

WORKSHOP SERIES | DAY ONE

Getting Started with SPSS

Variables, Data Entry, Computation and Descriptive Statistics

Facilitator: Assoc. Prof. Enoch Oladunmoye

College of Humanities and Social Sciences

Kampala International University

Email: enoch.oladunmoye@kiu.ac.ug

What we shall cover today

1

Understanding SPSS

The interface, the two views and what each one is for

2

Defining variables

What a variable is, its types by role and by level of measurement

3

The Variable View

All eleven elements and how to complete them correctly

4

The Data View

Entering data case by case without corrupting the file

5

Computing variables

Turning items into construct scores using Transform

6

Missing data and descriptives

Detecting, coding and handling gaps, then summarising

What exactly is SPSS?

SPSS (Statistical Package for the Social Sciences) is a menu driven statistical software used to enter, manage, transform and analyse quantitative data. It is popular in the social and health sciences because almost everything can be done through dialogue boxes rather than programming, although it also has a syntax language underneath.

- 1 Data management** Enter, label, recode, merge, split and clean your dataset
- 2 Transformation** Compute new variables, reverse score items, create scale totals
- 3 Analysis** From frequencies and means to regression, factor analysis and beyond
- 4 Output** Tables and charts that can be exported to Word for your report

Figure 1

The SPSS opening window

Show participants live:

- The title bar and menu bar
- The toolbar icons
- The data grid occupying the centre
- The two tabs at the bottom left: Data View and Variable View
- The separate Output Viewer window that appears after any analysis

SPSS has two faces: Variable View and Data View

Variable View

This is where you DESCRIBE your data

- Each row is one variable
- Each column is a property of that variable
- You tell SPSS the name, the label, the codes and the level of measurement
- Nothing is analysed here

Data View

This is where you HOLD your data

- Each row is one case, that is one respondent
- Each column is one variable already defined
- The cells contain the actual responses
- This is the spreadsheet you type into

Rule of thumb: you always finish Variable View before you touch Data View.

What is a variable?

A variable is any characteristic, attribute or condition that can take on different values across persons, objects, times or settings. If it does not vary, it is a constant, not a variable.



Age Varies from one student to another



Sex Takes two or more categories



Anxiety score Varies in degree along a continuum



Faculty Takes several named categories



Monthly income Varies in amount with a true zero



Academic performance Varies from very poor to excellent

Types of variable by the role they play in a study



Independent variable

The presumed cause. It is manipulated or observed as the predictor.

Example: Teaching method, hours of study



Dependent variable

The presumed effect. It is the outcome you measure.

Example: Academic performance, anxiety



Mediating variable

It carries the effect of the IV to the DV. It explains HOW or WHY.

Example: Motivation between method and performance



Moderating variable

It changes the strength or direction of the IV to DV link. It explains WHEN or FOR WHOM.

Example: Sex, age, socio economic status



Control variable

Held constant or statistically adjusted so it cannot distort the result.

Example: Prior grade point average



Extraneous variable

An uncontrolled outside influence that may confound findings.

