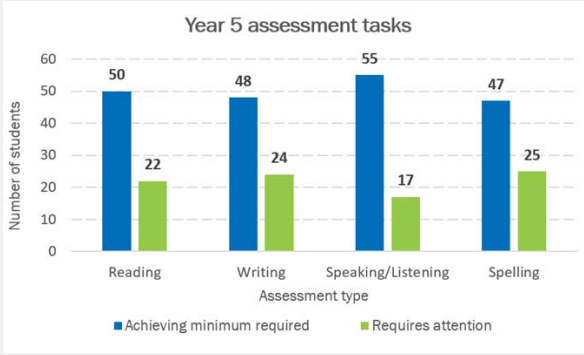
DATA GLOSSARY

Successful schools and teachers make use of multiple sources of data. They understand that using this information effectively can improve both instruction and student learning and wellbeing outcomes.

This glossary is intended to assist in understanding commonly used terms and concepts related to using data in schools.

A - D E - H I - L M - P Q - T U - Z

Glossary A - D

Term	Definition	Illustration
• **Aggregation**	The combination of data values (numbers) which are then expressed as a total for analysis purposes. For example, all Year 7 cohort data is aggregated to provide a total for the year group. Aggregation is often undertaken to de-identify data.	
• **Average**	The average or **mean** is used to summarise a dataset. It is calculated by adding up all the numbers and then dividing by how many numbers there are. It is used to represent the typical value in a **dataset**.	
• **Bar graphs**	Bar graphs are comprised of rectangular bars of lengths proportionate to the values they represent. They are most useful when comparing different categories of the same type of data. Bars are displayed horizontally, providing space for long axis labels.	
• **Baseline data**	The initial data collected against which subsequent data can be compared. Baseline	

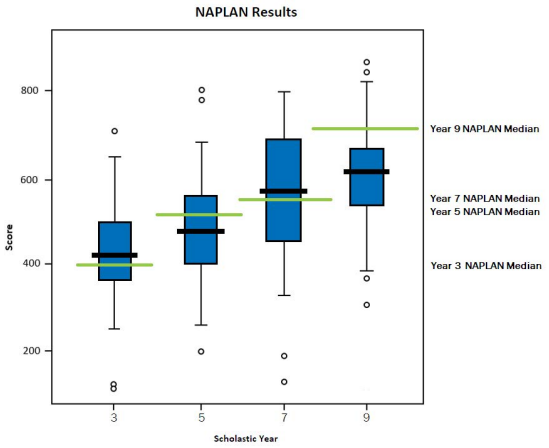
data is helpful when measuring impact and change.

- **Bias**

A statistic is biased if it is calculated in such a way that it is systematically different from the population being estimated. For example, if the **average** student performance in maths from one class (lets say the top maths class), then using this as the class average of all students at the school (the **population**) would be a biased estimate of the average performance in maths.

- **Box and whisker plots**

These are comprised of rectangular boxes and lines extending from the box. They are most useful when wanting to show both the values in the middle of a **dataset** and the **variability**. The lower whisker represents the bottom 25% of scores, the middle box the 50% of scores and the upper whisker the top 25% of scores. The line across the box shows the **median** score for the data set. These are also called box plots.



- **Causation**

One event (B) results from the occurrence of another event (A). There is a causal relationship between the two events insofar as event B will

	not occur if event A does not happen. Causation is different from **correlation**, a weaker relationship between two events or variables. Causation is more difficult to prove than correlation.
• **Census**	A census is a survey conducted on every individual in a given **population**.
• **Central tendency**	Describes what is typical for a set of data. It does not provide specifics about individual pieces of data, but rather a general overview of the entire **dataset**. There are three ways to show central tendency: **mean**, **mode** and **median**.
• **Cluster**	When data appear to gather around a single value. For example: in the following set of numbers 1, 2, 2.5, 4.5, 5, 5.5, 9 there is a cluster around the number 5.
• **Code**	A descriptor or label attached to the data used to represent a theme. This allows for text,

	images or audio (**qualitative datasets**) to be grouped into related categories for analysis.
• **Coding**	A data analysis process for organising, categorising and making sense of data. A code is a short summary or label that captures the overarching idea or meaning of a chunk of **qualitative data** (text, images, etc.). Data that display similar characteristics are labelled with the same code.
• **Cohort**	A group of individuals sharing a common characteristic. For example, age, gender or year level.
• **Column graphs**	Column graphs are comprised of rectangular columns of lengths proportionate to the values they represent. Columns are displayed vertically.
• **Content analysis**	A research method used to analyse meaning in text, images, audio and video (**qualitative data**) and to compare their various characteristics.

