The Routledge Handbook of Research Methods in Applied Linguistics

# Descriptive Statistics in Data Analysis

Jessica Briggs Baffoe-Djan, *University of Oxford*
Sara Ashley Smith, *University of South Florida*

Quantitative datasets are often very large, such that presentation of raw data is either impossible due to wordcount limitations (say, in a journal article or thesis) or impractical because the sheer number of figures makes meaningful interpretation of them via ‘eyeballing’ the raw data very difficult. The primary role of descriptive statistics (or *descriptives*) in data analysis is therefore to enable researchers to meaningfully describe and summarize quantitative datasets. In this chapter we discuss the ‘what’, the ‘why’ and the ‘how’ of descriptive statistics for data analysis in applied linguistics research: we define descriptives, discuss their strengths and limitations, and proffer a range of options for presenting them in theses, dissertations and/or journal articles.

## 1. Understanding descriptive statistics

Descriptives are intricately linked to the nature of the people who provide data in a given study. In some cases, the researcher has access to data from all of the people s/he is interested in researching (i.e. the entire population): for example, when a headteacher wants to analyse the exam scores of the pupils in their school. In other cases, the researcher does not have access to the entire population and instead recruits a sample of individuals to represent the larger population: for instance, when the researcher wants to make claims about all language learners of a particular profile but can only feasibly collect data from a proportion of those people. Descriptive values (such as means and standard deviations) applied to an entire population are termed parameters, whereas descriptives applied to a sample are termed statistics. It is important to note that regardless of whether the descriptive values reported in a given study are parameters or statistics, descriptives cannot and do not allow for making generalisations beyond the dataset itself. However, the distinction between population parameters and sample statistics is important: in applied linguistics (as, indeed, in the social sciences broadly) it is rare that a researcher will have access to an entire population, and therefore population parameters are largely unknown. For example, the population mean and standard deviation for L2 vocabulary size of all Japanese learners of English who have had, say, 100 hours of formal L2 English instruction in Japan will probably never be known. If a sample of the population is recruited to a study, and the sampling strategy adopted in the study means that the sample are representative of the wider population, then the descriptive statistics yielded from the dataset can be used in inferential statistical analyses to (1) yield estimates of population parameters (i.e. to generalise to the wider population) and (2) test hypotheses (also termed significance testing).

### 1.1 Frequencies

Measures of frequency are descriptive statistics used to indicate how often specific values occur in a given dataset. To calculate a *frequency distribution*, every value in a variable is listed (usually from lowest to highest) against the raw number of its occurrences. It is common to also report the *relative frequency* of each value, i.e. to state (usually in percentage) what proportion of the sample is represented by each value. The raw count and relative frequency of *missing values* (i.e. where one or more participants have not supplied a value) are also included in frequency distributions, such that the relative frequency always adds up to 100%. Frequency distributions are usually reported for variables which have *discrete values*: in other words, where participants have selected one option from a limited choice of responses (e.g., self-reported L2 proficiency by CEFR level). *Continuous values*, on the other hand, are those which the participant supplies from a potentially infinite number of options (e.g., hours of formal L2 instruction). Where a variable is comprised of continuous values, measures of central tendency and measures of spread are more appropriate than frequency distributions as means of characterising the data.

### 1.2 Measures of central tendency

Descriptives beyond frequencies can be categorised as measures of *central tendency* and measures of spread. Measures of central tendency are ways of expressing the central point of a set of data points and are therefore important descriptive measures for summarising quantitative data. There are a variety of measures of central tendency; deciding which to calculate and report depends on the type and nature of the data. The *mode* represents the most frequently occurring score in a set of values, such that if the set of values are plotted onto a bar chart, the mode is the highest point. The *median* refers to the middle score in a set of data values that have been ordered by magnitude (in other words, the median is the 50th percentile). Where there are an even number of scores, the middle two scores are averaged to find the median. The *mean* (denoted by the symbol $\overline{x}$, or x bar) is arguably the most commonly reported central tendency measure. It is calculated by dividing the sum of all values in a variable by the number of values in that variable. Note that the mean as denoted by the x bar is a sample mean: if the same value is calculated as a population parameter, the appropriate symbol is $\mu$ (mu). The mean is also commonly denoted by the letter M.

Which central tendency measure used to describe the data depends on a number of factors. The mode is commonly used with *categorical data* (e.g., location; sex). However, the mode can be problematic because in some cases there are two modes (i.e. two scores equally frequently occurring – termed *bimodal distribution*) and in other cases the mode is very far away from the rest of the values and is thus not a strong measure of central tendency. In these cases, measures of frequency are the best way to characterise categorical variables. The mode also is problematic to use with *continuous data* (e.g., test scores) because if a construct is measured continuously on a fine-grained scale (e.g., age in months), it is very possible that no one score will occur more than once. The median is commonly used to describe the central tendency of *ordinal data* (i.e. variables that have more than two ranked/ ordered categories).

Ordinal data (e.g., Likert-type items) are sometimes (and somewhat controversially) used to calculate a mean – thereby constituting an approximation of continuous data where the ‘gaps’ between all the scalar points (e.g., ‘agree’ and ‘somewhat agree’) are assumed to be equal. This is so that inferential analyses that require use of the mean (e.g., parametric tests such as the *t*-test) can be performed on the data. We recommend that unless a *Likert-type scale* is employed (i.e. the use of multiple similar Likert-type items that are combined to yield a composite score for a given construct), Likert-type data are treated firmly as ordinal, with frequencies or the median (not the mean) used as the measure of central tendency. Where a Likert-type scale is employed, a number of factors are likely to enhance the extent to which the data approximate equal intervals, thus strengthening the argument for using the mean as the measure of central tendency: (1) a greater number of items to be combined; (2) a greater number of response levels (i.e. more points on the scale); and (3) the use of numbered response levels (thus priming the participant to consider interval equality).

Either the mean or the median can be used as the measure of central tendency to describe continuous data; deciding which to choose will in part depend on the distribution of scores (or ‘shape’) of the variable. *Normal distribution* can be determined by plotting a set of values onto the x axis of a histogram against the frequency of occurrence of each value on the y axis: if the resulting shape is a symmetrical curve with one peak (i.e. resembling a perfect bell-shaped curve, with most of the values clustered at the centre), then the distribution of scores is normal. However, if there is a preponderance of values at the high end (right) or low end (left) of the x axis, the distribution is skewed: *skewness* to the right is termed positive skew and to the left, negative skew. A perfectly normal distribution of scores has a skewness value of zero, yet values between -0.5 and 0.5 are generally taken to indicate a distribution that is sufficiently symmetrical to be deemed normal. Another measure of shape is *kurtosis*, which refers to the peakedness or flatness of a distribution when compared against the normal distribution curve. The kurtosis value of a normal distribution is 3: a value higher than 3 indicates that the distribution is more peaked than the normal distribution as a function of *outliers* (i.e. scores at some distance from the other scores). A value lower than 3 suggests that there are fewer and less extreme outliers in the distribution as compared to a normal distribution. A important point to note, however, is that some statistical packages – including SPSS – compute 'excess kurtosis', whereby a value of 0 (and not 3) indicates normal distribution.