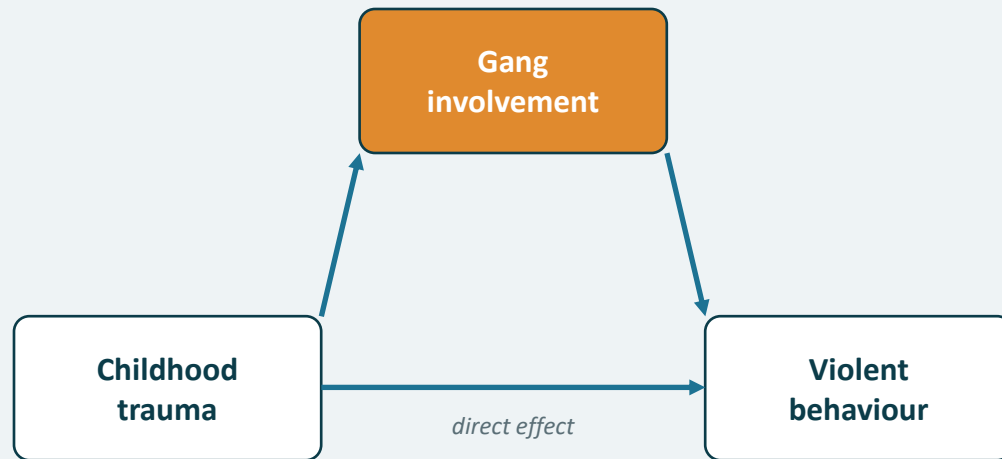
Example: Noise during testing, fatigue

FIGURE 2

Seeing the difference: mediation versus moderation

Mediation

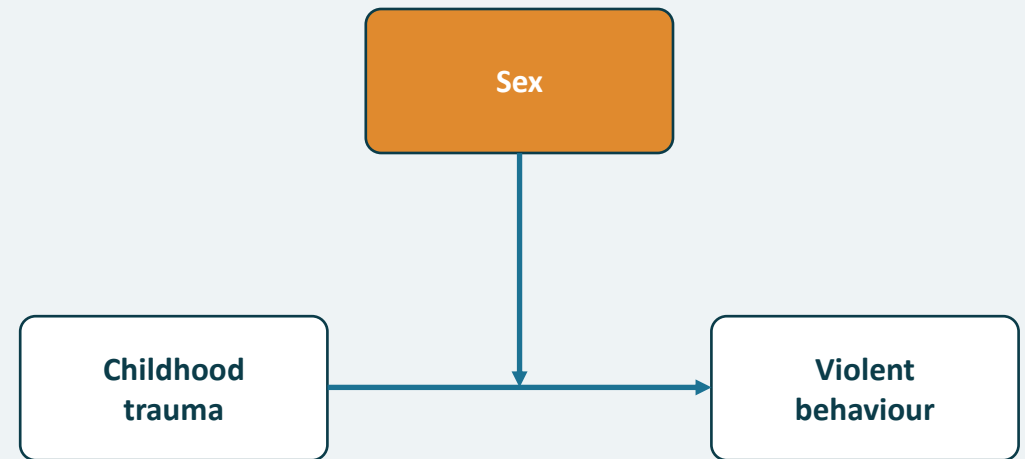
The IV affects the mediator, which in turn affects the DV.



Answers: HOW or WHY does the effect occur?

Moderation

The moderator alters the strength of the IV to DV link.



Answers: WHEN or FOR WHOM does the effect hold?

Figure 2. Schematic contrast between a mediation path and a moderation path.

Types of variable by level of measurement



Continuous variables

Take any value within a range, including decimals. Arithmetic makes sense.

- Age in completed years
- Anxiety score out of 60
- Monthly income in shillings
- Reaction time in seconds

Also called quantitative, numeric or scale variables



Categorical variables

Place a case into a group. The numbers are labels, not quantities.

- Sex: male, female
- Marital status: single, married, divorced
- Faculty: CHUSS, education, law
- Yes or no responses

Also called qualitative or grouping variables

Categorical variables split further into nominal (unordered) and ordinal (ordered). Continuous variables split into interval and ratio.

TABLE 1

The four scales of measurement

Scale	What it does	Key property	Example	Permissible statistics
Nominal	Names or labels categories with no order	Difference only	Sex, religion, faculty, tribe	Frequency, mode, chi square
Ordinal	Ranks categories in order but the gaps are unequal	Difference and order	Class of degree, Likert agreement, rank in class	Median, percentiles, Spearman rho
Interval	Equal intervals between points but no true zero	Order and equal distance	Temperature in Celsius, IQ score, many test scores	Mean, standard deviation, Pearson r, t test
Ratio	Equal intervals with a meaningful absolute zero	Order, equal distance and true zero	Age, income, weight, number of children	All of the above plus ratios and coefficient of variation

Remember the ladder: nominal is the weakest and ratio is the strongest. As you climb, you keep every property of the level below and gain one more. The higher the scale, the more powerful the statistics you may legitimately use.

TABLE 2

How the four scales map onto the SPSS Measure column

Your measurement scale	What you select in SPSS	SPSS icon	Typical variables in your study
Nominal	Nominal	Three coloured circles	Sex, faculty, marital status, religion, yes or no items
Ordinal	Ordinal	An ascending bar staircase	Level of study, class of degree, single Likert items
Interval	Scale	A ruler	Total anxiety score, mean self esteem score, standardised test score
Ratio	Scale	A ruler	Age in years, income, number of children, hours of study

Why this matters

SPSS uses the Measure setting to decide which variables it offers you inside certain dialogue boxes and which charts it draws by default. A wrongly set Measure will quietly limit or mislead your analysis.

