Evaluating scientific data

# Stages 4-6

This document provides guidelines for evaluating data and conclusions from scientific investigations. It describes a set of standardised criteria for such evaluations and is applicable for Stage 4-6 science instruction. Teachers may use the principles described in this guide to familiarise their students with the definitions and usage of the evaluation criteria.

The document references the [Science Year 7-10 Syllabus](https://educationstandards.nsw.edu.au/wps/portal/nesa/k-10/learning-areas/science/science-7-10-2018/!ut/p/z1/rVPLcoIwFP0WFywzuQmB4BIflVqtVkuVbJwI0eIjIDJq-_WFTtud0I7NKpk5r9ycYIHnWGh5itcyjxMtd8U5EPaC3XsAJtCBY3cYPLUJ6baoNem1GZ59AqhLbOIx8jDqOQTc5xGxaZ_TkW9hUfIJ7RHi0CFwysGdjNvjjndHYWB98eHKcuF3_AqAqM7_ggUWoc7T_BUH6XkRJjpXOjcgzZKNCnN0VksDtDpKA7aIgAE7JTMd6zWSmZJHA45hrHSofjaIFzBEgTildBrGEQ6YbUbLiBLkrJpLxHjTQhIYQTYo6hAZ8bD5PcqKrKJ6UrPSr-Yx6jSCIgO_noHj2SlWZ-zrJNsX9Zj-8YpencOU3uhQI2_dKN-v61PxYeLN4SDcolVllS45nv93rdK97_t7x3xD29Wwa7Kgf3pvPSIRuI3GByoo2EI!/dz/d5/L2dBISEvZ0FBIS9nQSEh/?urile=wcm%3Apath%3A%2Fpw_content%2Fproject-web%2Fnesa%2Fk-10%2Flearning-areas%2Fscience%2Fscience-7-10-2018) © 2018 NSW Education Standards Authority (NESA) for and on behalf of the Crown in right of the State of New South Wales.

## Table of contents

[Evaluating scientific data 1](#_Toc49937078)

[Stages 4-6 1](#_Toc49937079)

[Table of contents 2](#_Toc49937080)

[Scope 3](#_Toc49937081)

[The limitations of generating scientific data 3](#_Toc49937082)

[Evaluating experiments – an overview of accuracy, precision, reliability, and validity 3](#_Toc49937083)

[In-depth discussions on accuracy, precision, reliability, and validity 5](#_Toc49937084)

[Accuracy: emphasis on exactness 5](#_Toc49937085)

[Precision: emphasis on agreement 5](#_Toc49937086)

[Reliability: emphasis on stability 8](#_Toc49937087)

[Validity: emphasis on meaning 8](#_Toc49937088)

[Use of accuracy, precision, reliability, and validity in other fields of research 10](#_Toc49937089)

[Causality 10](#_Toc49937090)

[Reliability and validity of secondary-sourced information 11](#_Toc49937091)

[Errors of measurement 11](#_Toc49937092)

[Uncertainty 13](#_Toc49937093)

[Manipulating uncertainties 14](#_Toc49937094)

[References 16](#_Toc49937095)

## Scope

This document describes the following aspects of data analysis:

* evaluating data using accuracy, precision, reliability, and validity
* evaluating secondary-sourced information using reliability and validity
* errors of measurements
* calculating uncertainty.

## The limitations of generating scientific data

Scientific advances rely on a solid foundation of evidence. The strength of any scientific concept is only as strong as the quality of the evidence that supports it. Thus, scientists place great emphasis on gathering, manipulating, and analysing evidence.

Data are measurements and observations that scientists use to develop scientific conclusions. The term measurement refers to the amount or quantity of some particular property that a system possesses (e.g. the mass of an object). Often, scientists manipulate raw measurements before using them as data. Scientific investigations do not produce single measurements. Rather, the data are the averages of multiple, independent measurements (Eisenhart, 1968). Most measurements contain errors (note that scientific errors are not mistakes – this is discussed in a later section). It is vital to anticipate, identify and minimise (or even eliminate) sources of error in measurements. Indeed, proper data analysis requires an understanding of measurement errors. Every aspect of a scientific investigation must be scrutinised for errors, as they may affect the investigator’s conclusions. When experiments are repeated, the errors of measurement may compound. Therefore, scientists use several criteria to decide if an experiment, and the conclusions derived from it, are acceptable.