If the data are normally distributed, the mean or the median can be used as the measure of central tendency (perfect symmetrical distribution would denote that the mean, median and mode are all the same). However, the mean is the best option for normally distributed continuous data because it is the value that best diminishes error in predicting the values in the dataset. That is, unlike the median and the mode, the mean is a calculation derived from all of the values in the data, such that a change in an individual score will yield a change in the mean. However, because of its strong ties to all the values in the data, the mean is affected by asymmetry of distribution and the presence of outliers so where continuous data are non-normally distributed, it is better to report the median as the central value.

### 1.3 Measures of spread

Measures of spread are ways of expressing how dispersed (or 'spread out') a set of data points are. The range is the simplest method of expressing spread, representing the interval between the lowest and highest values (i.e. the lowest value subtracted from the highest value). The *range* is useful for detecting errors in the dataset (e.g., a value which is lower/higher than the actual options available) but has limited value as a descriptive statistic because of its susceptibility to outliers. That is to say, the range does not indicate the spread of most of the values in a variable, but rather only considers the lowest and highest. The *sample variance* ($s^2$) – arrived at by calculating the average of the squared differences from the mean – is another descriptive measure of spread. It tells very little by itself (other than that the higher the value, the greater the spread), but is a necessary step in calculating the *standard deviation* because the standard deviation is the square root of the variance. The standard deviation is the measure of spread used to accompany the mean and it provides the average distance of the scores from the mean. The standard deviation is appropriate to calculate and report for reasonably normally distributed continuous variables. Like the mean, there are two types of standard deviation: (1) the population standard deviation, used for data from an entire population, or if there are data from a sample of a population but the intention is not to generalise to the population; and (2) the sample standard deviation, whereby data have been collected from a sample and the intention is to generalise to the population (the sample standard deviation is the most common of these types). The formulae for calculating a sample and population standard deviation are slightly different, so if an online calculator is used to generate this statistic, the webpage needs to specify which type it is calculating. The symbol for population standard deviation is $\sigma$ (the lowercase Greek letter sigma); the sample standard deviation is usually denoted by the letter S or letters SD.

Datasets can be divided into up to 100 *percentiles* (cut-off points) that show what percentage of the sample has a value equal to or less than the percentile. *Quartiles* are the most commonly reported percentiles. Quartiles divide the scores in four to reveal the proportion of the sample whose scores fall into the 25th percentile (the lower quartile, or $Q_1$: i.e. the middle value between the lowest score and the median), the 50th percentile (the median quartile, or $Q_2$: i.e. the median), and the 75th percentile (the upper quartile, or $Q_3$ – the middle value between the median and the highest score). The *interquartile range* (IQR) is the difference between the upper and lower quartiles (i.e. $Q_3$ minus $Q_1$), thereby eliminating the influence of outliers by indicating the range of only the central 50% of the values. As it is impervious to outliers, the IQR can be seen as a more representative measure of spread than the range. The IQR is usually used in tandem with the median to describe the spread of ordinal data, or of non-normally distributed continuous data. It is common practice to report the IQR as a range (i.e. the $Q_1$ and $Q_3$ values) rather than as a value: this is to indicate to the reader the extent of any asymmetry in the IQR around the median (i.e. to show whether $Q_1$ or $Q_3$ is closer to the median).

Calculating descriptive statistics is a relatively straightforward endeavour, with most measures easily determined using Microsoft Excel. SPSS can also generate descriptive statistics such as frequency data, mean, standard deviation, minimum, maximum, skewness and kurtosis.

## 2. Using descriptive statistics

If a quantitative dataset is to be inferentially analysed (i.e. to generalise to the wider population and test hypotheses), then analyzing descriptives is vitally important. This is because the most powerful types of inferential statistical tests (*parametric* tests) operate under assumptions about the data that cannot be violated. Some assumptions are specific to certain types of parametric tests, whereas others are common to all or most (for a detailed discussion of assumptions by inferential test type, see, e.g., Pallant (2016), Field (2013) and Tabachnick & Fidell (2014)). The most common parametric assumption is normality of distribution, i.e. that the distribution of values resembles a symmetrical curve with one peak (a bell-shaped curve) – an assumption determined primarily via descriptives. If a variable is non-normally distributed, then a number of options are available. One option is the use of *non-parametric* inferential analyses: non-parametric tests do not operate under assumptions about distribution. They are useful to use with nominal and ordinal variables (because they do not assume that the mean is the best measure of central tendency); with smaller sample sizes (because a smaller sample less closely approximates the population distribution); and with continuous data that violate the assumption of normality. However, non-parametric tests are considered less powerful than their parametric counterparts because they do not offer the sensitivity afforded by the many parametric assumptions. As such, non-parametric tests may be less likely to detect a significant effect or relationship where one in fact exists.

The presence of non-normally distributed variables in a dataset does necessarily denote the abandonment of parametric tests: data may instead be manipulated into more normal distribution, and there are a number of techniques that can be applied for this. The dataset should first be checked for *outliers*. Outliers affect normality because if a value is well below the expected range, the shape of the distribution will be skewed to the left (and vice versa for outliers above the expected range). Outliers can be identified by eyeballing the histogram for data points at either extreme and/or by inspecting the boxplot or scatterplot. There are three main options for dealing with true outliers (i.e. those that actually derive from the data, rather than from a data entry error): (1) removal; (2) transformation; and (3) alteration. Outliers should be removed only if their removal will not compromise the generalisability of the findings to the population. In other words, care should be taken to determine whether the outlying values constitute a legitimate response from the population from which the researcher has sampled. If they are legitimate responses, they should be retained for reasons of generalisability. If not, they can be removed and descriptive analyses should be rerun on the trimmed dataset.

*Transformation* is a technique which alters the shape of a distribution such that the impact of outliers is mitigated. The choice of which type of transformation to apply to a variable is determined by the shape of the initial histogram (see e.g., Pallant (2016) and Tabachnick & Fidell (2014) for guidelines), and transformations can be easily computed using statistical packages such as SPSS. Once a transformation has been applied, descriptive analyses are rerun to gauge the extent to which the technique has improved the distribution. A note of caution, however: if a transformation is applied to a scale which is widely-used and has a canonical score interpretation (e.g., IELTS bands), then interpreting and explaining transformed scores is a complex endeavour. Another option for dealing with outliers is *score alteration*: this approach involves changing the outlying raw scores such that they deviate less from the majority of the other values. There are a number of options for exactly how to alter scores: for example, Tabachnick and Fidell (2014) recommend substituting the raw score of a low outlier with a score one unit larger than the next most extreme low (but not outlying) score (and vice versa for a high outlier), and Field (2013) suggests replacing extreme outliers with a score that is three standard deviations from the mean.

If a researcher decides to remove, transform or alter a dataset in order to make the distribution of one or more variables more normal, then these modifications should be clearly reported in any write up of the findings: specifically, the justification for and exact procedure of the modification should be detailed, and the descriptive statistics reported should pertain to the modified – rather than the original – dataset.

## 3. Presenting descriptive statistics

### 3.1 Presenting frequencies, percentages and other descriptive information

Text is often optimal for presenting your total number of participants and specifics across a finite number of characteristics or categories. For example, gender frequencies within sample and/or the total number of participants that fall into various background groups (i.e. language; school year) is often presented within the body of the manuscript.

> ‘The participant sample consisted of 35 young adults (15 female, 20 male) who live in the German federal state of Baden-Wüttemberg in the greater Tübingen area…’
>
> Poarch, Vanhove & Berthele, 2018, p.5.

