A Step-by-Step Research Process

Dr. Karanja Mutitu

Learning outcomes:

**By the end of this unit, the participant should be able to:**

i. **Describe an educational research process from a social research perspective**

ii. **Appreciate the importance of rigor in research so as to enhance the quality of research output**

CEMASTEA RESEARCH PROCESS

The Director research and administration through the CEO will constitute a team that:

1) Identify an area of research relevant to CEMASTEA's core mandate and in line with the quality objective on research.

2) Plan and design a research project from the identified area within 4 weeks of receiving it

3) Initiate research proposal writing process
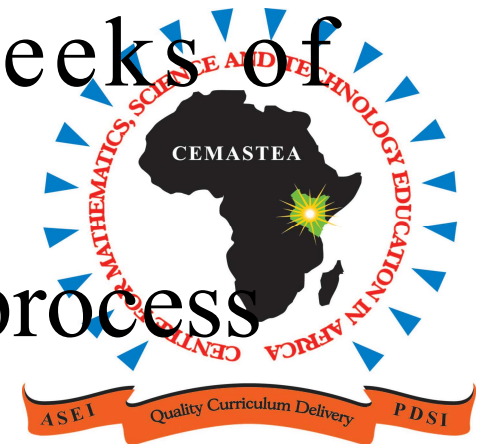
CEMASTEA RESEARCH PROCESS

4) On the date approved by the Head Research and Administration, share the research proposal with CEMASTEA staff for feedback and validation.

5) Share the validated research proposal and funding proposal with CEO CEMASTEA for approval within two weeks of receiving it.

CEMASTEA RESEARCH PROCESS

6) The Head Research and Administration will send an Application through the office of the CEO to NACOSTI for research Permit in accordance with the Science Technology and Innovation (Research Licensing) Regulation 2014.

7) The selected research team will carry out the pilot study for research tools validation

CEMASTEA RESEARCH PROCESS

8) The research team will conduct the research as per the approved research design and plan.

9) The research team will compile the research report within two months of carrying out the data collection. On the date approved by the CEO, the Head Research and Administration ensures that the research findings are presented to CEMASTEA staff.

CEMASTEA RESEARCH PROCESS

10) The Head of Research and Administration receives a soft-copy of the revised research report from the research team, submits it to the CEO for approval to circulate it within one month of presenting the research findings to CEMASTEA staff.

CEMASTEA RESEARCH PROCESS

11) The Head Research and Administration ensures that the research findings are published and disseminated to stakeholders using platforms such as CEMASTEA's website, workshop and conference presentations, policy briefs, practice briefs, and journal publications.

12) The Head of Research and Administration isolates and shares areas for consideration in TPD programs with the Head Training

CEMASTEA RESEARCH PROCESS

13) The Head of research and administration keeps a copy of the report in the repository of reports and, alerts staff that they are free to write papers for presentation at conferences and publication using data and information from the report.

Purposes of Research

- **Exploration**
    - gaining some familiarity with a topic, discovering some of its main dimensions, and possibly planning more structured research

- **Description**
    - Gives details of the characteristics of the phenomenon studied; more of what than why of the study.
    - Census report on number of citizens/population
    - Political poll predicting who will win an election

- **Explanation**
    - the overall strategy that choosen to integrate the different components of the study in a coherent and logical way to take it a step further