The process involves measuring the **frequency** and prominence of specific words and/or phrases.

- **Convenience sample** — Data is collected from or about a group of individuals because they are conveniently accessible. This is different from a **randomly selected sample** and the data they provide will not be representative of any larger **population**. A convenience sample, for example, might be all the students in a particular classroom. They may differ from students in other classrooms or schools.

- **Correlation** — Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more events or aspects. Correlation does not automatically mean that a change in one will cause, or is responsible for the change in the other.

- **Cumulative** — Data values that collected over time and added to the running total.

- **Data** — Data are information that we gather and analyse to inform decision making. Data take many forms, including words, numbers, sounds, and images. (See **qualitative** and **quantitative data**).

- **Data informed practice** — Data informed practice is the systematic use of educational data by schools, leaders and educators to improve student learning and wellbeing outcomes.

- **Data literacy** — Data literacy is the ability to understand and use data effectively to inform decisions. It consists of skills and knowledge that enable educators to transform data into meaningful information and actionable knowledge.

- **Data visualisation** — Data visualisation conveys information quickly and effectively. It can be useful when analysing and interpreting data to identify patterns and **trends** which may not be visible in the raw numerical data. For example, charts, graphs, infographics, and maps can visually transform large amounts of data into comprehensible information.

- **Dataset** — All of the data collected for a particular purpose or analysis.

- **Dependent variable** — The variable that is being measured to detect change due to the impact of an independent variable. For example a school introduces extracurricular programs to increase student engagement in science. Student engagement in science is the dependent variable, and the extracurricular program is the **independent variable**.

- **Disaggregation** — Disaggregation is the separation of data that has been combined to reveal individual values.

- **Dichotomous** — These data are **nominal** and only have only two categories, for example, yes/no.

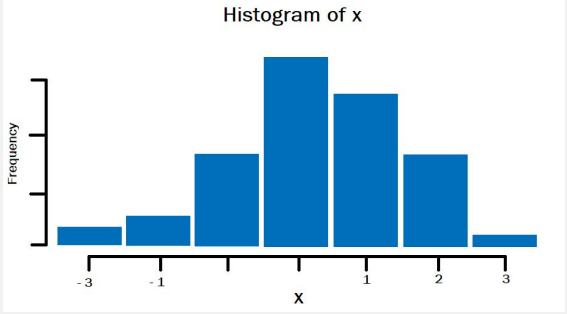
Glossary E - H

Back to top

Term	Definition	Illustration
Effect size	A statistical calculation which quantifies the difference between two **variables** and the impact of one variable on another variable. For example a school introduces extracurricular programs to increase student engagement in science. The effect size describes the amount of change in student engagement that is attributable to the extracurricular program.	
Frequency	The number of times an event occurs in a test or analysis of data.	
Frequency distribution	The count of how often something occurs. Frequency distributions are presented in tables which summarise the counts of particular data. For example, the count of gender of respondents in a **survey**.	

- **Generalisability** The extent to which findings and conclusions from data analysis can be applied to a wider **population**.

- **Histogram** A histogram represents the distribution of a continuous variable and allows for a visual inspection of the data for **skewness** and **outliers**. Each bar represents a range of the data. It differs to a **bar chart** because it focuses on a single **variable** that is continuous where as a bar chart may represent groups of discrete values.



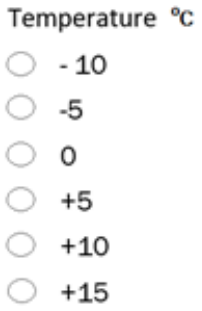
Glossary I - L

Back to top

Term	Definition	Illustration
Independent	The variable responsible for the change in a	

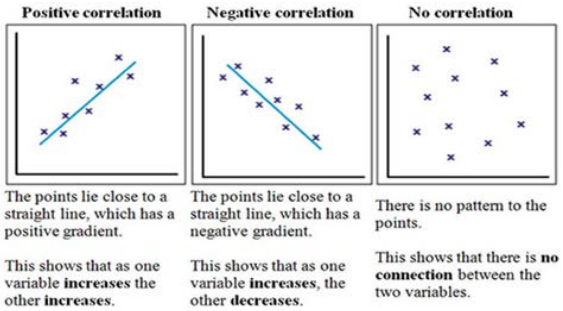
variable	**dependent variable**. For example, a school introduces extracurricular programs to increase student engagement in science. The extracurricular program is the independent variable and student engagement in science is the dependent variable.

• **Interval data**	Represent units of measurement that have the same difference between each point on the **scale**. These are numerical and ordered so we know the exact difference between the values at each point. For example, a thermometer measuring the temperature.



