Data Terms

Numerical Data: Data that are numerical have values that represent numerical quantities that can be ordered and compared mathematically. A quantitative variable takes numerical values for which arithmetic operation such as difference and averages make sense. Quantitative data is a useful alternative term for numerical data.

Categorical Data: Data that are categorical have values such as red and blue, or dog and cat. These values cannot be compared or ordered as quantities. The order in which they are placed on a graph or table is arbitrary.

Range: The range gives an indication of the spread of the data. The range of the data is technically the difference between the highest and lowest values in the data set. Often students talk about the minimum and maximum values in the data set rather than the distance between them.

Measures of Center: Students learn to use measures of center to summarize a data set. Building on children's informal understanding of what is the most, what is the middle, and what is typical, teachers can help students develop understanding about the mode, median. But students need to learn more than simply how to identify the mode or median in a data set. They need to develop an understanding of what these measures of center tell us about the data, and what each does and does not indicate about the data set. Children can see where the median is located among the data.

- **Mode:** The mode of a data set is the value at which more data occur than at any other value. A data set might have one or several modes. Modes are not generally used in statistics for the analysis of numerical data as they do not take into account all the values in

the data and so may not communicate anything useful about the data set as a whole. The mode is easy for students to identify. Students need to think about the mode in the context of the entire data set. Are there clumps around the mode? Mode is often used to describe categorical data, where the most frequent value may have more meaning.

- **Mean:** The mean is abstract as it has no clear identity within the data themselves. The arithmetic mean is a commonly used statistical average. It can be thought of as a "balance point" or as an evening out of the data. It takes into account all of the values in the data set. One way to think of the mean is as the value that would result if all the values in the data set were evened out. Another way to think of the mean is like the fulcrum of a seesaw. If all the pieces of data were represented on a line plot, the mean is the point at which the number line would balance. Students are typically taught to find the mean by finding the sum of all the pieces of data and dividing the total by the number of pieces of the data.

Median: The median is the midpoint of the data set. One way to think of the median is the value that would result if you listed all the pieces of data in order of their values and found the middle of the ordered list. The median cuts the data set in half. Half of the values in the data set are either equal to or greater than the median value, while half are either equal to or less than the median value. The median takes into account all the pieces of data in the data set, but it is not subject to much change by unusually small or large values because the median depends only on the order of values, not on the actual values. If there is an even number of pieces of data, the median is between the two middle values. The median is a stable indicator of the center of a data set and is often used for data, such as housing prices,

for which a middle value (rather than an “evening out”) is most useful. Use measures of center, focusing on the median, and understand what each measure does and does not indicate about the data set

Unusual Data Point or Outlier: An unusual data point is an outlier or a piece of data that is well removed from the rest of the data. An outlier is a value that is much higher or much lower than other values in the data set. An outlier may be an error in the data or an unusual value of interest that should be investigated further. Students should think about what might account for that value.

Values and Frequency: Every data set has both values and frequencies. Each piece of data has a value. This value might be a quantity such as ”6” or a categorical value such as “toys”. In both cases there are also frequencies which identify “how many pieces of data have a particular value”. There might be 5 students who have lost 3 teeth thus the value of 3 occurs with a frequency of 5. A graph that shows the frequencies of an ordered set of numerical values is called a frequency distribution.

Average: An average is a value that describes the center of a data set. In everyday usage, people usually assume that the word average refers to the arithmetic mean. However, average is also an inclusive term for any measure that is used to summarize the center of the data.

Measures of variability: An average or center of a data set gives an incomplete picture of the data. We also need to know how the data are spread around the center, or how they vary. It is important that students are able to describe the shape of the data. Two measures of variability are the standard deviation, used to show spread around the mean, and interquartile range, used to show spread around the median.

Shape of the Data: (Adapted from learningmath/data/session10 Describe the shape and important features of a data set and compare related data sets, with an emphasis on how the data are distributed. Children readily notice individual data points and are able to describe parts of the data where their own data falls on the graph, which value occurs most frequently, and which values are the largest and smallest. A significant development in children's understanding occurs as they begin to think about the set of data as a whole. Our goal for children is for them to see a data set as a distribution of values with important features, such as center, spread, and shape.

To focus students' attention on the shape and distribution of the data, it is helpful to build from children's informal language to describe where most of the data are, where there are no data, and where there are isolated pieces of data. The words *clusters, clumps, bumps,* and *hills* highlight concentrations of data. The words *gaps* and *holes* emphasize places in the distribution that have no data. The phrases *spread out* and *bunched together* underscore the overall distribution. Teachers must also continually emphasize and help students see that what they notice about the shape and distribution of the data implies something about the real-world phenomena being studied.

Questions:
Do you see any clumps of data?
Where are the data very spread out?
Where are data clustered together?
What values are more typical or usual in this data set?
What values are unusual?