> ‘Bilinguals were proficient in English and one of the following languages: Bengali (1); Cantonese (5); French (5); Greek (1); Korean (1); Mandarin (12); Portuguese (1); Spanish (5); and Vietnamese (1).’
>
> Chung-Fat-Yim, Himel & Bialystok, 2018, p.4.

More complex frequency and descriptive data with more categories should be presented in tables. There are cases in which your reader will need to know all categories and the number of participants in each, even when that number is zero. Table 1, for example, presents the total number of participants in a finite number of background categories. Tables allow comparisons across different groups and/or categories; for example, Table 2 presents dual language instruction ratios (percentage of instruction in each language) by grade level.

**Table 1**
*Participant maternal education*

| Highest level of education achieved | *n* |
|---|---|
| Less than high school | 25 |
| High school diploma or equivalency (GED) | 16 |
| Associate degree | 6 |
| Bachelor’s degree | 3 |
| Master’s degree | 0 |

| | |
|---|---|
| Doctorate | 0 |
| Professional degree (MD, JD, DDS, etc.) | 0 |
| Don’t know, prefer not to say | 0 |

Adapted from Smith, Briggs, Pothier & Garcia, 2017, p.7.

**Table 2**
*Percentage of subject-matter instruction in the two languages*

| Grade | English | Korean |
|---|---|---|
| K | 30% | 70% |
| 1 | 40% | 60% |
| 2 | 50% | 50% |

Adapted from Bae, 2007, p.304.

Frequencies can also be presented using *statistical graphics* (visual representations of quantitative data). A *pie chart***,** also called a *circle chart*, is a circular figure divided into “slices” that represent proportions within the data set. Each slice of the circle is proportional to the amount of the whole that it represents. Pie charts are helpful for visually representing subcategories that make up the whole. For example, Figure 1 presents the group ‘Refugee arrivals’ by language background subgroups. It is immediately apparent to the reader that no single language subgroup dominates; the largest subgroup is ‘Other’ (35%), multiple less commonly spoken languages with too few speakers to be represented as a single group. The second and third largest subgroups (Arabic speakers; Nepali speakers) each make up less than a quarter of the whole. Presenting these data visually underscores language diversity within the whole. *Side-by-side pie charts* can be used to demonstrate differences in demographic frequencies between two groups, as in Figure 2.

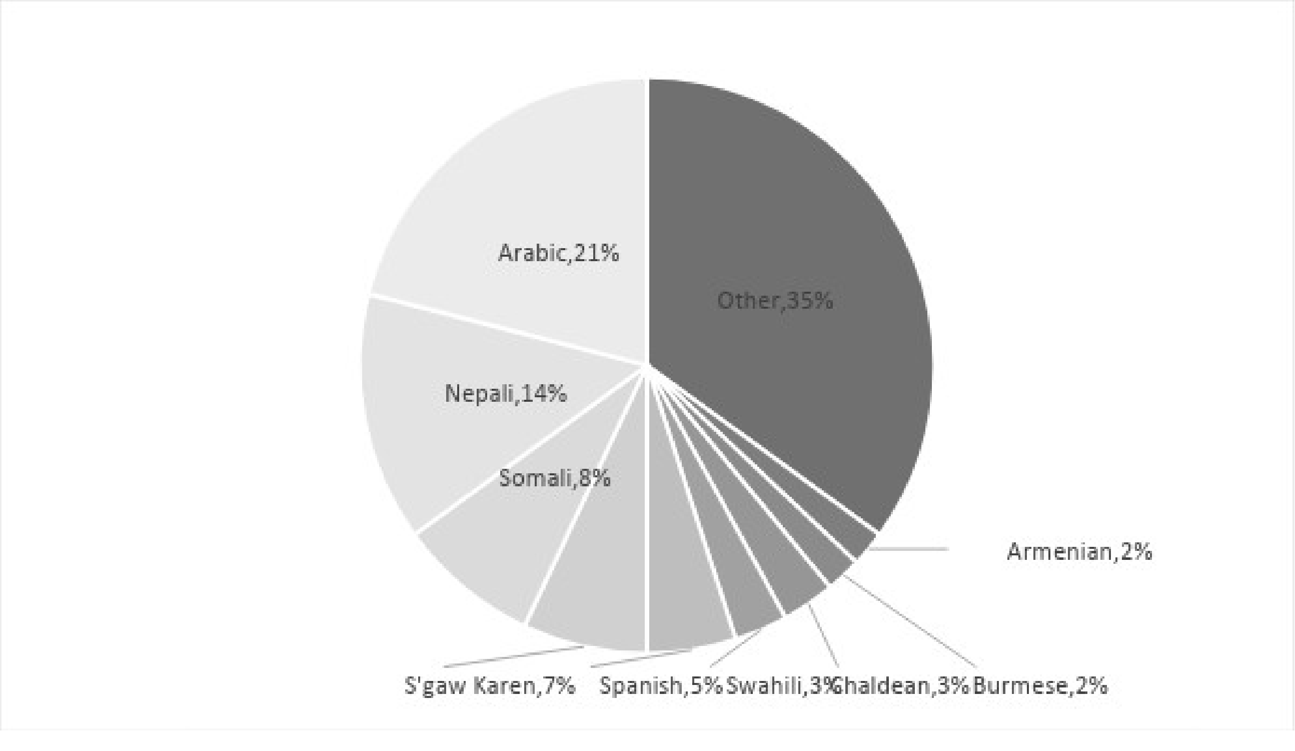

**Figure 1**. Top native languages of refugee arrivals, years 2008-2017. Adapted from: Park, Batalova & Zhong, 2018, p.24.

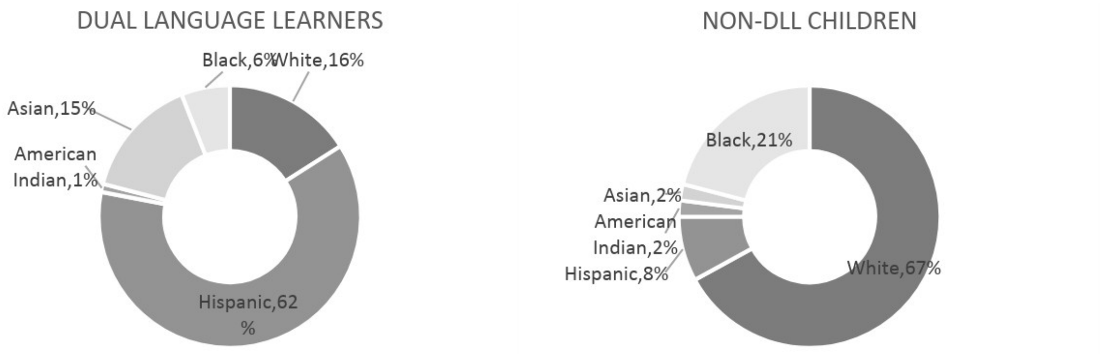

**Figure 2**. Race and ethnicity of dual language learner (DLL) and non-dual language learner children (ages 0-8 years). Adapted from: Park, Batalova & Zhong, 2018, p.11.

It is relatively easy to create visually appealing pie charts using Microsoft Excel, Microsoft Office or Microsoft PowerPoint. Additionally, most SPSS manuals (e.g., Pallant, 2016) include a section on how to generate graphs and charts in the programme for the purposes of presenting descriptive (and other) data.