Research

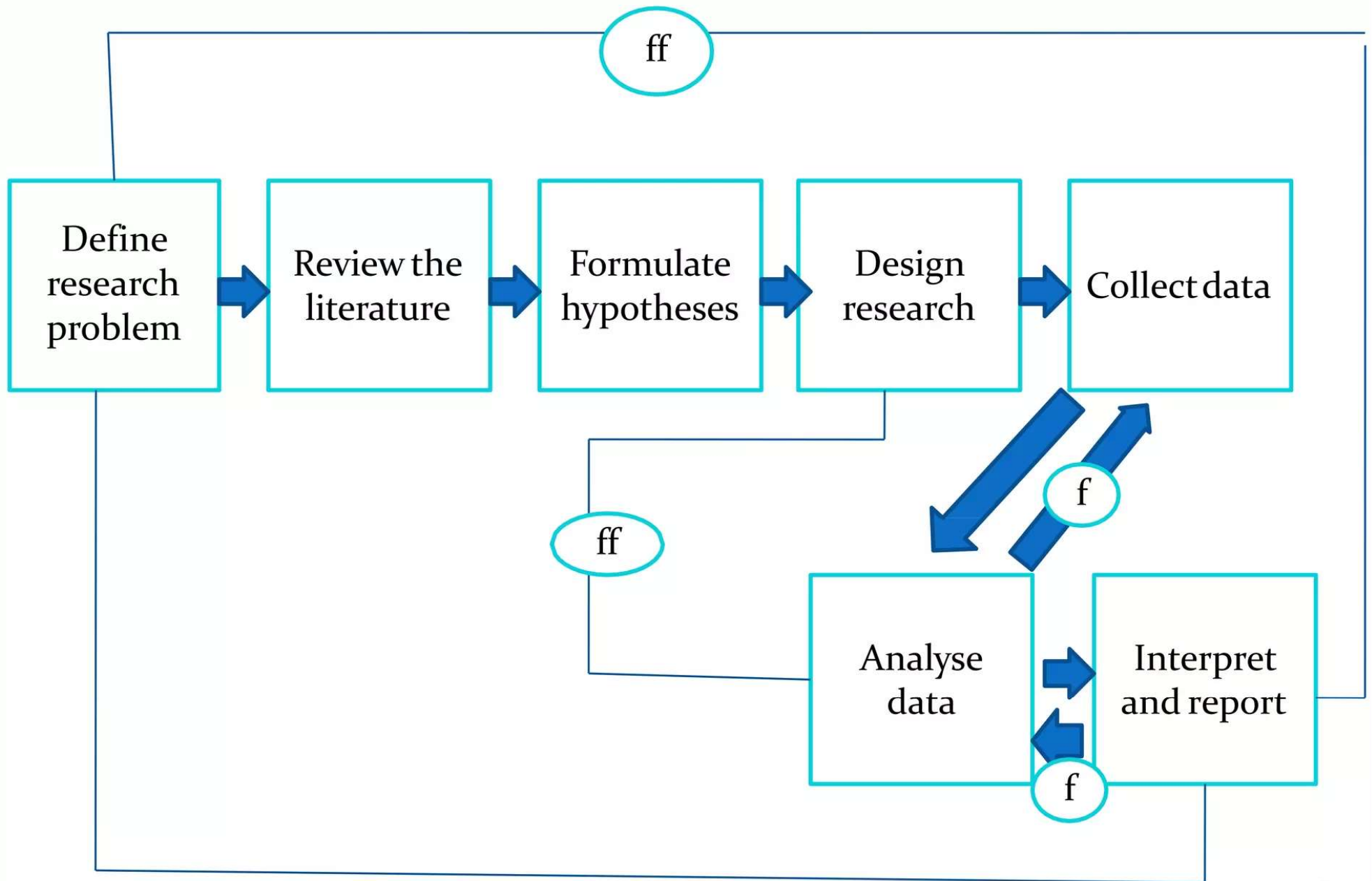**Before embarking on a research one must:**

- Clearly recognize/identify the Problem (through Observation, Decision dilemma, academic/daily/reading experience, field situations, consultations/brainstorming, research recommendations, intuitions)

- Determine information already available and what further information is required (gap), as well as the best approach for obtaining it

- Obtain and assess information objectively to help inform the decision

'Seven' Phases of Research

1. Define research problem

2. Literature review

3. Formulate hypothesis/specific research questions

4. Selection of research design, subjects, and data collection techniques

5. Data gathering/collection

6. Data processing and analysis

7. Interpretation and report writing (Implications, Conclusions, and Recommendations)

```
                              ┌──────────────┐
                              │      ff      │
                              └──────────────┘
```

| Define research problem | → | Review the literature | → | Formulate hypotheses | → | Design research | → | Collect data |

ff

f

| Analyse data | → | Interpret and report |

f

Where   f = feed back(helps in controlling the subsystem
        ff= feed forward(serves the vital function of providing criteria forevaluation

Operational Definitions

- Variables first defined by conceptual definitions that explain the concept the variable is trying to capture

- Variables then defined by operational definitions which are definitions for how variable will be measured

Unit of Analysis

- Major entity you are analyzing in your study

- It is the type of object that makes up each data point

    – Individuals
    – Artifacts (books, photos, newspapers)
    – Geographical units
    – Social interactions

Unit of Analysis Error

- In some studies people are allocated in groups, rather than individually. When this is done, the <u>unit of allocation</u> is different from the unit of analysis (usually).

