**Glossary of Data Terms**

**Access -** The process of retrieving data from the evidential database.

**Analytical Data Tool -** Software that provides analysis of data through the connection of data sources (e.g., SPSS, Excel, Fathom).

**Analysis -** A detailed examination of data and evidence intended to answer a question or reveal something.

**Aggregation -** A number of measures made into one.

**Attribute -** Otherwise known as a field. A characteristic of an entity, such as name, gender or ethnicity.

**Attribute Specification -** Specifications that control the kind of information that can be entered into the attribute, such as data type (e.g., number, text, date) or properties (e.g., field size, format, and validation rule).

**Benchmark -** A specific statement of knowledge and skills within a content area’s continuum that a student must possess to demonstrate a level of progress toward mastery of a standard.

**Box-and-Whisker Plot -** A visual display of the five number summary – minimum, lower quartile, median, upper quartile, and maximum. The box shows the quartiles and median. The whiskers extend all the way to the minimum and maximum values regardless of how far out they may be.

**Case** - The object (i.e., thing or person) of investigation for which data are collected; "the cases that we studied were drawn from two different communities".

**Categorical Data** - Data that represent different categories of a variable. For example, recording the colour of a car or the gender of a person is to collect categorical data.

**Census** - The process of systematically collecting data from a population.

**Central Tendency -** Measures of the location of the ‘middle’ or the ‘centre’ of a distribution. The mean is the most commonly used measure, but sometimes the median or mode are more appropriate measures.

**Code -** In most data files, text (e.g., ‘male’ or ‘female’) is converted into a numerical value for analysis. For instance, a "code" of 1 might be used for Maori students, a 2 for Pacific students, etc.

**Cohort -** A group of people sharing a common factor such as the same age or year level.

**Correlation Coefficient -** The correlation coefficient indicates the degree of linear relationship between two variables. The correlation coefficient always lies between -1 and +1. -1 indicates perfect linear negative relationship between two variables, +1 indicates perfect positive linear relationship and 0 indicates lack of any linear relationship.

**Cleaning Data -** The process used to determine inaccurate, incomplete or unreasonable data and then improving the quality through correction of detected errors and omissions.

**Criterion Referenced Tests (CRT) -** Evaluations that judge how well a test taker does on explicit objectives relative to a pre-determined performance level. There is no comparison to any other test takers.

**Cross Sectional Study -** In survey research, a study in which data are obtained only once.

**Cross-tab** - A table of survey results, with several rows and columns of figures. Can be created with data in a spreadsheet such as Excel (also known as pivot table).

**Cross Tabulation -** A table that contains two or more variables, usually with one across the top, and the others down the side.

**Dashboard -** A user interface that organises and displays critical data in a way that is easy to read. The asTTle ILP is an example of a dashboard.

**Data -** Data are the information collected and stored at the level at which the unit of analysis was observed. Data must be processed to be of practical use. Summaries of these data are usually statistics. Known facts or measurements, probably expressed in some systematic or symbolic way (e.g., as numbers).

**Database -** A system of complete, retrievable and organised information that is accessible electronically and that can be readily manipulated.

**Data Cleaning -** To "clean" a data file is to check for wild codes and inconsistent responses to verify that the file has the correct and expected number of records or cases, and to correct errors found.

**Data Integrity -** The quality of correctness, completeness, wholeness, soundness, and compliance with the intention of the creators of the data. It is achieved by preventing accidental o deliberate but unauthorised insertion, modification or destruction of the data in a database.

**Data Mining -** The process of searching for and finding valuable patterns (that are not always obvious) from a large collection of data. Sometimes software tools are used for those who typically do not know exactly what they are searching for but are looking for patterns or trends. Data mining is the process of sifting through large amounts of data to produce data content relationships. Also known as data surfing, or fishing expeditions.

**Data Teams -** A group of people who are committed to using data to improve student achievement. At the school level, data teams would include knowledgeable teachers who might serve as “school techs” who plan and coordinate professional development activities so all teachers could use the available data.

**Data Visualisation -** Graphical display of data and models which helps the user in understanding the structure and meaning of the information contained in them. Displays help use the high capacity of the human brain to visually recognise patterns and trends.

**Deciles -** The values of a variable that divide a frequency distribution into ten equal frequency groups. In New Zealand, schools are grouped into ten equally sized groups by a scale that is used for funding purposes according to the needs of the school. This is determined by taking a random sample of the student body and comparing their home address with census data. School decile is used as a proxy for socio-economic status.

**Demographics -** Data relating to characteristics of groups within the school’s population. Data that provides a *profile* of people at your school.

**Descriptive Statistics -** Figures which summarize or describe a data set, without making any inferences or generalizations.