Pie charts are not always optimal for representing multiple data sets, multiple time points, or categories that have values of zero or less than zero. Additionally, comparing different data across multiple pie charts or comparing proportions within the chart to each other can be challenging. For complex comparisons, bar charts are often a better way to visually represent data.

*Bar charts*, also called *bar graphs*, present data using vertical or horizontal rectangles and can be used to present frequencies, percentages, and other categorical data. The different length rectangles immediately display differences to the reader and allow for comparison across groups. Bar charts can demonstrate rankings, differences in frequencies between groups, and/or changes over time. For example, Figure 3 presents the percentage of conversational turns that contained code-switching (switching from one language to another) and compares conversational code-switching behaviour between two groups: (1) participants with 1-2 years' English language exposure; (2) participants with 3 or more years' English language exposure. The side-by-side vertical bars of different heights clearly demonstrate for the reader the difference in percentage of conversation that included code-switching.

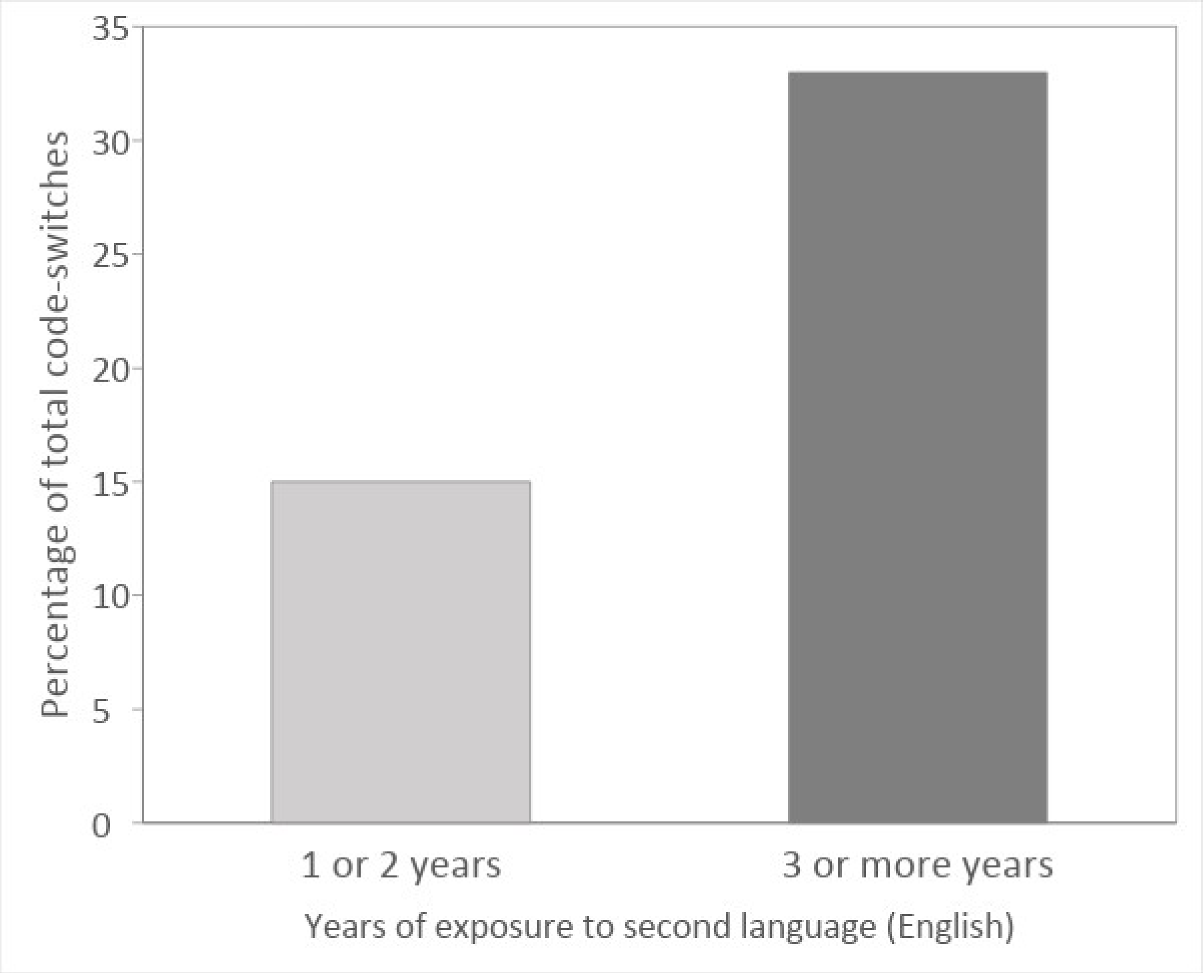

**Figure 3**. Percentage of 10-year-olds' conversational turns that used code switching, by number of years exposed to English. Adapted from: Reyes, 2004, p.88.

Bar charts allow comparisons between a subgroup and the entire sample. Figure 4 shows how one area within the U.S. state of Georgia differs from the state overall with regard to home languages spoken by parents of dual language learner schoolchildren: it clearly demonstrates that Fulton County has greater representation of children from Telugu, Hindi, Arabic, and Chinese speaking homes than the state as a whole. Bar charts can also show data from multiple time points: for example, Figure 5 presents the language backgrounds of parents of black dual language learner children, showing changes between the year 2000 and the years 2011-2015.

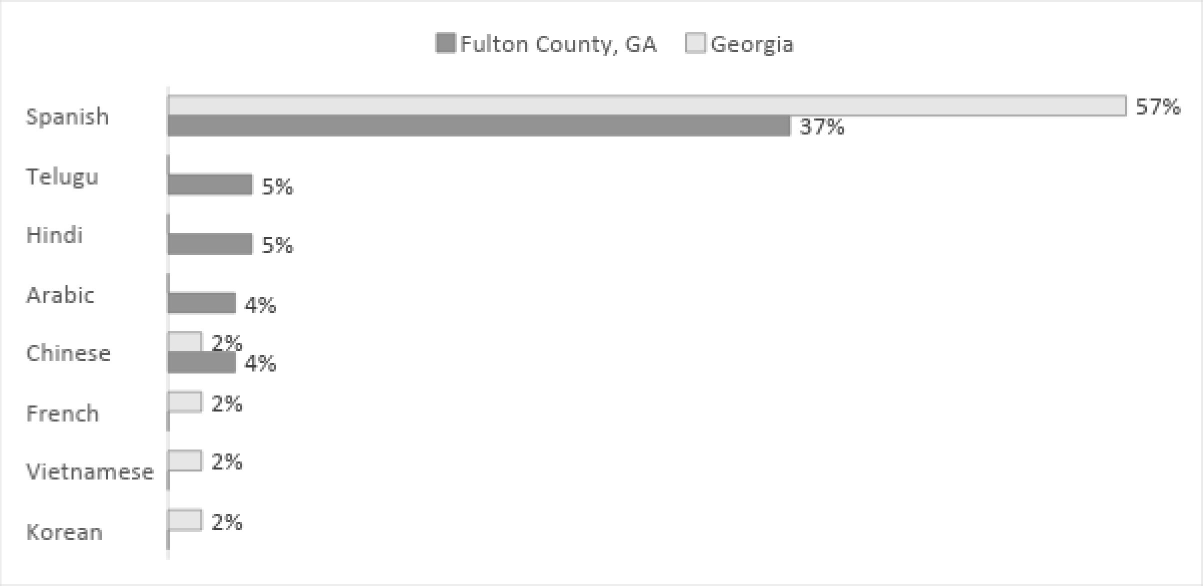

**Figure 4.** Top five non-English languages spoken by parents of dual language learner students, Georgia state and Fulton County. Adapted from Source: from Park, Batalova & Zhong, 2018, p.29.

