# LESSON 2: DATA COLLECTION METHODS (PART 01)

## 1- Data Collection Methods.

Data collection stands as a pivotal and sensitive endeavour, demanding precision and judicious methodology. It constitutes **a systematic process** aimed at gathering the necessary information to address research questions, solve specific problems, or test hypotheses. Choosing the appropriate data collection method is a decision of paramount importance, as it significantly influences the quality and validity of research outcomes. While data collection typically occurs in the field, it may also extend to libraries, archives, or online sources. Researchers are faced with a myriad of data collection methods, each possessing its unique strengths and weaknesses. Inadequate method selection can yield invalid findings and undermine the overall effectiveness of a study. Therefore, researchers must carefully consider various factors and issues in the multifaceted process of choosing data collection methods to optimize the efficiency and quality of their research projects.

In the field of applied linguistics and foreign language teaching, the significance of data collection cannot be overstated. Just as judges and generals rely on comprehensive data for informed decisions, educators and language researchers depend on data to enhance language learning and teaching strategies. While data collection isn't a novel concept, the modern era has ushered in an unprecedented wealth of data sources and forms. This wealth of data enables researchers and educators to better understand language learners' needs, preferences, and challenges, ultimately leading to more effective language instruction. Data collection in this context encompasses a spectrum of techniques, from traditional surveys and classroom observations to cutting-edge digital tools, all aimed at improving language education.

## 2- Examples from the Field of TEFL.

In the field of TEFL, data collection is pivotal, and various tools and methods aid in gathering valuable insights. Here are four examples highlighting both data collection methods and the specific tools associated with each:

### A- Language Proficiency Assessments:

**Tools Used:** Language proficiency tests like the TOEFL, IELTS, or the Common European Framework of Reference (CEFR) levels.

**Data Collected:** Quantitative data on a learner's language skills such as scores and level scales, encompassing reading, writing, listening, and speaking abilities.

These standardized assessments provide a robust measure of language proficiency, guiding educators and institutions in making informed decisions about students' language capabilities and placement.

### B- Classroom Observations:

**Tools Used:** Observation checklists, video recording equipment, and note-taking during classroom sessions.

**Data Collected:** Qualitative data offering insights into teaching techniques, student engagement, and overall classroom dynamics.

Using tools like checklists and video recordings, educators and researchers gain a firsthand view of instructional practices and learner interactions, facilitating the refinement of teaching strategies.

### C- Language Learning Apps and Online Platforms:

**Tools Used:** Digital language learning applications and platforms, such as Duolingo, Rosetta Stone, or Moodle.

**Data Collected:** User interaction data, encompassing time spent on lessons, responses to exercises, and progression through language modules.

These digital tools employ data analytics to personalize learning experiences, offer customized recommendations, and employ usage patterns to continuously enhance language instruction.

### D- Student Surveys and Questionnaires:

**Tools Used:** Online survey platforms like SurveyMonkey or Google Forms, in addition to paper-based questionnaires.

**Data Collected:** Data on student preferences, motivations, perceived challenges, and feedback regarding language learning experiences.

Through these survey tools, educators and researchers gather feedback from students, gaining insights into learner motivation, areas requiring improvement, and expectations, allowing for the adaptation of teaching approaches accordingly.

## 3- Main Data Collection Methods.

A diverse array of tools and techniques empowers researchers to gather information effectively. Here, we amalgamate insights from different perspectives to provide a general overview.

As a first categorization of these methods, we can rely on the source data as a distinguishing feature:

**A-    Primary Data Collection:** This entails collecting original data directly from the source or through direct interaction with respondents. It is tailored to research objectives and employs various techniques:

1.    **Surveys and Questionnaires:** Structured questionnaires and surveys enable data collection from individuals or groups, conducted through face-to-face interviews, telephone calls, mail, or online platforms.

2.    **Interviews:** Interviews involve direct interactions, whether in person, over the phone, or via video conferencing. They can be structured (predefined questions), semi-structured (flexible), or unstructured (conversational).

3.    **Observations:** Researchers observe and record behaviors, actions, or events in their natural setting, particularly useful for studying human behavior and phenomena without direct intervention.

4.    **Experiments:** Experimental studies involve variable manipulation to observe cause-and-effect relationships, with researchers controlling conditions to collect data.

5.    **Focus Groups:** Small groups engage in discussions moderated by a researcher, facilitating the exploration of opinions, perceptions, and shared experiences.

**B-    Secondary Data Collection:** This method uses existing data collected by others for purposes different from the original intent. Researchers analyze and interpret this data:

1.    **Published Sources:** Books, academic journals, magazines, newspapers, government reports, and other publications containing relevant data serve as valuable sources.

