypotheses are tested in controlled experiments. The results of the hypothesis testing are used to infer the validity of the underlying theory.

Newer fields of research rely on research questions that seek to develop the knowledge base (inductive approach). In well-established fields of research, there will be a large knowledge base on which research questions (hypotheses) are based (deductive approach). Thus, as research areas mature, there is a tendency towards using inquiry questions that are based on hypothetico-deductive approaches.

Limits to scientific inquiry

In scientific research, not all inquiry questions are of equal value. There are limitations to scientific inquiry. Some reasons for this include:

* Some phenomena cannot be investigated in controlled experiments (for example – supernatural and paranormal phenomena)
* Some phenomena are too complex to be investigated in a single research endeavour.
* The cost of conducting scientific investigations can be prohibitive.
* There may be resource and technological limitations.

Therefore, when developing research questions, scientists must consider if:

* the topic area is relevant to contemporary science;
* the research question is based on a substantial knowledge base;
* the research question is of interest not only to the researcher conducting the investigation, but also to the broader scientific and lay communities.

Types of research questions

Generally, there are three categories of inquiry questions:

* Relational questions: these questions explore the relationship between variables. For example,
  + Does the consumption of vitamin C reduce the severity of cold symptoms?
  + How is the efficiency of solar cells affected by temperature?
* Descriptive questions: these types of question are designed to collect information about a system. For example,
  + What is the territorial range of the blue whale?
  + What happens when high-velocity protons collide?
* Causal questions: A causal question asks how one variable affects another variable. Here, the goal is to develop a cause-and-effect relationship between the variables. For example,
  + What is the effect of salt consumption on blood pressure?
  + How does increasing the inhibitor concentration affect the activity of an enzyme-catalysed reaction?

From ideas to research questions

There are major differences in the way scientists and students approach the process of developing research questions. High school students generally do not possess the expert disciplinary knowledge that professional scientists do. Despite that, students can develop highly effective research questions.

Students’ interest in scientific issues may be a springboard for developing research questions, including:

* Interest in scientific topics or scientific stories that are described in the media (news, documentaries, science fiction and so on).
* Interest that arises from discussions with friends, family, teachers and others (for example – students may know of someone with a particular disease, which may prompt them to inquire about that condition; a family member who may be a scientist).
* Interest in social justice issues that may be addressed through scientific inquiry.

All of these provide ample motivation for scientific inquiry, and students may be able to transform their interests into research questions. Students who may not be intrinsically driven by personal interests can also be guided to develop research questions after consulting their teachers or mentors.

No matter what the starting point may be, students must convert their general ideas into specific research questions. The following steps may be used to guide students to develop their research questions:

1. Develop the initial overarching question
2. Develop inquiry questions from the initial question
3. Ground questions in theory
4. Convert the inquiry question into a research question

Step 1 – develop the initial overarching question

The initial overarching question is a highly generalised question in a specific topic area. At this stage of the research process, students should be encouraged to think broadly (blue sky thinking). For example, initial inquiry questions could be ‘can we live forever?’; ‘will climate change destroy life on Earth?’; ‘do parallel universes exist?’. The goal here is for students to tap into their innate curiosity or explore research ideas suggested to them by their teachers. A good initial question invites further exploration and discovery. The overarching question need not be framed as a question but may be stated as the overall goal for inquiry.

Step 2 – develop inquiry questions from the initial overarching question

At this stage, students will develop several specific, focused inquiry questions from the initial overarching question. These inquiry questions may eventually form the basis of the research project. This requires students to unpack their overarching question into sub-topics. For example, when exploring the effect of climate change on living things, students should select some aspect(s) of climate change (for example – carbon dioxide levels, temperature and weather) and the biosphere (for example – changing biodiversity, adaptation and ecosystems) for further exploration. Students can begin to construct inquiry questions by asking questions about their overarching question: “What do I want to know about climate change and its effect on living things?” “How does climate change affect biodiversity?” The goal here is to develop questions that can be answered through data collection and analysis.

Somethings to consider when developing inquiry questions:

* Avoid questions that have a simple yes/no answer.
* Avoid questions that have a simple factual answer.