## Evaluating experiments – an overview of accuracy, precision, reliability, and validity

Data and scientific conclusions possess some distinguishing characteristics, such as accuracy, precision, reliability, and validity (Houser, 2009). Scientists use these characteristics to evaluate scientific experiments. Those characteristics are described in the table below and explained in greater detail in the following sections.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term | Working definition  (a simple, plain language description of the terms) | NESA definition  (from Syllabus) | Joint Committee For Guides In Metrology (JCGM) definition | Synonym | Notes |
| Accuracy | The extent to which a measured value agrees with its true value (that is - reference value). | Accuracy estimated taking into consideration the evident sources of error and the limitations of the instruments used in making the measurements | The closeness of agreement between a measured quantity value and a true quantity value of the measurand (the quantity that is measured). | Exact | Requires prior knowledge about the measurand (for example reference values) |
| Precision | The extent to which multiple measurements, made under identical or similar conditions, agree with each other. | Not defined | The closeness of agreement between measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions | Internal reliability  Dispersion Spread | **Measurement precision**: applied to repeated measurements in a single experiment.  **Instrument precision:** the precision of measuring devices (analogue and digital) |
| Reliability | The extent to which the findings of repeated experiments, conducted under identical or similar conditions, agree with each other. | An extent to which repeated observations and/or measurements taken under identical circumstances will yield similar results. | Not defined | Consistency  Repeatability  Reproducibility  Stability | **External reliability:** the reliability of measurements across multiple experiments. |
| Validity | The extent to which an experiment addresses the question under investigation. | The extent to which the processes and resultant data measure what was intended (Stage 4 or 5). | Where the specified requirements [of an experiment or instrument] are adequate for its intended use | Internal validity | **External validity**: the extent to which the results of a study can be generalised (for example extrapolated) |

## In-depth discussions on accuracy, precision, reliability, and validity

### Accuracy: emphasis on exactness

The accuracy of measurement refers to its closeness to or deviation from its true value (Eisenhart, 1968; Ranstam, 2008). The acceptable level of measurement accuracy depends on their intended uses (for example, the significant numbers of measured values), as well as on the limitations of the measuring instruments. For example, the mass of an electron is 9.10938356 x 10-31 kg. This value has a high accuracy (9 significant figures). However, unless some application requires this level of accuracy, we can approximate the mass of an electron to 9.1 x 10-31 kg.

In science, there are several standard measurements called fundamental physical constants. The datasheet that accompanies some stage 6 science HSC examinations contains data on some physical constants. For example, as shown in the datasheet, the charge on an electron is -1.602 X 10-19 C. This is the accepted reference (true) value for the fundamental charge. Therefore, an experiment to measure the charge of an electron should show the same value as the reference value. If that happens, the measurement is accurate. Therefore, determining the accuracy of measurement data **requires prior knowledge of the true value or reference data.**

### Precision: emphasis on agreement

There are two different contexts to precision: measurement precision and instrument precision.

Measurement precision refers to the extent to which repeated measurements, made under identical conditions, agree with each other (Eisenhart, 1968; Ranstam, 2008). It describes the dispersion of values around the mean (average) or median value of a set of measurements (dataset). One way to look at the precision of measurement is to determine the range of values obtained. Consider a simple experiment in which students record the mass of an object (four repeated measurements):

* Student 1:
  + Measurements: 102.5g, 101.1g, 98.9g 101.5g and 101.2g
  + Range: 102.5g – 98.9g = 3.6 g
* Student 2
  + Measurements: 97.2g, 100.5g, 105.3g, 98.7g and 106.3g
  + Range: 106.3g – 97.2g = 9.1 g

**Note:** Range = maximum – minimum

Student 1’s results are more precise than Student 2’s because the former’s data span a smaller range of values.