    – This is sometimes called a unit of analysis error.

    – It can result in studies having narrower confidence intervals and receiving more weight than is appropriate.

Independent and Dependent Variables

- independent variable is what is manipulated

- a treatment or program or cause

- 'Factor'

- dependent variable is what is affected by the independent variable

- effects or outcomes

- 'Measure'

Problem Definition

A research problem is a question/statement/problem that the researcher wants to answer/solve. A problem is an interrogative statement or sentence that asks what relationship exists between two or or more variables

- Describe broader context (background)

- State the objectives or purposes

- Inform reader about the scope of the study, including defining any terms, limitations, or restrictions or scope

- State the hypothesis (es)/research questions
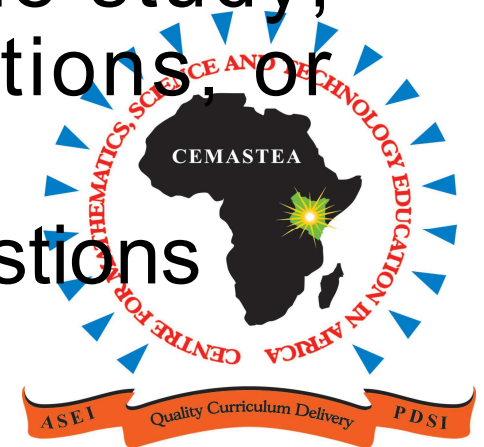
Literature Review

- Gives theoretical rationale of problem being studied, what research has been done and how it relates to the problem identified

  - Helpful to divide the literature into sub-topics for ease of reading

  - Quality of literature should be assessed

  - Be sure to include well respected 'researchers' in the research area (if they exist), recent publications (usually not more than 5 years older than the date of your study)

  - Paraphrase or quote verbatim. Ensure correct referencing is done
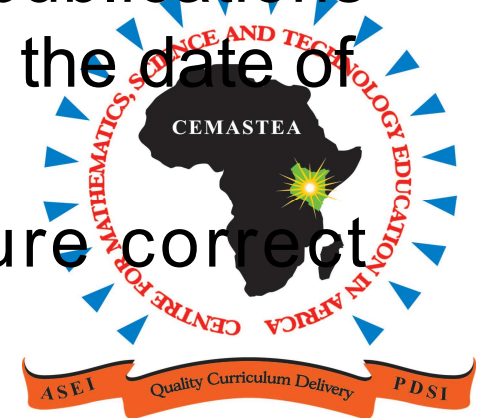
Selection of Research Design

- The research design indicates the steps that will need to be taken and the sequence they will occur

- Each design can rely on one or more data collection techniques

- Assess reliability and validity

- Critical consideration in determining methodology in the selection of subjects

Research Design and Methodology

- In general, a **research design** is like a blueprint for the research.

- **Research Methodology** concerns how the design is implemented, how the research is carried out and which tools are used

A few designs

- Cross-Sectional Design

- Longitudinal Design

- Time Series Design

- Panel Design

Cross-Sectional Design

- A cross-sectional design is used for research that collects data on relevant variables one time only from a variety of people, subjects, or phenomena.

- A cross-sectional designs provides a snapshot of the variables included in the study, at one particular point in time.

- Cross-sectional designs generally use survey techniques to gather data, for example, Availability of resource, Census. eg. Study on Coding

- Advantages: data on many variables, data from a large number of subjects, data from dispersed subjects, data on attitudes and behaviors, good for exploratory research, generates hypotheses for future research, data useful to many different researchers

- Disadvantages: increased chances of error, increased cost with more subjects and each location, cannot measure change, cannot establish cause and effect, no control of independent variable, difficult to rule out rival hypotheses, static

Longitudinal Designs

- A longitudinal design collects data over long periods of time.

- Measurements are taken on each variable over two or more distinct time periods.

- This allows the researcher to measure change in variables over time.

- In a longitudinal analysis you actually have the measurement of change either at the individual level or a group.