It is important to note that these types of questions may be scientifically valid. However, as a research inquiry, the answers to these types of questions do not provide much context (For example – why the question is important). Rather, students should ask questions that require them to explain, demonstrate/illustrate, argue or prove a conclusion.

Step 3 – Ground questions in scientific theory

The next step is to see how well the inquiry questions may be linked to existing scientific theories. To do this, students should explore the literature (for example – Google Scholar search or searching academic databases). This will indicate if other researchers have addressed the inquiry questions. If so, students should consider not pursuing those inquiry questions, or modify the inquiry to explore related aspects of the research problem. If students discover that there are only a few published articles on their inquiry topic, then it may indicate that the proposed inquiry may be too specific (in which case the questions need to be broadened) or not be based on established theories.

Science Extension students should develop their inquiry around topics that are grounded in well-established scientific theories. Inquires that are not based on established scientific principles may be difficult to investigate within the timeframe of the Science Extension course. Some advantages of theory-based inquiries include the following:

* The inquiries are easier to justify (for example – easily establish the relative importance of the research area).
* It is easier to develop questions that are not too general or too specific.
* The purpose of the inquiry can be easily established (for example – the research will provide data to validate a theory, or that the research will help resolve a contradiction and so on.).
* The inquiry question can be used to frame research questions and hypotheses.
* After analysing the data obtained in the research project, it may be easier to find cause-and-effect relationships between variables.

The following points suggest useful approaches to finding published articles.

* In the first instance, students should explore the internet with keywords that are relevant to the research topic. They should use a variety of databases (for example – Google Scholar, ISI Proceedings, JSTOR Search, Medline, Scopus, Web of Science), as different databases may identify different publications for the same search query. Many of these databases may be freely accessed through the State Library of NSW.
* As they progress through their literature research, students should use increasingly complex search strings. For example, rather than simply searching for information on ‘hypertension, searches such as ‘hypertension and new treatments’ may provide specific information. Boolean operators such as ‘and’, ‘or’, ‘not’ and quotation marks (“ ”) will increase the specificity of search results. When using the Google search engine, use the ‘[advanced search](https://www.google.com/advanced_search)’ feature, which allows users to construct complex queries.
* Some databases (for example – PubMed) use controlled vocabulary to identify research articles. For example, the medical research literature is archived as MeSH (Medical Subject Headings) terms (Baumann, 2016). This allows researchers to identify articles that contain a list of key terms that are of interest to them:
  + Search 1: Cancer
  + Search 2: Clinical trials
  + Search 3: Gene therapy
  + Search 4: search 1 + search 2 + search 3

The final search will return a list of all articles that contain information on gene therapy trials for the treatment of cancer.

* Students should check that the articles being reviewed are recent (generally, published in the past ten years\*, although older articles may still be important). To do this, try the following:
  + Open Google Scholar.
  + Type in the title of an article.
  + When the results appear, refer to the entry that is the same as the query (for example – the same article).
  + Look at the ‘cited by’ entry – this indicates if there are more recent articles on the same topic (note: a large number indicates that the search article is an important one as it has been cited by many other authors!).

\* The chronological age of the published articles is not as important as the currency of the ideas and concepts. In science, important ideas are reviewed and cited, even if those ideas are relatively ‘old’.

Step 4 – Converting the inquiry question into a research question

Most inquiry questions are too broad to initiate investigations. A research question is a focussed and concise statement of investigation. The research question is developed from the inquiry question. After narrowing the inquiry questions to a small number of possible questions, students need to select a specific question for subsequent research. This choice may be guided by several factors, such as:

* Can the inquiry question be evaluated through an investigation?
* Can a fair test be developed to answer the question?
* Are resources and other materials readily available for investigation?
* Can the data be analysed using the resources available?

Identifying the variables in an inquiry is an important step for converting the inquiry question into a research question. Apart from the dependent and independent variables, other variables need to be controlled so that the findings of the study are meaningful.

Consider the following example:

* Inquiry question: does the use of pesticides in agriculture affect bee populations?
* Research question: what is the effect of neonicotinoid pesticides bee colony collapse disorder (CCD)?