• **Interquartile range**	The interquartile range (IQR) is a measure of the spread of data around the **median**. The IQR is the difference between the upper (Q3) and lower (Q1) **quartiles**, and describes the middle 50% of values when ordered from lowest to highest. They can be represented by a distribution curve and **box and whisker plot**.

• **Linear correlation**	A measure of the relationship between two variables. The **correlation** is visible in how well data points fit a straight line when plotted



together. When all the points fall on the line it is called a perfect correlation. When the points are scattered all over the graph and there is no trend or pattern there is no correlation.

- **Likert scale**

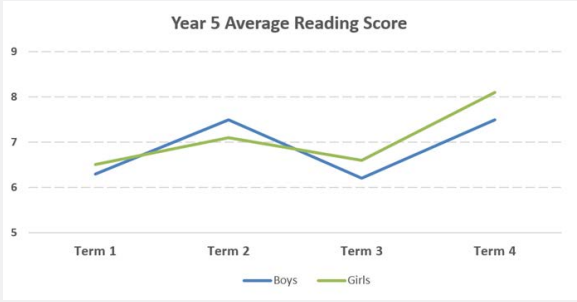
  Likert scales are used to measure respondents' attitudes to questions or statements. The responses are then coded and the value of the response is either higher or lower in value.

- **Likert-type scale**

  A type of rating scale used to measure attitudes or opinions using a list of response items that are more descriptive and granular than a yes or no. These scales are usually bipolar - they have an equal number of positive and negative responses around a neutral mid-point.

- **Line graphs**

  Graphs comprised of lines connecting different data points. They are most useful when showing **trends** over equal intervals of time such as weeks, terms or school years. These can also be used to compare changes of more



than one group over the same period of time.

• **Longitudinal study**	A study in which individuals are followed over time, and compared with themselves at different points in time.

Glossary M - P

Back to top

Term	Definition	Illustration
• **Mean**	One type of **average**, and a **measure of central tendency**. It is calculated by adding all the numbers in a dataset and dividing the sum by how many numbers there are. Means can be calculated on continuous data, not categorical data (for example we can calculate the mean of a series of numbers, but not a group of preferences).	

- **Measures of central tendency**

  Measures of central tendency show the centre of a data set. This includes the **mean**, **median** and **mode**.

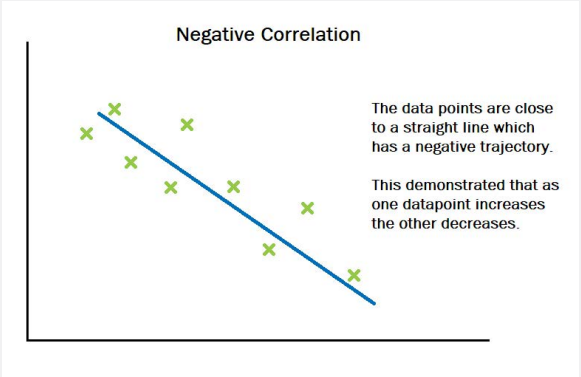
- **Median**

  A **measure of central tendency**. It is the middle score of a dataset when its scores are placed from lowest to highest.

- **Mode**

  A **measure of central tendency**. It is the value that occurs most often in a dataset. It is possible for a dataset to have multiple modes.

- **Negative correlation** — In a negative correlation, the two variables tend to go in opposite directions. As one variable increases, the other variable decreases, hence they are negatively correlated.



- **Nominal data** — Nominal data is categorical. Data represent counts of categories of objects.

- **Observation** — An occurrence of a specific data item that is recorded about a variable.

- **Operationalisat** — The process of identifying indicators (for example, behaviours, opinions, attitudes) that can be used to capture and measure an abstract concept (for example, motivation, engagement).

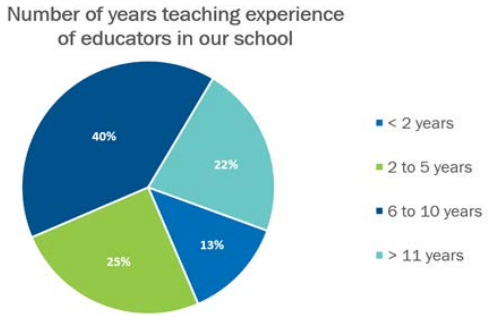
- **Ordinal data** — Ordinal data is categorical. These data can be

ranked.

- **Outlier** — An extreme, or atypical data value(s) that is substantially different from the rest of the data.

- **Percentile** — A number for which the corresponding percentage of scores fall below. For example, someone who scored in the 70th percentile scored better than 70% of those who took the test.