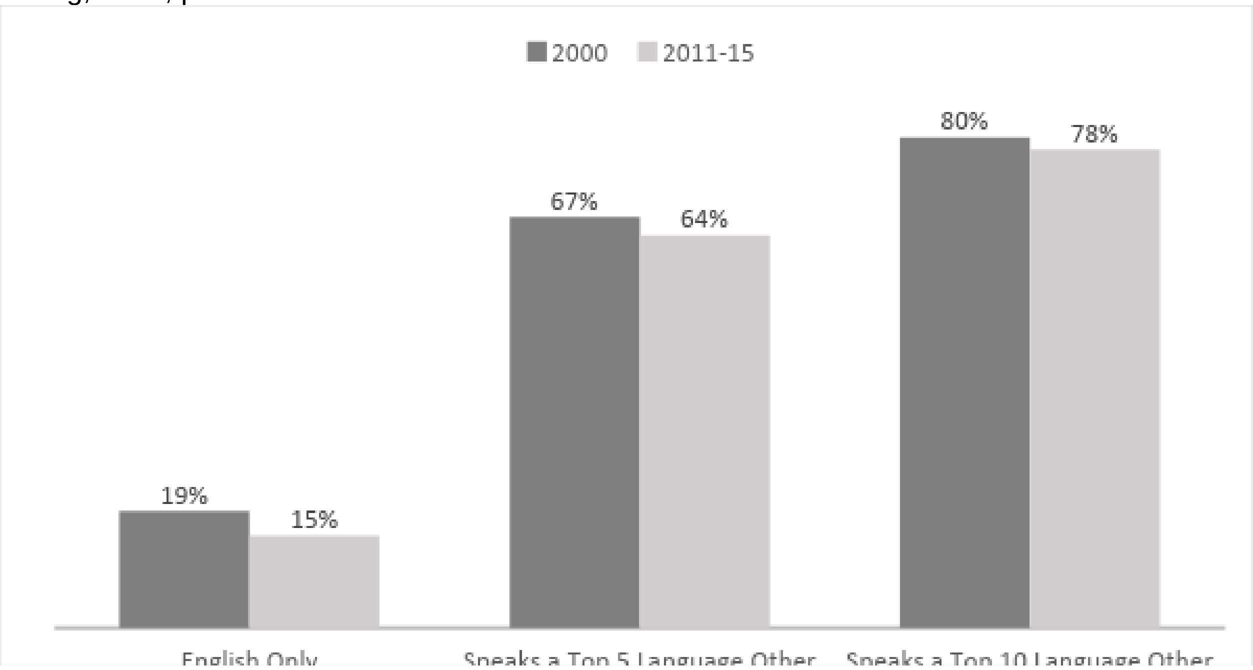

**Figure 5**. Linguistic diversity among parents of black dual language learner children, 2000 and 2011-2015. Adapted from: Park, Batalova & Zhong, 2018, p.20.

*Stacked bar charts* can be used to demonstrate how two categories together add up to the whole (e.g., two percentages that make up 100%). Figure 6 demonstrates the amount of Spanish and English used for reading among a sample of bilingual participants. While English is used more than 50% of the time, the ratio of the two languages depends on context.

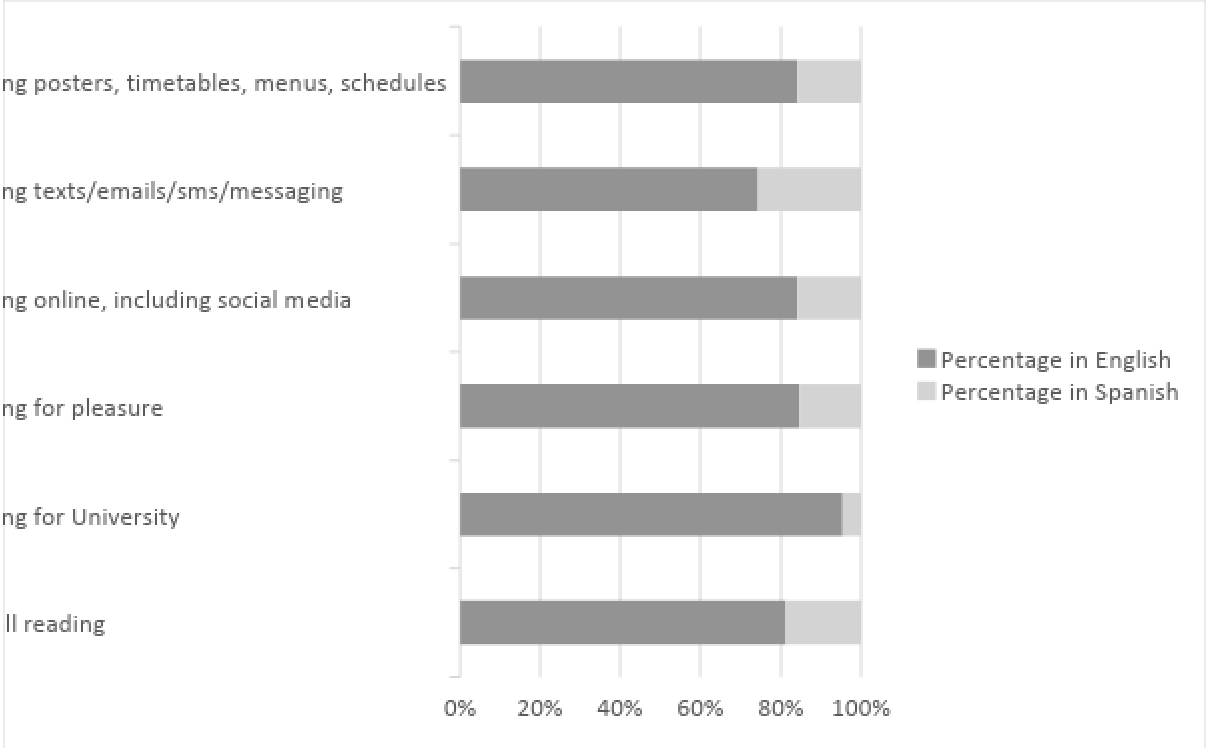

**Figure 6**. Spanish and English usage for reading behaviour. Adapted from: Smith, Briggs & Pothier, 2017, p.11.

### 3.2 Presenting descriptive self-report and survey data

Surveys and other means of self-report are commonly-used methods of data collection in applied linguistics research. Small amounts of these data can be presented in text, for example, results from two questions.

> ‘The amount of teaching experience varied, ranged from less than one year to more than twenty-five years. However, a majority of the respondents had fewer than ten years of teaching experience (70.6 percent). Whereas 36.4 percent of social studies teachers spoke English, about 15.2 percent of teachers indicated they were bilingual. Half of the respondents (48.5 percent) reported having learned a foreign language…’ Cho & Reich, 2008, p.237.

Data from ranking questions (questions that allow participants to order items based on a specified scale, e.g., most to least preferred) or questions in which participants select options from a longer list can be presented in text or tables. When presenting outcomes from a few questions, text is suitable; for example, data detailing teachers’ reported top three most desired supports for working with bilingual students. Tables are optimal for presenting multiple responses. For example, Table 3 lists each answer option (various challenges faced by teachers) and the percentage of participants who selected the answer option in their ‘top three most challenging’.