Here, the research question is a focussed statement derived from the inquiry question. Rather than exploring ‘pesticides’, the researcher proposes to specifically investigate ‘neonicotinoid pesticides’. Similarly, rather than investigating bee populations in general, the researcher will explore the phenomenon of CCD in beehives. Thus, the research question is a specific restatement of the inquiry question. The independent and dependent variables are evident in the research question.

If there is a substantial knowledge base on the topic area, then a hypothesis may be developed. A well-constructed hypothesis possesses the following features:

* The hypothesis is testable (conduct a fair test).
* The hypothesis is predictive (if certain conditions are met, then certain outcomes may be observed).
* The hypothesis is based on theory (the predicted behaviours occur because they are based on scientific theory).

Note: The scientific theory on which the hypothesis is based need not be presented in the hypothesis statement but may be provided in the background to the research.

In scientific research, hypothesis testing often involves statistical analyses. For statistical tests, a null hypothesis is derived from the scientific hypothesis.

Returning to the example described above,

* Inquiry question: does the use of pesticides in agriculture affect bee populations?
* Research question: what is the effect of neonicotinoid pesticides bee colony collapse disorder (CCD)?
* Hypothesis: Increasing the concentration of neonicotinoid pesticide concentrations will result in a great CCD effect in beehives.
* Independent variable: neonicotinoid pesticide concentration.
* Dependent variable: CCD effect.

If the literature does not contain a good knowledge base on the research topic, then it may be difficult to formulate a hypothesis.

Scientific Research Proposal

A research plan or research proposal is an organisational structure for the ways the inquiry will be carried out. The research plan provides the reader with an overview of the investigation and a strategy for addressing the research question and hypothesis. Research plans are normally written in the form of a report, although in a highly abbreviated form. The research plan should address the following:

* The problem.
* The method.
* The result (how the collected data will be analysed).

Developing the research strategy

The research strategy tells the reader about the investigator’s approach for answering the research question. Students may initiate the process by asking two questions about their inquiry:

* What data is needed to answer the inquiry question?
* How can the required data be collected?

By answering these questions, students can identify potential experimental designs and data analysis strategies.

The dependent variable

Consider a research project in which a student wishes to explore the effects of climate change on photosynthesis. The student’s research question examines if higher atmospheric carbon dioxide levels affect photosynthesis rates. Thus, in this investigation, the photosynthesis rate represents the dependent variable. During researching planning, the student should examine different perspectives for collecting relevant data.

Table 1. Exploring different data collection strategies during research planning

| Criteria | Data needed | Data collection strategy |
| --- | --- | --- |
| Photosynthesis rate | The volume of oxygen produced | Oxygen probe |
| Photosynthesis rate | Amount of glucose produced | Starch test |
| Photosynthesis rate | Change in dry mass | Measurement of the dry mass of leaves |
| Photosynthesis rate | Changes in the rate of the light-dependent reaction | Perform a Hill Reaction with DCIP (2,6-dichlorophenolindophenol) |

As shown in Table 1, the student has considered four strategies to collect data for the dependent variable. After this, the student must decide on the most suitable strategy for the project. The student may decide to select more than one strategy to collect data for the dependent variable. In the example above, the student may decide to measure photosynthesis rates by determining oxygen production and the Hill reaction. Using more than one type of measurement will increase the confidence in the outcomes of the experiment.

The decision to select one or more of the measures of the dependent variable depends on many factors, including:

* The difficulty of making the measurements.
* The availability and suitability of equipment, reagents and protocols.
* Selecting published methodologies (which have been verified by other researchers).
* The suitability of the measured variable for answering the research question – for example, variables that yield qualitative data will not be suitable for investigations that require quantitative data.

The independent variable

After deciding on the data needed to answer the research question, the student must consider the most suitable experimental system for generating the data. For the student conducting the investigation described above, the following considerations may be factored into the experimental design (Table 2).