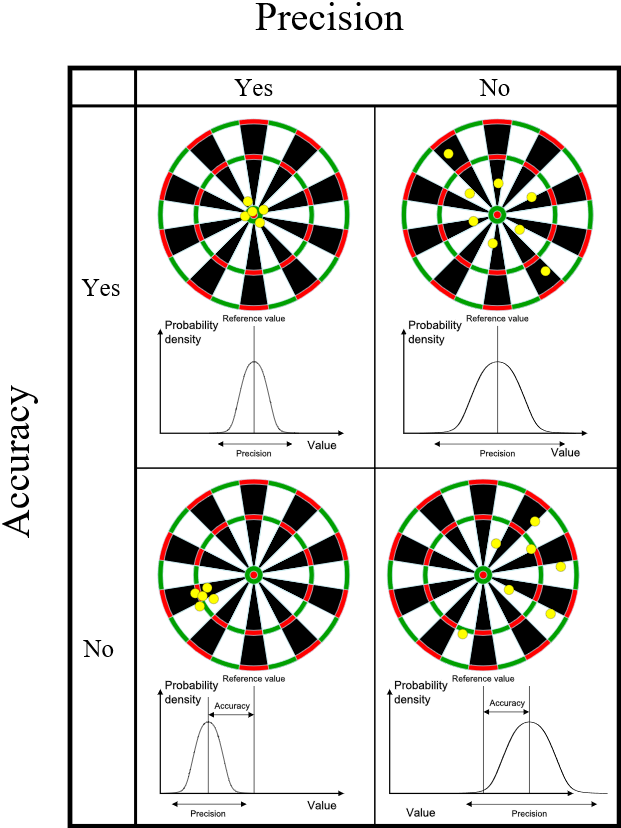
Statistics also provides other tools, such as variance, standard deviation and standard errors of means, to provide a measure of the precision of data in datasets. A discussion of these statistical tools is beyond the scope of this document.

Instrument precision refers to the error associated with measuring instruments. Measuring instruments may be analogue or digital. Digital measuring instruments are more sensitive and precise than analogue devices. Instrument precision and error are addressed later in this guide.

Sometimes, the lack of precision is not the fault of the measuring devices or the fault of the operator. Naturally chaotic systems (such as weather systems), or those in which large variations are inherent (for example biological systems), produce data that lack precision. In such cases, scientists often rely on another characteristic of measurements – reliability (see below).

**Notes**

* Although some publications use the terms ‘accuracy’ and ‘precision’ interchangeably, these terms refer to distinct characteristics of measurements. It is possible for measurements to be precise without being accurate (Houser, 2009).
* Precision is synonymous with another characteristic of measurements – internal reliability (discussed below).



This figure was obtained from [Arbeck, 2015](https://commons.wikimedia.org/wiki/File:Accuracy_and_Precision.svg) , [(CC BY 4.0)](https://creativecommons.org/licenses/by/4.0/deed.en)

The figure above describes the relationship between accuracy and precision. This activity aims to hit the bull’s eye in the dartboard repeatedly. Each of the four panels in the figure describes outcomes with varying levels of accuracy and precision. Besides the dartboard diagrams, each panel also contains a graph of the location (called the probability density) of the hits. The reference value shows the location of the bull’s eye (target), while the peaks of the graphs indicate the average locations of the hits. The further away the average location lines (peaks) are from the reference value, the less accurate the outcome of the activity is. Precision is related to the breadth of the graph – the broader the graph, the less precise the outcomes are.

### Reliability: emphasis on stability

Reliability refers to the repeatability or reproducibility of measurements (Stallings & Gillmore, 1971). Reliable measurements are those that are similar in value over multiple experiments (conducted under similar conditions). Such data are consistent and stable (Elasy & Gaddy, 1998; Houser, 2009).

Reliability is the cornerstone of the scientific process. When publishing research, scientists must describe all relevant information about their experiments (materials, methods, and all analytical procedures) so that others may repeat them and confirm the reliability of the findings.

Some authors distinguish between two types of reliability measures - internal and external:

* Internal reliability: the consistency of measurements within an experiment. Internal reliability is synonymous with precision
* External reliability: the consistency of measurements over multiple, independent experiments. Thus, if data from two or more identical experiments are plotted on a graph, there should be a strong correlation between the data points.

In the natural sciences, reliability refers to external reliability.

We can calculate the reliability of measurements and data mathematically using a correlation statistic called Cronbach alpha (similar to the Pearson correlation coefficient). Cronbach alpha is a number from 0 to 1, where 0 indicates no correlation between the data from replicate experiments, and 1 indicates complete correlation. Reliable data usually have an alpha value of 0.7 or greater (Downing, 2004; Houser, 2009).