2.    **Online Databases:** Accessible online databases provide an array of secondary data, including research articles, statistical information, economic data, and social surveys.

Government and Institutional Records: Government agencies, research institutions, and organizations often maintain databases or records useful for research.

3.    **Publicly Available Data:** Data shared on public platforms, websites, or social media by individuals, organizations, or communities can be accessed and utilized.

4.    **Past Research Studies:** Previous research studies and their findings can serve as valuable secondary data sources, offering insights or foundations for new research.

From another perspective, we can rely on the nature of data to distinguish two prominent data collection methods, qualitative and quantitative, that facilitate diverse research approaches:

**C-    Qualitative Methods** are concerned with non-numerical data to grasp deeper meaning, context, and subjectivity. They are ideal for exploring social phenomena, human behaviour, and interactions:

1.    **Interviews:** In-depth insights, opinions, and experiences are gathered through individual or group interviews.

2.    **Focus Groups:** Small groups engage in guided discussions, illuminating perspectives and shared experiences.

3.    **Observations:** Behaviours, interactions, and phenomena are observed and recorded in natural or controlled settings.

4.    **Case Studies:** In-depth investigations provide detailed insights into specific subjects, such as individuals, groups, or organizations.

5.    **Ethnography:** Researchers immerse themselves in social groups or cultures, employing participant observation and interviews to understand behaviours, practices, and beliefs.

**D-    Quantitative Methods** focus on collecting numerical data to measure relationships, patterns, and trends:

1.    **Surveys:** Structured questionnaires gather standardized data from a large sample for statistical analysis.

2.    **Experiments:** Variable manipulation in controlled settings observes cause-and-effect relationships and quantifies effects.

3.    **Existing Data Analysis:** Pre-collected numerical data, such as government statistics or organizational records, is analyzed.

4.    **Statistical Analysis:** Statistical techniques like regression analysis, t-tests, or chi-square tests are applied to analyze and interpret numerical data.

# LESSON 2: DATA COLLECTION METHODS (PART 02)

## 1- Descriptive Statistical Analysis of Data Collected from a Questionnaire

Questionnaires have earned their popularity as a prominent data collection tool in the field of TEFL among many other fields. Their widespread use is often attributed to several inherent advantages, making them a preferred choice among researchers and educators. They offer cost-effective data collection and simplicity in data analysis. In the coming section, we will try to understand how data collected from a questionnaire can be statistically processed using SPSS. Before we start, let's recall some information about parametric and non-parametric data.

**Parametric Data:** Parametric data refers to data that follows a specific probability distribution, usually the normal distribution (bell-shaped curve). It assumes that the data has a known and fixed population parameter, such as a mean and standard deviation. Common parametric statistical tests, like t-tests and analysis of variance (ANOVA), are based on these assumptions. Parametric data is characterized by equal intervals between data points and continuous, quantitative measurements.

**Examples:**

_Test Scores:_ When assessing the effectiveness of a new language teaching method, researchers often collect quantitative data, such as scores on language proficiency tests like the TOEFL or IELTS. These scores are typically normally distributed and can be analyzed using parametric tests like t-tests or ANOVA.

_Language Fluency Time:_ In a study measuring how long it takes language learners to achieve fluency, the time (in months) it takes for learners to reach a certain fluency level could be considered parametric data. Researchers can calculate means, variances, and conduct parametric analyses on this data.

**Non-Parametric Data:** Non-parametric data, on the other hand, does not adhere to the assumptions of a specific probability distribution. These data types often include nominal or ordinal data and measurements that do not have equal intervals. Non-parametric statistics are used when data cannot be easily transformed to meet parametric assumptions or when dealing with categorical or rank-ordered data. Non-parametric tests include the Mann-Whitney U test and the Wilcoxon signed-rank test. Non-parametric data may also be called non-parametric or distribution-free data.

**Examples:**

_Language Proficiency Levels:_ Instead of using standardized test scores, language researchers might collect data on language proficiency levels that are ordinal or categorical in nature. For instance, they might categorize language learners as "beginner," "intermediate," or "advanced." Since these categories lack equal intervals, it's non-parametric data. Researchers may use non-parametric tests like the Kruskal-Wallis test to compare groups.

_Student Surveys:_ Surveys or questionnaires with Likert scale questions (e.g., strongly agree, agree, neutral, disagree, strongly disagree) are common in applied linguistics. The responses from these surveys are ordinal data. Non-parametric methods, such as the Mann-Whitney U test, can be applied when comparing responses between groups.