Time Series Design

- A **Time Series Design** collects data on the same variable at regular intervals in the form of aggregate measures of a population.
- Time series we can measure the overall change in the measurement over time (or by group) - STEM study

- Time series designs are useful for:
  - establishing a baseline measure
  - describing changes over time
  - keeping track of trends
  - forecasting future (short term) trends

- **Advantages:** data easy to collect, easy to present in graphs, easy to interpret, can forecast short term trends

- **Disadvantages:** data collection method may change over time, difficult to show more than one variable at a time, needs qualitative research to explain fluctuations, assumes present trends will continue unchanged

Panel Designs

- **Panel Designs** collect repeated measurements from the same people or subjects over time.

- Panel studies reveal changes at the individual level.

- Panel data are a type of longitudinal data where the observed units are the same eg. Tracer Study

- Advantages: reveals individual level changes, establishes time order of variables, can show how relationships emerge

- Disadvantages: difficult to obtain initial sample of subjects, difficult to keep the same subjects over time, repeated measures may influence subjects behavior

Data Collection

- Obtain ethical approval

- Must pretest/pilot the tools for validity and reliability before main data collection

- Design the sampling scheme

- Tools should be coded

- After pilot restest the data using dummy tables/graphs

Language of Sampling

- **Population**: entire collection of people/things

- **Parameter/statistic**: number that results from measuring all units in population

- **Sampling frame**: specific data from which sample is drawn

- **Unit of analysis**: type of object of interest

- **Sample**: a subset of some of the units in the population

Data processing and analysis

- Data preparation and cleaning

- Justify the statistical technique to be adopted

- Describe demographics of the data

- Carry out quantitative then qualitative data analysis (mixed method analysis) - if applicable

- Compare behavior (if applicable)

- Choose an appropriate statistical technique (if applicable)

- Look for patterns in data (if applicable)

Statistics

1. **Descriptive statistics**
   a. Percentages, means, ratios and proportions
   b. Reduce massive amount of data into manageable sizes, summarization, organisation and description of data
   c. Enable visualisation of trends

2. **Inferential statistics**
   a. Correlation, Regression, Chi-Square, T-Tests, ANOVA
   b. Allow generalizations, conclusions, estimates, decisions, predictions and inferring characteristics of the population from the sample characteristics

Statistics

1. **Descriptive Analysis**
   a. Means, Frequency, Percentages
2. **Data Quality Checks:** Tests of Normality, Reliability of scales, Factor Analysis
3. **Assess Relationships:** Correlation, Regression, Chi-Square
4. **Group Comparisons**
   a. T-Tests: Independent, Paired samples, One Sample
   b. ANOVA: One Way ANOVA, ANCOVA, MANOVA
4. **Trend Analysis**
   a. **Repeated Measures ANOVA**
   b. **Times Series**

Interpreting the Results

- Indicate the findings inline with the research problem and with specifics to the specific research objectives

- Interpretation should be both direct and indirect relationships; other studies findings, similarity and differences and explanations

- Make sure to consider the audience of your findings

- Discuss implications for the population of interest and future research

3 KEY STEPS TO DATA CLEANING

Step 1. Safely store your data:

Original raw data should be stored safely with a reliable backup; both locally and in a secure online storage.

Step 2: Tidy your Dataset

Tidy datasets are easy to manipulate, model and visualise, and have a specific structure:

i. Do all your columns have the right data format, e.g. numbers as numerical and not as text?

ii. Does every column have a unique, correct and simple header? For example: "Age" instead of "What is your age?"

iii. Have you cross-checked the character encoding? When you import data from CSV for example, the character encoding can be wrong which can result in special characters not being displayed correctly, e.g. "%"appearing as "&". UTF-8 encoding does well with most languages

Step 3: Cleaning your data

Clean your data for data quality

Check for inconsistencies and inaccuracies in your data to ensure - accuracy, consistency, completeness, uniqueness, validity and precision (by filtering outliers).

Each of these elements should be checked for each and every variable as well as a combination of variables.

i) Accuracy

How close is the value of the data to the true value? How accurately does the value of the data describe the object or event being described? There are two types of accuracy.