There are other measures of reliability in disciplines such as education and psychology research. Some examples include subject reliability (e.g. subjects of a research experiment), interrater reliability (for example performance in surveys) and test-retest reliability (such as students in pre- and post-tests) (Houser, 2009). A discussion of these measures of reliability is beyond the scope of this guide.

There are many examples of unreliable scientific findings. For example, in 1989, Martin Fleischmann and Stanley Pons announced that they performed a nuclear fusion of hydrogen atoms at room temperature, a process called ‘cold fusion’. Other scientists could not replicate Fleischmann and Pons’ experiment. Hence, Fleischmann and Pons’ discovery of cold fusion is an unreliable finding (Jones et al., 1989). According to a study published in the journal Nature, 70% of researchers surveyed could not reproduce another scientists’ experiment (Baker & Penny, 2016).

### Validity: emphasis on meaning

The validity of an investigation refers to whether the procedures in it accurately measure what was intended to be measured (Eisenhart, 1968). It also describes the extent to which the findings of a study may be generalised or extrapolated to other situations. Thus, validity examines the ‘meaning’ or the ‘interpretation’ of the findings of an investigation (Streiner & Norman, 2006). Accuracy, precision, and reliability are all essential attributes of valid scientific findings.

There are two aspects to experimental validity: internal and external.

* Internal validity: whether the experimental procedures measure what the investigators set out to measure (Thatcher, 2010). For an experiment to be valid, it must satisfy the following criteria:
  + The hypothesis is properly constructed (Elasy & Gaddy, 1998).
  + All variables (dependent, independent and controlled) are identified. A well-designed experiment can establish the relationship between the dependent and independent variables because all other variables are controlled.
  + Systematic and random errors are minimised or eliminated.
  + The experimental procedure incorporates steps to repeat measurements.
  + The correct instruments or procedures are used in an experiment so that the measurements are relevant
  + Where grouping is involved, investigators randomly allocates samples to the different groups.
  + The findings of the investigation stem from the independent variable. In other words, there are no other plausible reasons for those results.
  + All analytical tools used for the analyses of data, including data manipulation, must be relevant and appropriate for the data types generated in the investigation.
  + All underlying assumptions must be carefully evaluated, including the settings in the software used for data analyses.
* External validity - the results of an investigation can be generalised or extrapolated beyond the immediate study. To accomplish this, the researcher must:
  + Address all the criteria described for internal validity.
  + Ensure that the samples used in the investigation are representative of the wider population.

Experimental outcomes that are internally valid allow researchers to establish causation (see below) – that is, establish the cause-and-effect relationship between the independent and dependent variables. Externally valid results allow researchers to predict future outcomes and extrapolate current findings to new conditions.

Validity relies on the accuracy of all the different components of a scientific investigation. It is a complex criterion that holistically evaluates many aspects of the scientific process. The instruments of measurements must operate with high accuracy. Thus, as described by NESA, validity depends on “… data, inferences, and actions produced from tests and other processes [that] are accurate”. These measures enhance the validity of scientific outcomes.

It is possible for reliable measurements to be invalid, but an unreliable cannot be valid. For example, an incorrectly calibrated instrument may generate precise and reliable measurements, which are invalid, as they do not measure the intended objective accurately (Houser, 2009; Mohajan, 2017; Twycross & Powls, 2006).

One example of an invalid study was published in the New England Journal of Medicine in 2013 (Estruch et al., 2013). The researchers announced that Mediterranean diets lowered the risk of heart attacks and strokes by 30% diet (external validity). In 2018, the journal retracted the article when other scientists discovered a flaw in the study’s design. Hence, its poor design invalidated a reliable experiment (Estruch et al., 2018). Other scientists have also cautioned about the reproducibility of scientific experiments (Baker & Dolgin, 2017). John Ioannidis, who is a medical scientist, points out that many reliable investigations may not be valid (Ioannidis, 2005). Investigations that have a small number of subjects, observations or measurements are at risk of being invalid. Using a statistical tool called Bayesian Analysis, Ionnaides predicted that many scientific claims might be false (Ioannidis, 2005).