- **Pie charts** — Circular charts which are divided into slices to visualise numerical proportions.



- **Population** — A group of objects or individuals defined by a set of specific characteristics, for example, students at a school.

- **Probability** — The measure of the chance that an event will occur.

- **Probability/ran sampling** — A method of selecting a sample using a random selection process. It has many forms, including simple random sampling, simple stratified sampling and multi-stage **cluster** sampling. Randomisation increases the likelihood that the results will be generalisable to a wider population.

- **Purposive (or non-random) sampling** — A method for selecting a sample in order to fulfil a particular purpose, and which is not random. For example, participants may be chosen on the basis of particular characteristics.

Glossary Q - T

Back to top

Term	Definition	Illustration
• **Qualitative data**	Data that are non-numerical and often in the form of text, images, and objects.	
• **Quantitative data**	Data that are numerical. Numbers are used to represent values or counts.	
• **Quartiles**	Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).	
• **Quota sampling**	A method of sampling which aims to gather representative data for a group based on certain characteristics. It involves setting quotas of units for each characteristic of concern (e.g. age, gender etc.). Quota sampling is different to **random sampling**, because it requires that the individuals are chosen out of a specific subgroup (such as age, gender and ethnicity).	

- **Random sampling** — See **probability sampling**.

- **Range** — The difference between the smallest value and the largest value in a dataset.

- **Rate** — The occurrence of events over an interval of time, or the frequency of a phenomenon of interest.

- **Ratio** — Compares the frequency of a variable compared to the frequency of another.

- **Ratio data** — Data that have an equal and fixed ratio between each data point. It also has an absolute "zero" as a point of origin. Consequently there are no negative numerical values in the dataset.

- **Reliability** — Data reliability is when a dataset is sufficiently complete and error free, and repeated analyses will yield the same results. Data collection tools are considered reliable if they generate the same data from the same **populations** at different points in time.

- **Sample** — An identified subset of a population. Samples are selected based on characteristics of importance (age, gender or ethnicity).

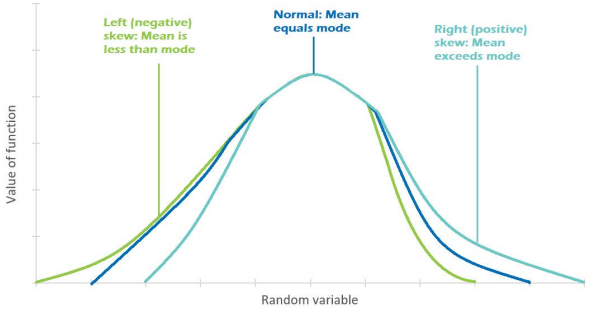
- **Sampling** — A technique used to select a subset of individuals or units from an identified population for data collection.

- **Saturation** — The moment during data collection when new data no longer reveals new information.

- **Scale (measurement)** — A list of items that respondents can choose in response to a question. Examples include: **nominal**, **ordinal**, **interval**, **ratio** and **Likert type**.

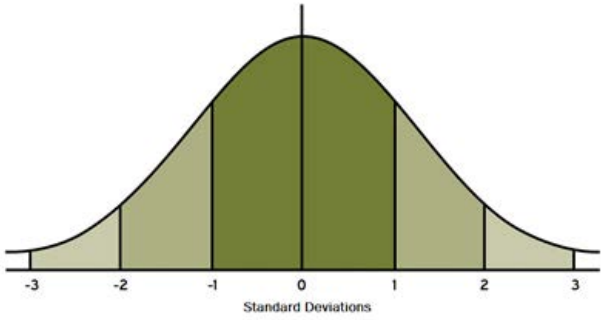
- **Scale (graphs and charts)** — The subdivision of each axis. The scale may be numerical or categorical.

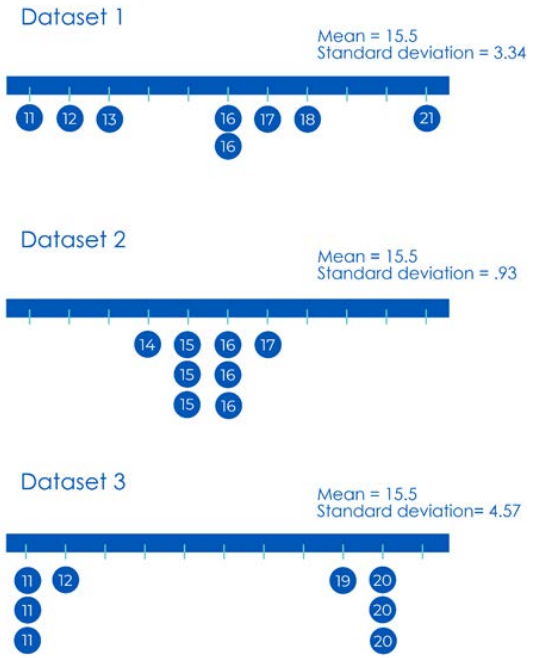
- **Skew** — Refers to the distribution of data values. Skewed data occurs when one side of the distribution curve deviates further from the "middle".