**A- What to do?**

When it comes to analyzing data for Master's research in TEFL, it's important to note that the data often doesn't conform to parametric assumptions. Instead, a more effective strategy involves a <u>combination of descriptive statistics, visual representations, and nonparametric inferential statistics</u>. This approach is particularly advantageous when dealing with questionnaire data, especially when the responses are better characterized as scores rather than precise measurements. While there are ongoing debates in the statistical community about choosing between parametric and nonparametric methods, it is recommended to initiate your analysis with a foundation in nonparametric methods, considering <u>their suitability for the specific data in question</u>. This approach aligns with the nature of language-related research, where nonparametric tests like the Mann-Whitney U test, Wilcoxon signed-rank test, or Kruskal-Wallis test are especially relevant for group comparisons or assessing the significance of score differences, given the ordinal or non-normally distributed nature of language proficiency data.

Below is an example from the questionnaire we will analyze in the coming sections. The three questions yield categorical data that will be coded for the statistical analysis.

**What school are you studying in?**

| | | |
|---|---|---|
| EDS | Education | ☐ |
| HSC | Health and Social Care | ☐ |
| SCI | Science and Mathematics | ☐ |
| SED | Environment and Development | ☐ |
| SLM | Sport and Leisure Management | ☐ |
| CMS | Computing and Management Sciences | ☐ |
| SSL | Social Science and Law | ☐ |
| ENG | Engineering | ☐ |
| SCS | Cultural Studies | ☐ |

**What is your Gender?**

| Male | Female |
|---|---|
| ☐ | ☐ |

**How old are you?**

| 18-24 | 25-30 | 31-40 | 41-50 | 51-60 | 60+ |
|---|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**B- How to?**

When conducting data analysis on your questionnaire responses, it's essential to transfer the collected data into a computer file. To facilitate this process, consider the following guidelines:

**1-** Typically, each research subject or respondent should correspond to a single row within your data sheet.

**2-** Each question posed should generally occupy its own column, representing an individual variable.

**3-** Responses are best stored as numerical values, often using a scale (e.g., 1 to 5 for Likert scales), with the application of "**Value Labels**" to assign meaningful descriptive text labels to these numeric values.

**4-** For ranking or ordering questions, each option or choice should be treated as a separate variable. The same principle applies when respondents are required to select multiple options.

To illustrate these concepts, let's refer to the "Students data23**.sav**" data file as an example. This dataset encompasses 2614 entries from first-year students in the year 2023, with each entry occupying a distinct row in the data sheet. Depending on your preferred data view, you may encounter either <u>text labels</u> or <u>numeric values</u>. You can switch between these views by accessing the "Value Labels" option in the "**View**" menu.
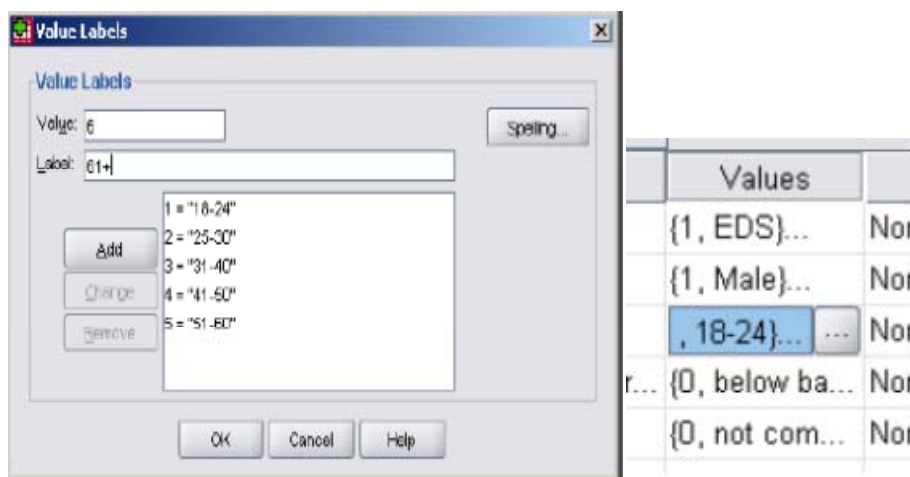


To establish the meanings associated with numeric values, navigate to the "Variable View" tab at the bottom of the screen. Here, by selecting the "Values" column for the relevant variable and using the "Value Labels" dialog box, you can assign specific text labels to each unique value. After entering each pairing, click "**Add**" to incorporate it into the list. It's important to ensure precision in your labels, as they will feature in your graphs and output. Although SPSS does not automatically spell-check your labels, it's wise to keep them concise and error-free.