Table 2: Exploring the materials and conditions for the independent variable

| Component | Considerations | Options |
| --- | --- | --- |
| Choice of plant | * Fast growing – experiment can be completed with the duration of the course. * Easy to handle in large numbers – need to create multiple groups for treatment and control. * Others have worked with the plant – established protocols will be available. | * Beans – meet all these criteria * Algae – meet most criteria, but growth is difficult to control * Arabidopsis thaliana (mustard) – a well-established model system |
| Carbon dioxide source | * Cheap * Controlled | * Dry ice – readily available, but difficult to control the amount of carbon dioxide released. * Carbon dioxide gas bottles – these are fitted with regulators, which control the release of the gas. Maybe expensive. |
| Setup | Need a sheltered space so that plants obtain all the components needed for photosynthesis, but where the carbon dioxide levels can be controlled. | Greenhouse – may be purchased from hardware stores or nurseries. |

Controlled variables

The student then identifies all other variables that may influence the outcome of the investigation and then develops procedures to keep those variables constant during the experiment.

Experimental design

Teacher note: this section provides an overview of experimental designs. A detailed discussion of experimental design is beyond the scope of this document and is a difficult concept for most high school students. However, it is important that students understand the importance of experimental design in scientific research. All experimental designs are developed to ensure that the observed effect of manipulating a variable is real, unbiased and valid.

A good experimental design ensures that the observed effect (dependent variable) occurs solely due to the manipulations of the independent variable. The three aspects of good design are:

* Manipulation – manipulating the independent variable to observe its effect on the dependent variable.
* Control – all other variables are controlled to the best extent possible; also, when comparing the effect of manipulation, the control group does not receive the same manipulation as the experimental group.
* Randomisation – where grouping is involved (for example, experimental and control groups), individuals are randomly assigned to those groups. This is important for removing bias and enhancing the validity of the investigation.

Here is a brief overview of commonly used experimental designs in scientific research:

| Test | Design | Example |
| --- | --- | --- |
| Post-test | * Experimental group → treatment → post-test measurement * Control group → post-test measurement (no treatment applied to the control group) | Measuring the severity of cold in a group of students taking regular doses of vitamin C, compared to a group that does not consume vitamin C. |
| Pre-post test | * Experimental group → pre-test measurement → treatment → post-test measurement * Control group → pre-test measurement → post-test measurement (no treatment applied to the control group) | Blood pressure measurement before and after a treatment regime. |
| Factorial design | Multiple independent variables are tested (for example 2 x 2 or a 2 x 3 design). Usually, controls are built into such designs.  Multiple independent variables are tested (for example 2 x 2 or a 2 x 3 design). Usually, controls are built into such designs. | Testing the effect of an inhibitor (at different concentrations) and reaction temperature on the rate of a catalytic reaction. This is a 3 x 3 factorial design. An inhibitor concentration of 0 mM and a temperature of 10oC represent control conditions. |
| Randomised block design | This is a variation of the factorial design and explores complex interactions between multiple variables. This is a variation of the factorial design and explores complex interactions between multiple variables. | Testing the effect of 3 anti-hypertensive drugs in patients with simple or complex disease profiles. |

The experimental design is also important for the type of statistical analysis that will be applied to the data. This is discussed further below.

Writing the research plan

After developing the research strategy, students should construct the research plan. All of the steps described above will contribute to the research plan. One possible structure to the research plan is as follows (Coroller, 2017):

* Abstract**: 100-200 words, describing the key features of the research project.**
* Introduction**: a summary of the following points:**
  + Statement of the problem addressed.
  + Purpose of the project – what the project will accomplish.
  + Key ideas from the literature – include the major relevant publications, as well as any contradictions/controversies and the progress in the field. Note that this is not the comprehensive literature review that is prepared for the research report.
  + Research question and hypothesis
  + The significance of the research – how the research community and the general community will benefit from the research.

**Each of the abovementioned parts of the introduction may be presented as individual paragraphs. Be concise. Paragraphs should be arranged logically and display a progression of ideas.**

* Materials and methods**:** Here, students should provide the methodology and approach that they intend to use to answer the research question. Technical details should be avoided, although sufficient information should be provided so that the reader understands the procedure. The experimental design (including variable and controls), experimental techniques (for example – the use of gas sensors to measure photosynthesis in elevated carbon dioxide conditions), data analysis methods (for example – the use of t-tests) should be indicated here. If the methodology has been published in academic journals, then students can refer to those articles (reference them), except indicating any modifications they intend to make. Other methodological information that can be described include:
  + Statistical methods
  + Software (including version)
  + Hardware (for example – microscope, LASER)
  + Biological organisms
  + Chemicals

Students can also highlight any anticipated problems with the proposed methodology, as well as their possible remedies. This indicates a high level of critical thought on the part of the students. The design and methods should also reflect the specific aims of your study. The methodology should anticipate possible pitfalls, suggest potential controls, and ensure that the collected data is clean and unbiased.