The practical shortcut

SPSS collapses interval and ratio into one option called Scale. So you have only three choices to make: Nominal, Ordinal or Scale.

THE GOLDEN RULE

**You must tell SPSS about your variables before
you enter a single response.**

SPSS is not Excel. It does not guess. An undefined variable becomes VAR00001 with the wrong type, no label, no codes and the wrong level of measurement. Everything you build on it afterwards inherits that error.

The eleven elements of the Variable View

1 Name

The short internal name SPSS uses

2 Type

Numeric, string, date and so on

3 Width

Total characters or digits allowed

4 Decimals

Decimal places displayed

5 Label

The full descriptive name shown in output

6 Values

The code for each category, such as 1 = male

7 Missing

Codes reserved for missing responses

8 Columns

Display width of the column in Data View

9 Align

Left, right or centre alignment

10 Measure

Nominal, ordinal or scale

11 Role

Input, target, both, none, partition, split

FIGURE 3

What a properly completed Variable View looks like

Name	Type	Width	Decimals	Label	Values	Missing	Measure	Role
id	Numeric	4	0	Respondent identity number	None	None	Nominal	Input
sex	Numeric	1	0	Sex of respondent	{1, Male}...	9	Nominal	Input
age	Numeric	3	0	Age in completed years	None	999	Scale	Input
faculty	Numeric	1	0	Faculty of respondent	{1, CHUSS}...	9	Nominal	Input
level	Numeric	1	0	Level of study	{1, Year one}...	9	Ordinal	Input
se1	Numeric	1	0	Self esteem item 1	{1, Strongly disagree}...	9	Ordinal	Input
se2	Numeric	1	0	Self esteem item 2	{1, Strongly disagree}...	9	Ordinal	Input
selfesteem	Numeric	8	2	Total self esteem score	None	None	Scale	Target

Figure 3. Extract of a completed Variable View. Note that every variable carries a full label and every categorical variable carries value codes.

Notice: the last row, selfesteem, is a computed variable. It does not come from the questionnaire, it is created later using Transform then Compute Variable.

Elements one to four: Name, Type, Width, Decimals

1 Name

- Must begin with a letter, not a number or symbol
- No spaces: use sex1, not sex 1
- Avoid reserved words such as ALL, AND, BY, NOT, OR, TO, WITH
- Keep it short and systematic: se1, se2, se3
- SPSS treats Age and age as the same name

2 Type

- Numeric is the default and the one you will use most
- String is for text such as open ended comments
- Date is for dates of birth or dates of interview
- Never store codeable answers as string: SPSS cannot compute with text

3 Width

- The total number of characters or digits SPSS will store
- Default is eight, which is enough for most social science data
- Increase it only for long identity numbers or long strings

4 Decimals

- The number of decimal places displayed
- Set to zero for counts, codes, age and Likert items
- Set to two for computed means and scale scores
- It affects display only, not the stored precision

Elements five to seven: Label, Values, Missing

Label

- A full descriptive name in ordinary English
- It is what appears in every output table
- Write it as you want it to read in your report
- Example: Perceived academic stress among finalists

Values

- Where you attach a meaning to each numeric code
- Click the cell, then the small blue box on the right
- Type the value, type the label, click Add, repeat, click OK
- Example: 1 = Male, 2 = Female, 9 = No response

Missing

- Codes you reserve to mean a response is absent
- Discrete missing values: enter up to three codes
- Range plus one optional discrete value for scale variables
- Choose codes that cannot occur naturally: 9, 99, 999

Why value labels are not optional

Without them, your frequency table reports 1 and 2 instead of Male and Female. You will then have to interpret from memory, and readers of your report cannot verify anything. Enter labels once and every table from that point onward is self explanatory.