## Use of accuracy, precision, reliability, and validity in other fields of research

The definitions of descriptors such as accuracy, precision, reliability, and validity described above are consistent in science, engineering, and statistics. However, publications in the fields of psychology, education and sociology research use the descriptions in slightly different contexts (Stallings & Gillmore, 1971). As described above, publications in the fields of psychology and education research use the terms ‘precision’ and ‘reliability’ interchangeably. Precision is also used to mean accuracy. Both reliability and validity of investigations are further elaborated using criteria such as subject reliability, inter-rater reliability, test-re-rest reliability, content validity, construct validity and criterion-related (Houser, 2009). A discussion of these criteria is beyond the scope of this guide.

## Causality

Some scientific investigations set out to determine the relationship between dependent and independent variables. If the variables correlate, then a relationship exists between them. However, correlation does not imply causation. To establish causation:

* the independent variable (cause) and the dependent variable (effect) must show a correlation
* the cause must precede the effect
* no other variable should be involved in causing the effect (specifically the cause is the result of the independent variable).

## Reliability and validity of secondary-sourced information

The definitions of reliability and validity described above refer to data from primary (or first-hand) investigations. However, these definitions also hold for secondary-sourced information. Thus, reliability refers to the consistency of information, while validity evaluates the appropriateness of the information (for the topic under investigation). Both reliability and validity of secondary-sourced information may be evaluated using the CRAAP framework, as shown in the following table.

|  |  |  |
| --- | --- | --- |
| Acronym | Term | Meaning |
| C | Currency | How recent is the information? \* |
| R | Relevance | Is the information related to the topic under investigation? |
| A | Authority | Who published the information? |
| A | Accuracy | Is the information accurate and reliable? |
| P | Purpose | What is the intention of the information? |

Adapted from [The CRAAP Test: Critically evaluating information sources](https://www.library.qut.edu.au/transcripts/craaptest.jsp) (accessed on 16 July 2018)

\* Currency refers to whether the ideas being evaluated are part of the contemporary research paradigm, rather than their dates of publication. Many important scientific ideas that were published a long time ago (more than ten years) are still relevant today and are cited in contemporary publications. For example, the paper describing the sequencing of the human genome was published 17 years ago but continues to be cited by researchers from the field today (there are more than 22,000 citations of that paper). Conversely, many newly published articles may be too tentative to be widely cited by the scientific community.

## Errors of measurement

In science, measurement errors refer to the difference between the measured quantity value and its true or reference quantity value (Bell, 1999). It is important to note that a scientific error is not a mistake in making the measurements.

Generally, there are two classes of errors in measurement (Bell, 1999). [Note that errors of measurement should not be confused with statistical errors].

* Systematic errors: These are the ‘component of measurement error that, in replicate measurements, remains constant or varies predictably’. Systematic errors are deviations from the true value by a constant amount. They are also called ‘biases’. Systematic errors affect the accuracy, but not the reliability of measurements. Repeating the measurements will not improve the accuracy of the data. Some examples of systematic errors include:
  + Calibration error. If a piece of equipment does not read zero when it should, it is not calibrated correctly. If a bathroom scale reads 2 kg when no one is standing on it, it will give a person’s mass as being 2 kg heavier than it actually is. Calibration error can be minimised by zeroing the equipment before it is used or by subtracting the false reading (what it reads when it should read zero) from each measurement.
  + Parallax error. When the object being measured is viewed at an angle, the scale being used to measure the object may not line up exactly with the object. This will lead to the measured value being higher or lower than the actual value. For example, when a car speedometer is viewed from the passenger’s seat, it will appear to show a higher speed than when viewed from the driver’s seat. Parallax error can be minimised. When reading from a scale, the scale needs to be viewed from directly in front. Some electrical meters have a mirror behind the needle; if the reflection of the needle can be seen, the observer is not viewing the scale from directly in front.
  + Limits of measurement. The increments on a scale determine the accuracy of a measurement. The length of a pencil can be measured more accurately with a 30 cm ruler marked in millimetres than with a 30 cm ruler marked in centimetres. The limit of reading for measuring equipment is half the smallest increment. For a thermometer where the increments are 1°C apart, the limit of reading is 0.5°C. Therefore, if the height of the column of liquid in the thermometer is closest to 24°C, the temperature is 24 ± 0.5°C. The equipment used in an experiment will affect the accuracy of the data collected. The manufacturer of the equipment will often provide specifications that indicate the expected uncertainties in measurements. For electronic sensors, this information will be provided when the sensors are purchased or are often available from the manufacturer’s website.
* Random errors: These are the ‘component of measurement error that in replicate measurements varies in an unpredictable manner’. Random errors cause deviations from true values by varying amounts. An example of a random error is the effect of environmental factors on measurement (e.g. temperature or humidity). Random errors affect the precision and reliability of measurements. Repeating the measurements can reduce random errors of measurement.