> ‘When asked what type of support teachers would like to receive... teachers indicated bilingual instructional materials as *most important*, followed by professional training and development.’ Cho & Reich, 2008, p.238.

**Table 3**
*Challenges facing teachers English as a Second Language Teachers.*

| Challenge | % of respondents* |
|---|---|
| Language barriers between you and English Language Learners (ELLs) | 58.8 |
| Cultural differences between you and ELLs | 5.9 |
| ELLs' lack of background knowledge of content area | 70.6 |
| ELLs' lack of motivation | 20.6 |
| Lack of guideline and/or support systems at school levels | 35.3 |
| Lack of time and resources to devote to ELLs | 41.2 |
| Assessment/grading of ELLs | 14.7 |
| *The sum of the percentiles does not reach 100 percent because we asked participants to mark their three biggest challenges (*N*=33) | |

Adapted from Cho & Reich, 2008, p.237.

### 3.3 Presenting behaviour frequency scales

When sharing results on Likert scales relating to frequency of a behaviour (*always, sometimes, never* etc.), the number of participants who selected each frequency choice should be presented. Means are not relevant for this type of data, as the mean of *rarely* and *always* is not *sometimes*, and presenting frequency of behaviour data using the mean would not covey findings in a meaningful way. For these types of data, presenting the percentage of participants who chose each agreement category better captures findings. Results from a single item, or a small number of items, and the corresponding responses can be presented within the text:

> 'About 65.6 percent of social studies teachers indicated they *always* or *often* allowed [English language learners] to have extra time to complete tasks.'
>
> Cho & Reich, 2008, p.237.

A table can help the reader make sense of findings from multiple items. For example, Table 4 presents how frequently teachers reported using various pedagogical accommodations in their teaching.

**Table 4**
*Accommodations made by teachers when working with English Language Learners (ELLs).*

| Accommodation | Rarely/never (%) | Sometimes (%) | Often/always (%) |
|---|---|---|---|
| I adjust my rate of speech for ELLs. | 24.2 | 48.5 | 27.3 |
| I provide different tasks and assignments for ELLs. | 78.1 | 18.8 | 3.1 |
| I allow ELLs to have extra time in completing tasks. | 9.4 | 25 | 65.6 |

| | | | |
|---|---|---|---|
| I provide different instructional materials for ELLs. | 65.6 | 31.3 | 3.1 |
| I assess/grade ELLs differently from the native English-speaking students. | 62.5 | 15.6 | 21.9 |
| I pair up (or group) ELLs so they can help each other. | 38.7 | 29.0 | 32.3 |
| I consult with ESL teachers in order to better help ELLs. | 31.3 | 28.1 | 40.6 |

Adapted from Cho & Reich, 2008, p. 238.

### 3.4 Presenting means

As prior discussed in this chapter, the mean is arguably the most commonly reported measure in applied linguistics research and provides the foundation for subsequent inferential statistics. Text can be used to communicate a single mean, or a small number of means. This is commonly done for presenting the mean age of a participant group. For example:

> ‘All the secondary students were selected from the first-, second-, and final-year populations (Years 11-13 in the UK system), with an average age of 16.5 years.’
>
> You & Dörnyei, 2016, p.5.

Where there are multiple means to present, a table can be used to clearly present these data. For example, a table can be organised to present mean scores on multiple measures, mean scores for different participant groups, or mean performance at different time points. See Table 5 for an example of means for two different groups (English language learners; English monolinguals), at two different time points (Autumn; Spring).

**Table 5**
*Means for ELLs and native speakers in breadth of vocabulary as measured by the Peabody Test (receptive vocabulary) English version, standard scores*

| | *Autumn* | | *Spring* | |
|---|---|---|---|---|
| *Group* | Mean | N | Mean | N |
| English language learners | 76.16 | 106 | 75.03 | 63 |
| English only | 110.41 | 205 | 115.45 | 84 |

Adapted from August, Carlo, Dressler & Snow, 2005, p.51.

Means are also meaningful when presenting data from Likert scales (symmetrical numerical scales on which participants specify agreement/disagreement level with a statement). Likert scale scores should be clearly labelled as such, so as to not confuse readers, as seen in Table 6.

**Table 6**
*Student responses*

| |
|---|
| Each figure gives the average response for each country on a scale from ‘Strongly agree’ |

<table>
<tr><td colspan="11">(=5) to ‘Strongly disagree’ (=1).</td></tr>
<tr><td colspan="11">a. In the classroom I see the teacher as somebody whose authority should not be questioned.</td></tr>
<tr><td>Brunei</td><td>Mainland China</td><td>Finland</td><td>Germany</td><td>Hong Kong</td><td>Japan</td><td>South Korea</td><td>Malaysia</td><td>Spain</td><td>Thailand</td><td>Vietnam</td></tr>
<tr><td>3.00</td><td>2.29</td><td>2.39</td><td>2.56</td><td>2.58</td><td>2.52</td><td>2.41</td><td>2.42</td><td>2.76</td><td>2.61</td><td>2.27</td></tr>
<tr><td colspan="11">The average for all countries was 2.47. The response chosen most frequently as 2 (‘Disagree’). For Asian countries the average was 2.46. For European countries the average was 2.53.</td></tr>
</table>

Adapted from Littlewood, 2000, p.33.

A *line chart*, also called a line graph, is a useful way to display data, such as means, over a period of time. A line chart presents dots or points on an x and y axis, with each point representing a specific value and points connected by a line. Line charts are useful for demonstrating continuously occurring data and facilitate visual representation of change over time. For example, Figure 7 presents the mean number of *restarts* (a form of spoken language error) at time points 6, 7, and 8 years old and shows decreasing means as a function of age. A single line chart can use multiple lines to represent different types of data (for example, different measures, as seen in Figure 8) or the performance of different groups.