- **Syntactic accuracy**

In this case, the value might be correct, but it doesn't belong to the correct domain of the variable. eg. A negative value for teaching experience or age or a percentage higher than 100

- **Semantic accuracy**

In this case, the value is in the correct domain, but it is not correct. eg. The attribute gender is given the value "female" for a person called John Wambui.
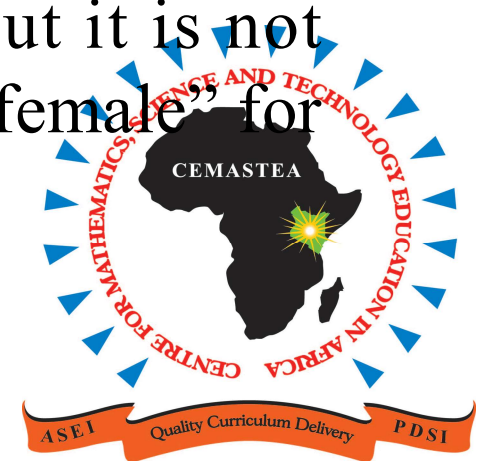
ii) Consistency

- Do all the values of one variable represent the same definition? For example: Age recorded in the same unit (months, years) throughout the dataset.

- This enables you to fix any measurement errors

iii) Completeness

- *Variable values*: Are there values missing for certain variables? Remove the observation or input the missing data?

- *Records*: Is the dataset complete for the analysis at hand? eg. you set out to study 10 schools but only have 8 completed.

The respondent might have refused or forgotten to answer a question in a questionnaire, or a variable might not be applicable to a certain object or the researcher may have forgotten to enter the data

iv) Uniqueness

- Is there any duplicate data? Checking the uniqueness in a dataset consists of identifying and correcting duplicated observations (mostly rows). eg. A school/respondent entered twice.

v) Validity

- Does the data conform to the defined data type rules? eg. age should be numbers and not letters $1 \neq I$, $0 \neq O$ letter are therefore invalid.

- Numbers are the most common data type that you will need to convert when cleaning your data.

- Text, is classified as a string and your analysis algorithms cannot perform mathematical operations on them.

- The same is true for dates that are stored as text.
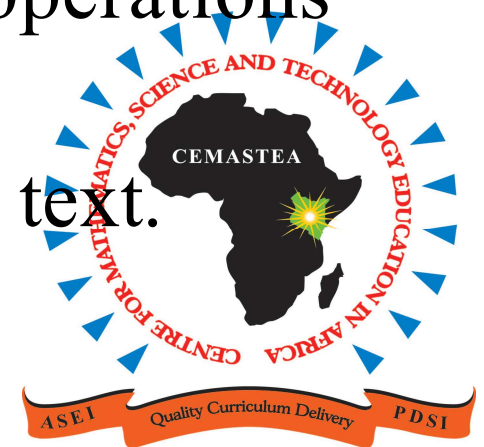
vi) Outliers

An outlier is a data point that differs significantly from other observations.

Here, you can see an example of an outlier. While most schools have learners of between 20 and 50, school A has been entered 500. First, you should check with the data collector. Do they really have this number or was it a typo? If this isn't possible, you'll have to make the decision yourself:

If you are sure it is a typo, modify the value to 50. If NOT sure, leave the value (there might be a school that actually has these learners) or maybe use mode to avoid getting a skewed view of the data.

It can be time consuming to check for outliers manually, especially when there are many observations. A simple trick to speed this process up is to visualise the data by sorting/filtering.
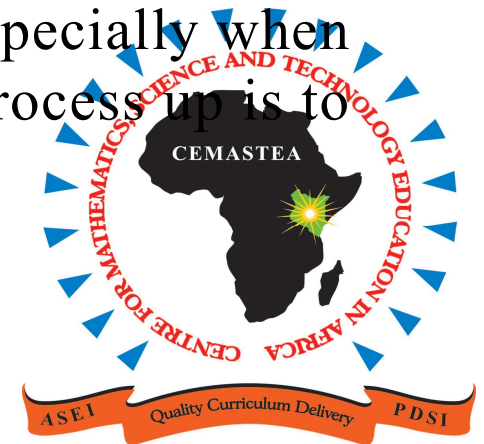
vii) Clear Formatting

- Clear formats, before and after data scrubbing. Heavily formatted data from a range of sources/data with different document formats can make your data analysis confusing and incorrect.

- Varied format removal is normally not a difficult process, both Excel and Google sheets, have a simple standardization function to do this.

Minimising data cleaning

Data cleaning is time consuming. You can significantly minimise data cleaning by:

i) designing your study with data quality in mind such as by limiting the possible answers to a certain question without influencing the actual response. eg. instead of asking the actual years of teaching experience we asked the teaching experience range 0 - 5 yrs, 6 - 10 yrs to avoid outliers or invalid data entry.

However, data cleaning process is worthy some reasonable time to provide correct insights for making sound decisions.