- **Snowball sampling** — A sampling technique that relies on social networks and referrals. For example, three teachers are interviewed and are then requested to ask their colleagues if they too would be available to be interviewed.

- **Standard deviation** — The standard deviation is a measure of **variation** in the data. It's the average distance from the mean of each unit of data.





- **Statistical literacy** — Refers to the knowledge and skills that enable data users and producers to understand, evaluate and communicate statistical data and information.

- **Survey** — Generally, a survey involves collecting information from a group of individuals for further analysis. Specifically survey tools are designed to gather information from a group of people using a set of standardised questions and response options.

- **Thematic analysis**

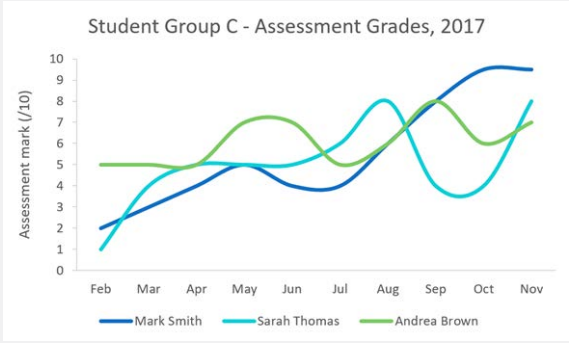
  A generic approach to analysing **qualitative data** by identifying patterns of meaning. It involves labeling chunks of data using **codes**, which can then be grouped together into themes.

- **Timeliness**

  Timely data accurately reflect the reference period for which they were gathered, and become available for use within an acceptable period of time.

- **Trend**

  The pattern of change over time for a specific **variable**. For example, the change in student assessment results over time. These data can be displayed graphically.



- **Triangulation**

  Using two or more sources of data, methods of collection or researchers to gather, analyse and interpret data. Triangulation is useful to verify analysis findings and interpretation, and increase and enrich understanding of the issue or problem being investigated.

Term	Definition	Illustration
T-test	A statistical test that compares two **sample averages** to determine if they are statistically different. It is appropriate for both small and large samples.	

Glossary U - Z

Back to top

Term	Definition	Illustration
Validity	The data actually show what they claims to show. In data collection, the collection tool measures what it claims to measure.	
Variable	A variable is any characteristic, number, or quantity that can be measured or counted.	

- **Variance** — A measure of how far each value in a dataset is from the **mean** (**average**). Variance can be used to measure **variability** (**volatility**) and spread. There are four commonly used measures of variability: **range**, **mean**, variance and **standard deviation**.

- **Volatility** — The measure of how much data fluctuates over time. For example, stock prices going up and down.

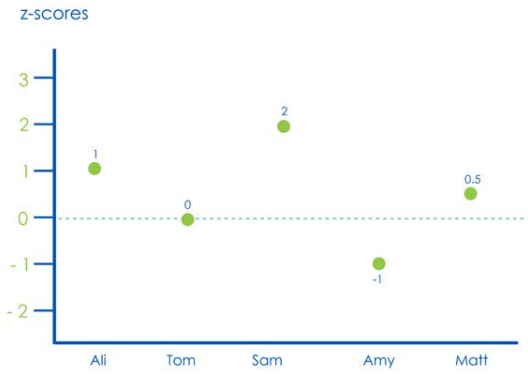
- **Variability** — How spread out a set of data is. Variability provides a way to describe how much data sets vary and allows the use statistics to compare one set of data to other sets of data. The four main ways to describe variability are range, interquartile range, variance, and standard deviation. Variability is also called spread or dispersion.

- **X-axis** — The horizontal number line on a graph.

- **Y-axis** — The vertical number line on a graph.

- **Z-scores**

Measures how many **standard deviations** a value is from the mean. For example, in a class of students with an **average** test score of 30 and a standard deviation of 5, if a particular student's test mark is 40, the z-score for that student would be 2 ((40-30)/5).



Contact AIS

The Association of Independent Schools of NSW Ltd
Level 12 / 99 York Street Sydney NSW 2000

Our postal address is the same as above

Telephone: (02) 9299 2845
Facsimile: (02) 9290 2274
ABN: 96 003 509 073