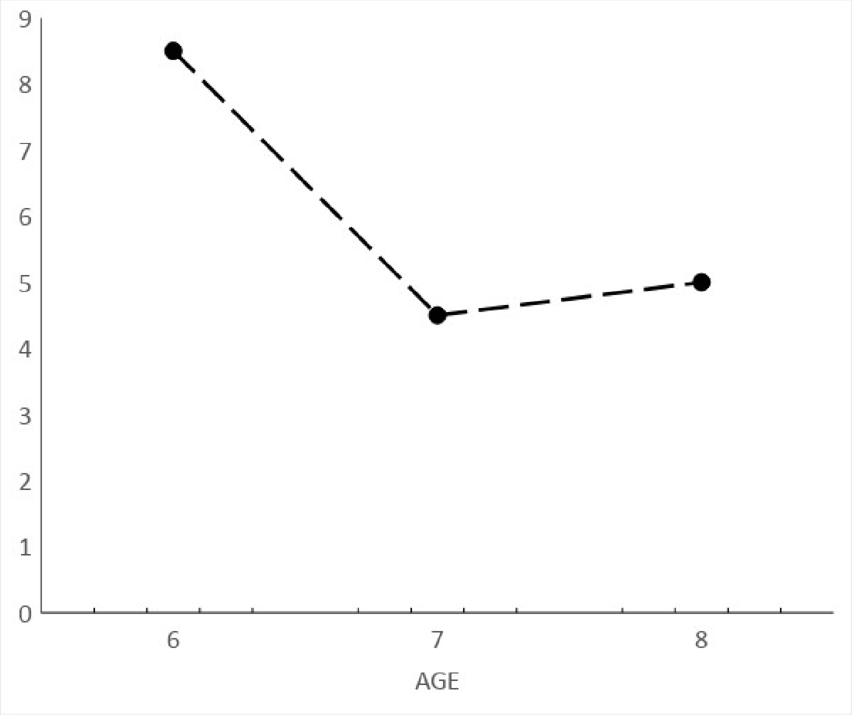

**Figure 7**. Mean number of restarts in 100 utterances as a function of age. Adapted from Verhoven, 1989, p.148.

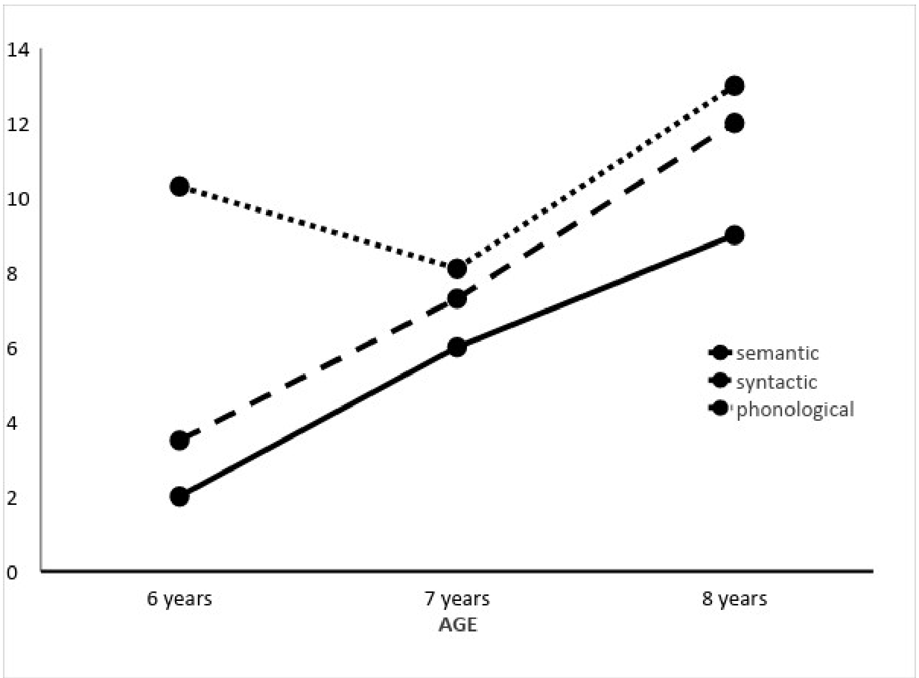

**Figure 8**. Mean number of phonological, syntactic, and semantic corrections in 100 utterances as a function of age. Adapted from Verhoven, 1989, p.149.

### 3.5 Presenting standard deviation

This chapter has previously underscored the dangers of accidentally misrepresenting data via descriptive statistics. Standard deviation (SD) provides a more complete understanding data. A small number of means and SDs can be presented within the text.

> ‘Participants were between 18 and 30 years of age, M = 21.8 years, SD = 2.47 years.’
>
> Smith, Briggs, Pothier & Garcia, 2017, p.6.

When comparing groups of participants to each other, standard deviation can also be essential for demonstrating comparability or difference. Text can still be used for only one or two means and standard deviations to compare between two groups; however, tables are a clear and efficient way to present means and standard deviations for multiple groups and/or measures. Depending on personal preference for format, standard deviation can be presented in a separate column (see Table 7), or next to the mean in parentheses.

**Table 7**
*Means and standard deviations of repairs in 100 utterances as a function of age.*

| | 6-year olds | | 7-year olds | | 8-year olds | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Repairs | 29.00 | 16.36 | 25.73 | 16.40 | 36.22 | 15.10 |

Adapted from Verhoven, 1989, p.147.

Different means and standard deviation can be presented for the same group’s performance on different measures, so readers can compare. Table 8 shows means are across different modalities, different languages; standard deviations demonstrate how spread differs between them.

**Table 8**
*Self-reported ability (scale from 0 to 100 'native' or 'native-like').*

| Language, skill | M | SD |
|---|---|---|
| *Spanish* | | |
| Speaking | 88.36 | 13.05 |
| Listening/understanding | 94.18 | 6.88 |
| Reading | 82.65 | 15.38 |
| Writing | 72.65 | 23.58 |
| *English* | | |
| Speaking | 97.04 | 6.28 |
| Listening/understanding | 98.06 | 3.93 |
| Reading | 97.35 | 4.9 |
| Writing | 96.33 | 5.9 |

Adapted from Smith, Briggs, Pothier & Garcia, 2017, p.11.

There are also statistical graphics that can be used to visually present the spread of the data. A *box and whisker plot*, also called a *box plot* or a *box and whisker diagram*, is a figure that represents the four quartiles of a data set. It consists of a vertical rectangle (the box) that extends from first to third quartile, bisected by a line at the median. Vertical lines (whiskers) extend from the ends of the box to represent data that falls outside of the first and third quartile, representing the minimum and maximum. In some box and whisker plots, extreme outliers are included as individual points. In Figure 9, a box and whisker plot is used to demonstrate differences in Spanish vocabulary scores for three groups of Spanish language students.

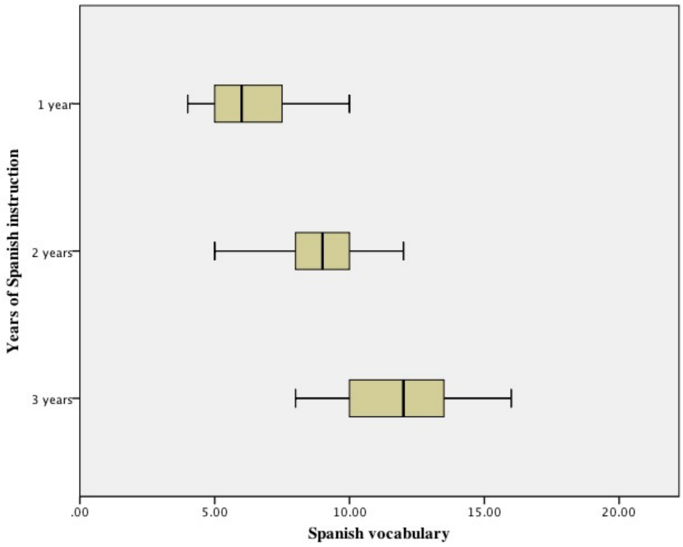

**Figure 9**. Spanish vocabulary score by years of Spanish language instruction.

### 3.6 Other statistical graphics and visual representations

There are more ways to visually represent descriptive data than can be detailed here. *Scatter plots*, for example, display all data points on a plot with variables along the horizontal and vertical axes. Scatter plots are ideal for showing reader the relationship between two variables. Different groups can be represented in the same scatter plot by using a demarcation system to differentiate data points from each group (i.e., colour; shape), as seen in Figure 10. Maps are another visually appealing way to present geographical demographics, regions can be colour-coded to represent demographic data (for an example, see Park, Batalova and Zhong, 2018, p.25). Other options include *Stem and Leaf* plots, and *QQ plots*.