**Disaggregation -** See aggregation. When you disaggregate data, you take aggregated data apart to see what you can discover from the component parts. For example, a student may do moderately well across a whole subject, but you need to disaggregate the year’s result to see where her weaknesses lie.

**Drill Down -** Allows the user to move between levels of data ranging from the most summarised (up) to the most detailed (down).

**Export Data -** To shift data out of a database/repository.

**Evaluation -** Any process of reviewing or making a judgement about a process or situation.

**Evidence -** Any facts, circumstances or perceptions that can be used as an input for an analysis or decision.

**False Negative -** The error that occurs when a decision maker accepts a result as true when it is actually false. Also referred to as a Type II error.

**False Positive -** The error that occurs when a decision maker rejects a result when it actually is true. Also referred to as a Type I error.

**Formative Assessment -** On-going assessments designed to monitor student achievement over time; allows staff to determine what part of the task a student knows and does not know, and enables adjustments to be made to the teaching/learning process.

**Frequencies -** In survey research, the number of respondents who responded to each of the possible answers to a question.

**Gain Score -** The change or difference between two administrations of the same test. Gain scores are calculated by subtracting the previous score form the most recent score. Gain scores can be negative, which indicates the student knows less than previously estimated by the assessment.

**Grade Point Average (GPA) -** The average grade earned by a student calculated by dividing the grade points earned by the number of credits attempted. For NCEA results, GPA can be calculated by assigning 4 points to Excellence, 3 to Merit, 2 to Achieved, and 0 to all other results.

**Heat Maps -** A colour-coded data chart where colors are used to differentiate values in a data set. Useful in seeing patterns in large datasets.

**Historical Data -** Data from previous time periods, in contrast to current data. Historical data is used for trend analysis and for comparison with previous periods.

**Hypothesis -** A tentative explanation for a phenomenon, used as a basis for further investigation.

**Import Data -** To shift data into a new database/repository.

**Information** - Knowledge gained from analysing data and making meaning from evidence.

**Interoperability -** With respect to software, describes the capability of different programmes to exchange data via a common set of business procedures, and read and write the same file formats and use the same protocols.

**Interpretation -** An explanation or establishment of the meaning or importance of something as the result of an investigation.

**Inter-subject Analysis -** A detailed examination of data and evidence gathered from more than one learning area.

**Intervention -** Any action that you take to change a situation, generally following an analysis of data and evidence.

**Intra-subject Analysis -** A detailed examination of data and other evidence gathered from within a specific learning area.

**Item Analysis -** Identifying how many students within a population selected each answer to a specific question on an assessment.

**Likert Scales -** A way of measuring phenomena (e.g., attitudes or events) using a response scale such as "strongly agree", "agree", "undecided", "disagree", "strongly disagree". There are a variety of scale types – for example, agreement, intensity or frequency.

**Longitudinal** **Analysis -** A detailed examination of data and evidence to reveal trends over time. Data is recorded for a panel of individuals repeatedly over a period of time.

**Longitudinal Study -** A study in which individuals are followed over time, and compared with themselves at different times, to determine, for example, the effect of aging on some measured variable. Longitudinal studies provide much more persuasive evidence about the effect of aging than do cross-sectional studies.

**Margin of Error -** A measurement of the accuracy of the results of a survey. Most commonly reported for political polls.

**Maximum -** The highest actual score or the highest possible score on the test.

**Mean -** Average score of a set of scores. Calculate by summing all the scores and dividing by the total number.

**Median -** The score that splits the distribution in half: 50 percent of the scores lie above and 50 percent lie below the median. If the number of scores is odd, the median is the middle score. If the number of scores is even, one must add the two middle scores and divide by two to calculate the median.

**Metadata -** The descriptive information about the data in the database – data about data. Metadata describe the structure and relationships of the data pieces. This includes such items as the names of the source fields that the data come from, the calculations to transform the data before loading them into the target database, the names of the columns that the data are going into in the target database, and the meaningful end-user orientated descriptions of the target database tables and columns.

**Minimum -** The lowest score or the lowest possible score on the test.

**Missing Data -** Occurs when no data value is stored for the variable in the current observation. Missing values are a common occurrence (e.g., because a student is absent for a test).

**Mode -** The score which occurs most frequently in the scoring distribution.

**Normalisation -** The process of decomposing large tables (files) into smaller table in order to eliminate redundant data and duplicate data and to avoid problems with inserting, updating, and deleting data. Normalisation is a way of organizing the data to prevent data duplication and to preserve strict relationship semantics.

**Normal Curve Equivalent -** Standard scores with a mean of 50 and a standard deviation of 21.06 and a range of 1 to 99.

**Normal Distribution -** Also known as a normal curve. A bell-shaped distribution of scores where most of the scores group in the middle.

**Norms -** Representative standards or values for a given group against which individual scores can be compared.

**Norm-Referenced Tests (NRTs) -** Evaluations in which the scores are compared to a norm group, a representative sample of a specified population.