**C- Now start to look at the data.**

SPSS can accomplish in a few seconds what would take an entire evening to handle manually with questionnaires strewn across your dining room floor. This efficiency allows us to engage with the data, investigating its significance. In our extensive sample of 2614 participants, it's beneficial to perform a fundamental demographic analysis, serving as a preliminary introduction to the subjects underpinning our research findings. In here, you will have preliminary descriptive information about the individuals forming the basis of the results. To compute basic percentages, we can employ the "**Frequencies**" command by navigating through

the "Analyse" menu and selecting "**Descriptive Statistics**," then "Frequencies." As an illustration, I've transferred the "Gender" variable to the variables box. Give it a try and click the OK button.
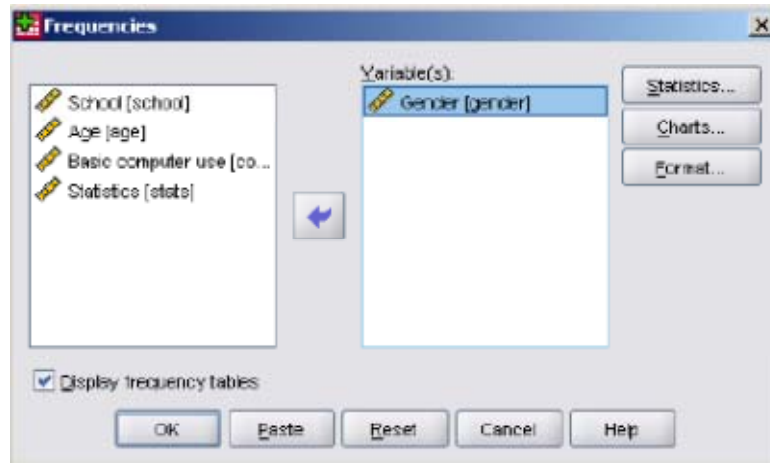


Please note that the OK button's location varies depending on the SPSS version you use, although the functionality remains unaltered. The ensuing output presents the frequency table, indicating that out of the total 2614 respondents, 1244 are female, 1342 are male, and data on gender is missing for 28 individuals. This accounts for the entire pool of 2614 subjects.

The percentage columns, particularly the "Valid Percent," hold our attention as they are computed after accounting for missing values. Although the "Cumulative Percent" is not applicable to this analysis, in cases involving data like an ordinal satisfaction scale, it could provide valuable insights. For instance, it might support statements such as "76% of responders expressed a lack of satisfaction." Sometimes, considering the statements you intend to make about the results can serve as a useful guide for your analysis.
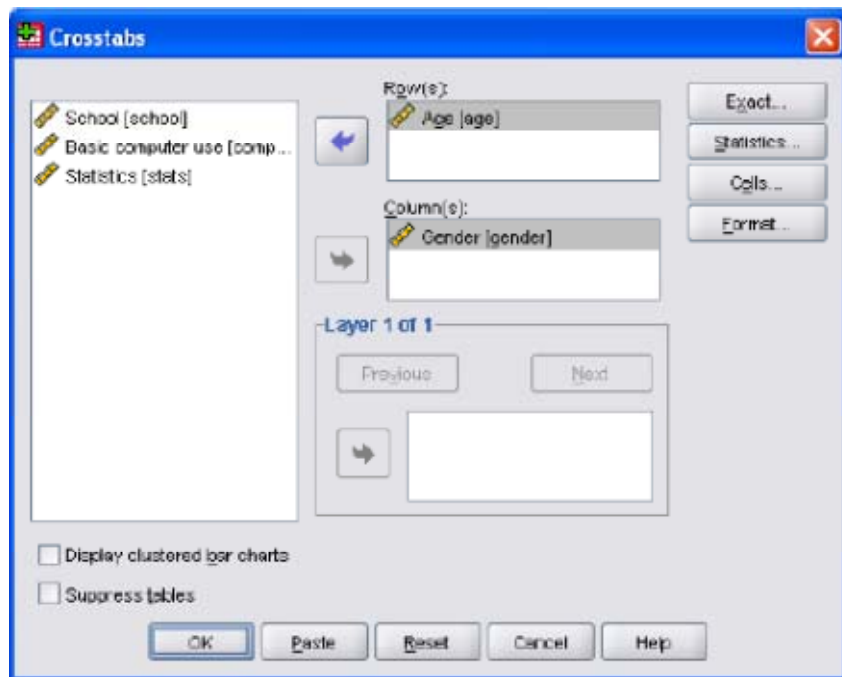
**Gender**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Male | 1342 | 51.3 | 51.9 | 51.9 |
| | Female | 1244 | 47.6 | 48.1 | 100.0 |
| | Total | 2586 | 98.9 | 100.0 | |
| Missing | System | 28 | 1.1 | | |
| Total | | 2614 | 100.0 | | |