* Results **–** Since this is a research plan, students will not have any results to describe. However, they may include a short description of how the data may be interpreted. For example, students may suggest that ‘increased oxygen production by plants exposed to higher carbon dioxide levels will indicate higher photosynthetic rates’. It is important that student not speculate or make predictions about their findings.
* Conclusion **– This should be a restatement of the intent of the project (purpose), as well as the significance of the work (why the project will be important for science)**
* References **– A list of references listing all articles used in developing the research plan. These references should be formatted according to accepted citation formats.**

Planning for data collection and analysis

Once an experimental design has been selected, students should consider the type of data that will be generated in the project, as well as how the data will be analysed. Data collection strategies should be designed with considerations, such as:

* The research questions.
* Data analyses tools.
* Statistical analyses.
* Accuracy, precision, reliability and validity.

Types of quantitative data

There are two types of data: qualitative and quantitative. Qualitative data refers to information that is based on the qualities of the property being measured. Quantitative data measure the size or number of some property being measured. Quantitative data is classified as categorical (data takes a limited number of possible values – for example, eye colour in a group of students) or continuous (data can have many possible values). The type of quantitative data collected in an investigation determines the type of statistical test that can be used to analyse the data.

Collecting primary data

Primary data is collected by the student through observations, experiments or surveys. Students should think about the optimal sizes and numbers of groups (for example, the numbers of experimental and control groups), as well as the number of times the experiment should be repeated for consistency of results. Time and cost issues must also be factored into the experimental design. If statistical analyses of the data are to be performed in the investigation, students also need to be aware of factors such as sample sizes and the variability of the data to be collected. For example, to compare two datasets using a 2-sample unpaired t-test, the data should be large (>25 data points in each dataset) and be normally-distributed. Generally, students can get a good idea of these requirements by examining experimental designs in published papers in similar research areas to their own.

Collecting secondary data

Secondary (secondary source) data refers to data that has been collected by others. Secondary data analysis is a cost-effective method of conducting scientific inquiry. Such data may be obtained from a variety of sources, such as scholarly articles, technical reports, review articles, reference book and online repositories (databases).

The quality of secondary data can vary considerably. Students must determine the suitability of the data before using them in data analyses. Some important factors to consider are (Mccaston, 1998):

* Purpose of the original data collection – this helps to determine if there is a significant bias in the data.
* Identify the authors and sources of the data – the quality of the data will be better if it was collected by experts who are fluent in the processes of data collection. For example, data collected through citizen science projects may contain inconsistencies that affect the accuracy of the information in the dataset.
* Collection method – datasets should contain metadata, which provides information about the procedures used in the collection and annotation of the data. These metadata are crucial for determining the suitability of the dataset for analysis.
* Intended audience – data collected for the general public may not be sufficiently rigorous for scientific analysis.
* Primary or secondary data – some databases aggregate data from other sources. This type of data is not primary data and are sometimes considered to be tertiary data. They may contain errors and inconsistencies. Primary data are produced directly by the investigators who use them for analysis. Secondary data is not produced by the user, but is extracted from databases and other data repositories. If the data is not primary information, then references to the sources should be provided.

Teaching resources and for data collection and analysis

The following resources may be used in the classroom to demonstrate various aspects of data analysis.

* [IB Physics: Uncertainties and errors](https://www.youtube.com/watch?v=6XZsfV5FCwc.). Note that in the final example the line of worst fit is outside the error bars, but otherwise, the content is good.
* [Chemistry Worksheet Accuracy vs Precision and Random vs Systematic Errors](https://www.quia.com/quiz/5112927.html)
* [Read Percentage Error](https://www.mathsisfun.com/numbers/percentage-error.html) (includes a quiz at the end of the web page)

Processing Data for Analysis

Before data are analysed, they must be processed. Data processing refers to the steps taken to ensure that the measurements are of good quality and integrity. It involves a three-step process that involves screening of the datasets, diagnosis of problem measurements, and treatment of faulty measurements in the datasets.