**Outliers -** Unusually large or unusually small values in the data. Outliers often arise from some mistakes in data-gathering or data-recording procedures (e.g., year of birth recorded as 5009 – probably a typo and should be 2009). The first step in data cleaning is to check for outliers.

**Percent Correct -** A calculated score implying the percentage of students meeting and exceeding some number, usually a cut score, or a standard. Percent passing=the number passing the test divided by the number taking the test.

**Percentile Rank (PR) -** Percentage of students in a norm group (e.g., national or local) whose scores fall below a given score. Range is from 1 to 99. A 50th percentile ranking would mean that 50 percent of the scores in the norming group fall below a specific score.

**Perceptions Data -** Information that reflects the opinions and views of students, staff, or parents, often gathered through questionnaires.

**Pivot Table -** A statistical software programme output table, which summarises data that can be pivoted interactively. Some pivoting options include rearranging rows and columns, creating multidimensional layers, and showing and hiding cells.

**Population -** The entire set of cases or individuals under consideration in a statistical analysis (e.g., all students in a school, or all Year 11 students).

**Quantitative -** Relating to or based on the amount as a result of measuring, or number of something as a result of counting.

**Qualitative -** Relating to or based on the quality or character of something, as opposed to its size or quantity.

**Quartiles -** There are 3 quartiles – Q1, Q@, Q# - that divide the distribution into 4 equal groups. (Q1=25th percentile; Q2=50th percentile; Q3=75th percentile).

**Query -** A request one makes to a database that is returned to the desktop. Understanding and knowing how to set up queries or to ask questions of the database is very important to the information discovery process.

**Range -** A measure of spread between the lowest and the highest scores integrate the distribution. Calculate by subtracting the lowest score from the highest score.

**Raw Scores -** A person’s observed score on a test or subtest. Number of questions answered correctly on a test or subtest; simply calculated by adding the number of questions answered correctly.

**Relational Database -** A type of database that allows the definition of data structures, storage, retrieval operations, and integrity constraints (requirements that must be satisfied for the database to maintain integrity). The data and relations between then are stored in the table form. Relational databases are powerful because they require few assumptions about how data are related and how they will be extracted from the database. Most school systems are relational databases.

**Reliability -** A measure of the degree to which a measurement can be trusted because it is consistent in different situations. Indicates the consistency of an assessment instrument to obtain similar scores over time.

**Rubric -** A scoring tool that rates performance according to clearly stated levels of criteria. The scales can be numeric or descriptive. For instance, school process rubrics are used to give schools an idea of where they started, where they want to be with respect to implementation, and where they are right now.

**Sample** - A sample is a portion or sub-set of the population. A sample is chosen to make inferences about the population by examining or measuring the elements in the sample.

**Scalability -** The ability of a system or data warehouse to accommodate future needs by “growing.”

**Scaled Score -** A mathematical transformation of a raw score.

**Server -** Hardware that houses server applications.

**Spread -** The degree to which scores on the variable differ from each other. There are many measures of spread (e.g., range, inter-quartile range, standard deviation).

**Standard -** Something that is considered the basis of comparison or a guideline that is used as a basis for judgement.

**Statistical Significance -** The word ‘significant’ should only be used in describing some phenomenon if a statistical test (called a null-hypothesis test) has been conducted, and you are confident (within specified limits) that the phenomenon is not likely to be the result of a chance event.

**String -** A variable type that consists of alphanumeric characters (e.g., Noddy or sirv003 or l8m8).

**Table -** A data structure for relational databases, comprised of rows and columns, like a spreadsheet.

**Tabular Reports -** Reports formatted to look like a table. Also used in asTTle to present results in a form similar to a teachers mark book.

**Triangulation -** Use of two or more methods in a study with a view to double (or triple) checking results. Term used for combining three or more measures to get a more complete picture of student achievement.

**T-score -** A calculated standard score with a mean of 50 and a standard deviation of 10.

**User Interface -** The way a user interacts with data or communicates with the system.

**Validity -** A measure of the degree to which the instrument being used for measuring a phenomenon is “fit for purpose”. Indicates the effectiveness of an assessment instrument to measure what it is intended to measure.

**Variable -** Each item of data (e.g., age of person, ethnic identity, Term 1 test score) is called a variable.

**Variance -** A measure of dispersion, or variability, of scores about their mean. The population variance is calculated by taking the average of the squared deviations from the mean – a deviation being defined as an individual score minus the mean.

Terms taken and/or adapted from:

Bernhardt, V. L. (2007). Translating data into information to improve teaching and learning.

Ministry of Education (200X). Consider the Evidence. Retrieved 28 May 2010 from<http://www.tki.org.nz/r/governance/consider/>