**D- Which column serves the purpose?**

The valid percentage column provides valuable statements like "51.9% of those who responded to the question were male." To enhance clarity, consider offering the response level (e.g., 98.9%) or, as a preferred approach, present the results in a table with actual figures in brackets beside the percentages. It's also possible to create a column for the response rate for each question if you prefer. We can also examine the age distribution among our respondents. Take a look at this now. By examining the cumulative percentage column, we observe that over 90% of respondents (91%) are 30 or younger. More significantly, it offers a comprehensive breakdown of the responses.

**Age**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 18-24 | 2161 | 82.7 | 85.1 | 85.1 |
| | 25-30 | 150 | 5.7 | 5.9 | 91.0 |
| | 31-40 | 160 | 6.1 | 6.3 | 97.3 |
| | 41-50 | 62 | 2.4 | 2.4 | 99.8 |
| | 51-60 | 6 | .2 | .2 | 100.0 |
| | Total | 2539 | 97.1 | 100.0 | |
| Missing | System | 75 | 2.9 | | |
| Total | | 2614 | 100.0 | | |

When we analyze two variables simultaneously, like comparing age profiles within different genders, crosstabulations prove to be quite useful. To create a crosstabulation, navigate to the "**Crosstabs**" option in the **Analyze** menu under **Descriptive Statistics**. Place the "Age" variable in the rows box and "Gender" in the columns. The output displays the number of people in each age group, with columns for each gender as well as a total column. This straightforward cross-tabulation allows us to discern gender and age group relationships. For a more insightful comparison, you can request SPSS to calculate the percentage of each gender in each age group.



**Age * Gender Crosstabulation**

Count

| | | Gender | | Total |
|---|---|---|---|---|
| | | Male | Female | |
| Age | 18-24 | 1159 | 989 | 2148 |
| | 25-30 | 73 | 75 | 148 |
| | 31-40 | 60 | 98 | 158 |
| | 41-50 | 15 | 46 | 61 |
| | 51-60 | 2 | 3 | 5 |
| Total | | 1309 | 1211 | 2520 |

Return to the Crosstabulation dialog box (Analyze, Descriptive Stats, Crosstabs) and click the "**Cells**" button. Add column percentages. The resulting table may appear more complex as it

includes both the raw number of respondents in each combination of gender and age group. If you wish to display percentages exclusively, you can deactivate the "**Observed Counts**" option in the cells dialog.
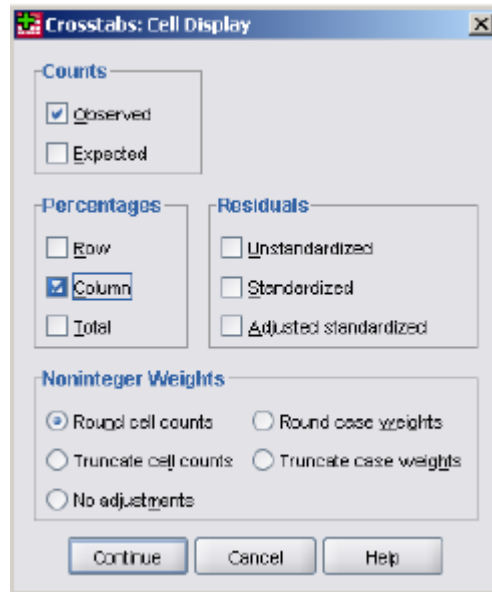


In a results section you wouldn't simply copy and paste the output tables into the document, you might create a table including the output but in a more readable format, for example;

| Crosstabulation of Age and Gender showing percentages within each gender. | | |
|---|---|---|
| Age | Gender | |
| | Male | Female |
| 18-24 | 1159 (88.50%) | 989 (81.70%) |
| 25-30 | 73 (5.60%) | 75 (6.20%) |
| 31-40 | 60 (4.60%) | 98 (8.10%) |
| 41-50 | 15 (1.10%) | 46 (3.80%) |
| 51-60 | 2 (0.20%) | 3 (0.20%) |

Ensure that your table's title effectively conveys its intended purpose. In this specific instance, the data reveals that a higher proportion of females over the age of 30 are enrolling as students. It's important to note that we should expect some variability in the percentages due to random chance. To assess the likelihood of encountering such a disparity in percentages by chance, we can turn to inferential statistics. In this context, we will employ **the Chi-square** statistic for this purpose. The coming section will take you into the inferential statistics procedure of Chi-square test.