Data processing consists of two related procedures: data cleansing and data editing. Data cleansing is the process of detecting and editing faulty measurements in datasets, while data editing refers to replacing incorrect data. These ideas are elaborated below:

* Faulty data: faulty data arise from incorrect data recording. This results in the incompleteness of datasets. Missing values or incorrect measurements must be rectified. For example, measurements may be missing from a dataset because a data logger may have failed during an experiment. If this is suspected, then the experiment must be repeated so that complete datasets are obtained. Some missing measurements may be added to the datasets retrospectively if it is reasonable to do so: for example, retrospectively adding the missing gender of a participant in a study is reasonable if the information is available. Care must be taken to ensure that any information that is added retrospectively are not guesses or extrapolations. Faulty data can occur during manual data entry, or when transposing data from one file type to another. For example, if the raw data is collected as a CSV file and then imported into a spreadsheet (for example – Excel), the data must be carefully inspected for errors (for example – blank space or rounding off). Excel can also change certain number formats (for example, from alphanumeric to date formats).
* Validity: this refers to whether the intended measurements were collected in an investigation. For measurements to be valid, appropriate instruments must be used in experiments. Relevant units of measurements must be used (for example, when measuring the voltages of batteries, the measurements in the dataset should indicate values in volts or millivolts). In some types of experiments, measurements must occur within certain ranges or permissible values. Readings that are beyond the measurement scales of instruments are invalid data.
* Reliability: when experiments are repeated, measurements in those datasets should be similar (similar means or medians and standard deviations).
* Faulty data may be diagnosed in several ways:
  + Using a spreadsheet to sort data tables and then to browse the data for anomalies
  + Plotting graphs of the data to identify anomalies
  + Examining the descriptive statistics of the datasets (mean, median, mode, and so on)

Detecting and removing outliers

Detecting and removing outliers is a common procedure in data analysis but should only be done cautiously. Most procedures for outlier detection involve statistical analyses of the data:

* 3 x SD: a rule of thumb is which states that measurements that are less than or greater than three times the standard deviation may be outliers. However, this is only an approximation.
* The interquartile range (IQR): below [Q1 – (1.5 x IQR)], or above [Q3 + (1.5 x IQR)], where Q1 is the first quartile and Q3 is the third quartile of the dataset.
* The Chauvenet’s criterion: a method based on calculating the  (tau) statistic of a dataset.

All steps taken to process the data should be mentioned in the material and methods section of the research report. The most important consideration of data processing is that the processing should not alter the qualities of the datasets. If datasets require considerable processing, students should consider repeating their experiments to obtain new data.

Teaching resource

View [Microsoft Excel data analysis tool](https://www.youtube.com/watch?v=4lAvbp-yVs8) (Analysis ToolPak for Excel) video which describes how to use the tool to determine the mean, median, hypothesis and regression of datasets

References

Baumann, N. (2016). How to use the medical subject headings (MeSH). *International Journal of Clinical Practice*, *70*(2), 171–174. https://doi.org/10.1111/ijcp.12767

Coroller, T. (2017). Step-By-Step Guide to Writing a Research Proposal. Retrieved November 20, 2018, from https://realclearenglish.org/blog/guide-to-writing-a-research-proposal-stem-edition

Mccaston, M. K. (1998). TIPS FOR COLLECTING, REVIEWING, AND ANALYZING SECONDARY DATA. Retrieved November 21, 2018, from https://www.ands.org.au/\_\_data/assets/pdf\_file/0003/713235/Tips\_for\_Collecting\_Reviewing\_and\_Analyz.pdf

Medawar, P. B. (1967). Two conceptions of science. *The Art of the Soluble. London: Methuen*.

1. This document references the Stage 6 Science Extension syllabus © 2019 [NSW Education Standards Authority (NESA) for and on behalf of the Crown in right of the State of New South Wales.](http://syllabus.nesa.nsw.edu.au/copyright/) [↑](#footnote-ref-1)