Mathematically, errors are represented as absolute errors or relative errors.

The vertical lines in the formula indicate that only the numerical value of the calculation, and not the sign, is considered. Therefore, even if the absolute error calculation returns a negative value, the minus sign is ignored.

These calculations can only be carried out if the true values are known.

Predictably, the closer the measured value is to its real value, the lower the error of measurement. Conversely, the lower the error, the greater the accuracy of the measurement (Streiner & Norman, 2006). Therefore, we must be aware of the errors of measurements when conducting and analysing experiments.

## Uncertainty

Uncertainty refers to the spread or dispersion of measurements (Bell, 1999; Roberts & Johnson, 2015). It implies that the true value of a measurement is not known accurately, but that it lies within a range of values. When reporting measurements, the uncertainty is shown with a ± symbol. Therefore, the measurement is reported as

Measurement ± uncertainty (for example 0.86 ± 0.05 s)

There are different ways of determining and reporting the uncertainties of measurements. The reported uncertainty falls into one of two categories (Bell, 1999; Eisenhart, 1968; Joint Committee For Guides In Metrology, 2012):

* Statistical uncertainty (sometimes referred to as Type A uncertainty (Bell, 1999))
  + The standard deviation from a mean (average)
  + The standard error of a mean
* Non-statistical uncertainty (sometimes referred to as Type B uncertainty (Bell, 1999))
  + Uncertainties in reference values
  + Uncertainties arising from instrument calibration or drift
  + Uncertainties resulting from random errors

In publishing scientific experiments, it is important to indicate the type of uncertainty of measurement (statistical or non-statistical uncertainty). For example, by stating that ‘data are reported as means ± standard deviations’, it becomes clear the uncertainty it is indicating.

There are two ways of determining the uncertainty associated with measurements:

**Analogue instruments** (for example a ruler): the uncertainty is half of smallest division. For example, standard laboratory thermometers show gradations of 1°C. Therefore, the uncertainty of measurements associated with this device is 0.5°C (that is, temperature ± 0.5°C).

**Digital instruments** (such as a voltmeter): the uncertainty is half of the highest significant figure. Consider a digital voltmeter that provides measurements to two decimal places. For example, if a voltmeter provides a reading of 1.55V, then the error of measurement is associated with the value of the number in the second decimal place (in this case, 0.05V). Since this number varies by 0.01V (the reading lies between 1.55V and 1.56V), the uncertainty of measurement is half of this value, or 0.005V. Therefore, this measurement is reported as 1.55 ± 0.005V.

## Manipulating uncertainties

When performing calculations with measurements, the uncertainties should also be modified accordingly. Measurement and uncertainty are treated separately. The following table shows how to manipulate uncertainties.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| What to do with the measurements? | How to manipulate the uncertainties? | Example | Strategy  M: measurement  U: uncertainty | Calculation  M: measurement  U: uncertainty | Answer |
| Add | Add the uncertainties | (5.0±0.1 mm) + (2.0±0.1 mm) | M: Add measurements  U: Add uncertainties | M: 5 + 2 = 7  U: 0.1 + 0.1 = 0.2 | 7±0.2 mm |
| Subtract | Add the uncertainties | (5.0±0.1 mm) + (2.0±0.1 mm) | M: Add measurements  U: Add uncertainties | M: 5 - 2 = 3  U: 0.1 + 0.1 = 0.2 | 3±0.2 mm |
| Multiply by a constant | Multiply the uncertainty by the constant | (5.0±0.1 mm) x 2 | M: multiply by the constant  U: multiply by the constant | M: 5.0 x 2 = 10.0  U: 0.1 x 2 = 0.2 | 10.0±0.2 mm |
| Multiply | Add the percentages of the uncertainties | (5.0±0.1 mm) x (2.0±0.1 mm) | M: multiply measurements  U: add percentage uncertainties | M: 5.0 x 2.0 = 10.0  U\*\*: 2% + 5% = 7% | 10±7% mm = 10±0.7 mm |
| Divide | Add the percentages of the uncertainties | (5.0±0.1 mm) ÷ (2.0±0.1 mm) | M: divide measurements  U: add percentage uncertainties | M: 5.0 / 2.0 = 2.5  U\*\*: 2% + 5% = 7% | 2.5±7% mm = 2.5±0.2 mm |
| Raised by a power | Multiply the percentage uncertainty by the power | (5.0±0.1 mm)2 | M: measurement raised to the power  U: uncertainty multiplied by the power | M: (5.0)2 = 25  U\*\*: 2% x 2 = 4% | 25±4% mm = 25±1 mm |