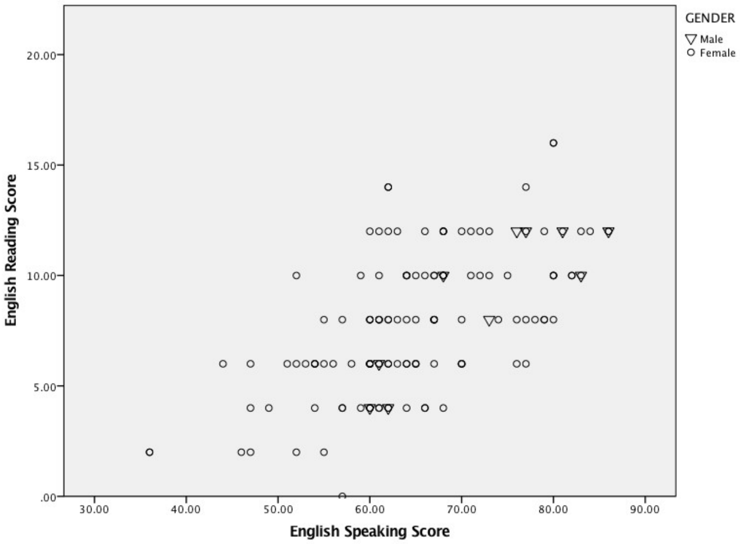

**Figure 10**. The relationships between English reading score and English speaking score for 2 gender groups.

### 3.7 Deciding what to include

There are a number of factors to consider when deciding whether or not particular descriptive results should be presented. First and foremost, descriptive statistics that directly address the research question(s) should be presented in depth. Additionally, one of the roles of descriptive statistics is to allow comparisons across studies. Features such as sample background characteristics, means, normality distributions, etc., often differ from one study to the next, even when two studies have similar designs, procedures, and measures. This is particularly relevant for reconciling contradictory findings to the same or similar research questions in different studies. Descriptive statistics provide foundational information for comparison across studies, and this allows the field to ultimately converge on a general consensus. To compare findings to previous literature, in particular if done in the discussion section, descriptive features of the original data should be presented so that readers can clearly see how the sample was (or was not) comparable to the past studies described. This is also relevant for interpreting one’s own work through a theoretical model; whether the findings do or do not align with a proposed theoretical framework, it is important to share the descriptive statistics relevant to the components of that framework.

Finally, descriptive statistics are foundational for inferential statistical analyses. If planning to conduct inferential analyses (discussed in the following chapter), it is essential to provide the descriptive statistics for the data intended for further analysis. Descriptives confirm that data are suitable for inferential analyses (i.e., normally distributed) or, alternatively, provide justification for other approaches (i.e. non-parametric analyses, conversions for non-normal distributions). Many studies use inferential statistics to compare performance on specific outcome variables between two or more participant groups, often attributed to an independent variable(s). Before examining differences, it is vital to first demonstrate groups comparability by providing descriptive statistics regarding relevant characteristics, including but not limited to: number of participants in each group, age, gender, language background, proficiency level, and education background. For example, imagine a hypothetical study that seeks to compare French vocabulary test outcomes between two groups of learners receiving different instruction methods, Method A and Method B. The study must present descriptives demonstrating that performance is normally distributed and confirm comparison group comparability. If, for example, the Method A group consists of fifteen male participants under the age of 10 with no previous French instruction and Method B group consists of one hundred and fifty adult female participants with intermediate level French, it is immediately apparent that subsequent inferential statistics demonstrating group-level performance differences are irrelevant. Descriptive statistics provide crucial information to confirm that findings from subsequent inferential statistics are meaningful.

## 4. Summary

Emerging researchers sometimes mistakenly view descriptives statistics as “too simple” and consequently overlook essential information in the rush to begin inferential analyses. Descriptive statistics are both independently informative and the backbone of subsequent analyses. Descriptives provide essential information and are needed for: 1) presenting participant characteristics; 2) demonstrating transparency; 3) establishing normality; 4) confirming group comparability; 5) comparing across studies to situate findings within the field; and 6) interpreting theoretical implications. Less is written about descriptive statistics, the metaphorical hard-working older sibling to flashier, attention-grabbing analyses. Though descriptive statistics may not be as frequently used to directly answer research questions, descriptives are the basis on which inferential statistics are assumed to work – to test a hypothesis via inferential statistics, descriptives are needed. To borrow an oft-used phrase in the field, inferential statistics are “necessary but not sufficient”: descriptives are required in order for results found via inferential statistics (e.g., comparing means) to have meaning.

## References

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice*, *20*(1), 50-57.

Bae, J. (2007). Development of English skills need not suffer as a result of immersion: Grades 1 and 2 writing assessment in a Korean/English Two-Way Immersion Program. *Language Learning*, *57*(2), 299-332.

Cho, S., & Reich, G. A. (2008). New immigrants, new challenges: High school social studies teachers and English language learner instruction. *The Social Studies*, *99*(6), 235-242.

Chung-Fat-Yim, A., Himel, C., & Bialystok, E. (2018). The impact of bilingualism on executive function in adolescents. *International Journal of Bilingualism*, DOI: 1367006918781059.

Field, A. P. (2013). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)* (4th Ed.). Los Angeles/London: Sage.

Goriot, C., Broersma, M., McQueen, J. M., Unsworth, S., & Van Hout, R. (2018). Language balance and switching ability in children acquiring English as a second language. *Journal of Experimental child Psychology, 173*, 168-186.

Littlewood, W. (2000). Do Asian Students Really Want To Listen and Obey? *ELT Journal, 54*(1), 31-36.

Park, M., Zong, J., & Batalova, J. (2018). *Growing superdiversity among young US dual language learners and its implications*. Washington, DC: Migration Policy Institute.

Pallant, J. (2016). *SPSS survival manual: a step by step guide to data analysis using IBM SPSS* (6th edition). Maidenhead: McGraw-Hill Education.

Poarch, G. J., Vanhove, J., & Berthele, R. (2018). The effect of bidialectalism on executive function. *International Journal of Bilingualism*, DOI: 1367006918763132.

Reyes, I. (2004). Functions of code switching in schoolchildren's conversations. *Bilingual Research Journal*, *28*(1), 77-98.

Ribot, K. M., & Hoff, E. (2014). “¿Cómo estas?”“I’m good.” Conversational code-switching is related to profiles of expressive and receptive proficiency in Spanish-English bilingual toddlers. *International Journal of Behavioral Development*, *38*(4), 333-341.

Smith, S. A., Briggs, J. G., Pothier, H., & Garcia, J. N. (2017). ‘Mental Workouts for Couch Potatoes’: Executive Function Variation among Spanish–English Bilingual Young Adults. *Applied Linguistics*, DOI: 10.1093/applin/amx038

Smith, S. A., Briggs, J. G., & Pothier, H. (2017). Exploring variation in reading comprehension among young adult Spanish–English bilinguals: The role of environmental language contact and attitudes toward reading. *International Journal of Bilingualism*, DOI: 1367006917690913.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th Ed.). Harlow: Pearson Education.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech, 53*(4), 510-540.

Verhoeven, L. T. (1989). Monitoring in children's second language speech. *Interlanguage Studies Bulletin (Utrecht)*, *5*(2), 141-155.

You, C. J., & Dörnyei, Z. (2016). Language learning motivation in China: Results of a large-scale stratified survey. *Applied Linguistics*, *37*(4), 495-519.

**CHAPTER WORD COUNT: 6,738**