\*\* To convert the uncertainties into percentages, express the uncertainty as a fraction of the actual measurement:

0.1/5.0\*100 = 2% or 0.1/2.0\*100 = 5%

## References

* Baker, M., & Dolgin, E. (2017). Reproducibility project yields muddy results Gates Foundation demands open access. Nature, 541, 269–270.
* Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? Nature, 533(7604), 452–454.
* Bell, S. (2001). Measurement good practice guide no. 11 (issue 2). A Beginner’s Guide to Uncertainty of Measurement. National Physical Laboratory Teddington, Middlesex, United Kingdom.
* Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. Medical Education, 38(9), 1006–1012.
* Eisenhart, C. (1968). Expression of the Uncertainties of Final Results. Science, 160(3833), 1201–1204.
* Elasy, T. A., & Gaddy, G. (1998). Measuring subjective outcomes: Rethinking reliability and validity. Journal of General Internal Medicine, 13(11), 757–761.
* Estruch, R., Ros, E., Salas-Salvadó, J., Covas, M.I., Corella, D., Arós, F., Gómez-Gracia, E., Ruiz-Gutiérrez, V., Fiol, M., Lapetra, J. and Lamuela-Raventos, R.M. (2018). Retraction and Republication: Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. The New England Journal of Medicine, 378(25), 2441–2442.
* Estruch, R., Ros, E., Salas-Salvadó, J., Covas, M.I., Corella, D., Arós, F., Gómez-Gracia, E., Ruiz-Gutiérrez, V., Fiol, M., Lapetra, J. and Lamuela-Raventos, R.M.PREDIMED Study Investigators. (2013). Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. New England Journal of Medicine, 368(14), 1279–1290.
* Houser, J. (2009). Precision, reliability and validity: essential elements of measurement in nursing research. Journal for Specialists in Pediatric Nursing, 13(4), 297–299.
* Joint Committee For Guides In Metrology. (2012). Vocabulaire international de métrologie. VIM3: International Vocabulary of Metrology, 3(third), 104.
* Jones, S.E., Palmer, E.P., Czirr, J.B., Decker, D.L., Jensen, G.L., Thorne, J.M., Taylor, S.F. and Rafelski, J., (1989). Observation of cold nuclear fusion in condensed matter. Nature, 338(6218), p.737. Mohajan, H. K. (2017). Two Criteria for Good Measurements in Research: Validity and Reliability. Annals of “Spiru Haret”. Economic Series, 17(4), 59.
* Ranstam, J. (2008). Methodological note: Accuracy, precision, and validity. Acta Radiologica, 49(1), 105–106.
* Roberts, R., & Johnson, P. (2015). Understanding the quality of data: a concept map for ‘the thinking behind the doing’ in scientific practice. The Curriculum Journal, 26(3), 345–369.
* Stallings, W. M. W. M., & Gillmore, G. M. (1971). A note on “accuracy” and "precision. The Journal of Educational Measurement, 8(2), 127–129.
* Streiner, D. L., & Norman, G. R. (2006). “Precision” and “accuracy”: Two terms that are neither. Journal of Clinical Epidemiology, 59(4), 327–330.
* Thatcher, R. W. (2010). Validity and Reliability of Quantitative Electroencephalography. Journal of Neurotherapy, 14(2), 122–152.
* Twycross, A., & Powls, L. (2006). How do children’s nurses make clinical decisions? Two preliminary studies. Journal of Clinical Nursing, 15(10), 1324–1335.