Elements eight to eleven: Columns, Align, Measure, Role

Element	What it controls	What you should normally do
Columns	The visible width of the column in the Data View grid	Leave at the default. Widen only if the label is being cut off
Align	Whether the value sits left, centre or right within the cell	Leave at the default. It is cosmetic and has no analytical effect
Measure	The level of measurement: Nominal, Ordinal or Scale	Set this deliberately for every variable. It shapes what SPSS offers you
Role	How the variable is used in dialogues that support roles: Input, Target, Both, None, Partition, Split	Leave as Input for predictors, set Target for outcomes. Optional in most classical procedures

The takeaway

Of these last four, only Measure carries real analytical weight. Columns and Align are display settings. Role is a convenience that saves you dragging variables in newer dialogue boxes. Get Measure right and move on.

Practice one: define this codebook in Variable View

Variable	Label	Codes to enter under Values	Measure
id	Respondent identity number	None	Nominal
sex	Sex of respondent	1 = Male, 2 = Female, 9 = No response	Nominal
age	Age in completed years	None. Use 999 as the missing code	Scale
faculty	Faculty of respondent	1 = CHUSS, 2 = Education, 3 = Law, 4 = Business, 9 = No response	Nominal
level	Level of study	1 = Year one, 2 = Year two, 3 = Year three, 9 = No response	Ordinal
se1 to se10	Self esteem item 1 to item 10	1 = Strongly disagree to 5 = Strongly agree, 9 = No response	Ordinal

Your task: create all fourteen variables, attach every label, enter every value code, set every missing value and set every Measure. Save the file as workshop_practice.sav before you go any further.

Moving to the Data View: entering the responses

Once every variable is defined, click the Data View tab. Each column is now one of your variables and each row is one respondent. You simply type the coded response into the cell where a case meets a variable.

	id	sex	age	faculty	level	se1	se2	se3	se4	se5
1	001	1	21	1	2	4	5	3	4	5
2	002	2	23	1	3	5	4	4	5	4
3	003	2	20	2	1	3	3	2	3	9
4	004	1	999	3	2	4	4	.	4	3
5	005	2	22	1	3	5	5	5	4	4

Figure 4. Extract of the Data View. Row 3 shows a user missing code of 9 on se5. Row 4 shows a system missing cell, displayed as a full stop, on se3, and a user missing code of 999 on age.

One row, one respondent

Never split a single respondent across two rows

Enter codes, not words

Type 1, not Male. The label does the explaining

Move with Tab or Enter

Tab moves across a case, Enter moves down a variable

Save constantly

Press Ctrl and S after every ten cases

Data entry discipline: what to do and what to avoid

Do this

- Give every questionnaire a serial number before you start typing
- Enter data in the same order as the questionnaires are stacked
- Enter the whole of one case before moving to the next
- Use a single consistent missing code across the whole file
- Run frequencies on every variable immediately after entry to catch impossible values
- Keep the raw questionnaires until analysis is complete

Avoid this

- Typing text where a code belongs
- Leaving a cell blank when the respondent actually refused: code it instead
- Using different missing codes for the same kind of gap
- Inventing a value because the handwriting is unclear
- Deleting a row to fix a mistake, which shifts every case below it
- Editing the file without keeping a backup copy

Missing data: what it is and why it matters

System missing

A cell you simply left empty. SPSS shows a full stop. It has no meaning attached to it.

User missing

A code you defined, such as 9 or 999, which records the reason a response is absent.

The consequence of ignoring it

If 9 is not declared as missing, SPSS treats it as a genuine score of nine on a five point scale. Your mean is then inflated, your standard deviation is distorted and your correlations are wrong. Declaring missing values is therefore a matter of validity, not of tidiness.

Pattern	Meaning	Implication
Missing completely at random	The gap is unrelated to anything in the study	Least damaging. Deletion is usually safe
Missing at random	The gap relates to other observed variables	Imputation methods are appropriate
Missing not at random	The gap relates to the missing value itself	Most damaging. Requires cautious reporting

A practical threshold

As a broad working guide, missingness below about five per cent on a variable is usually treated as negligible, while high missingness on a key variable should be reported and justified rather than hidden. Always state in your methods chapter how you handled it.

Handling missing data in SPSS

1

Declare the codes

Variable View, click the Missing cell, choose Discrete missing values, enter 9 or 999, click OK

2

Detect the extent

Analyze, Descriptive Statistics, Frequencies. Read the Missing row of the statistics table

3

Decide the strategy

Listwise deletion, pairwise deletion, mean substitution, regression or multiple imputation

4

Apply and record

Most dialogue boxes carry an Options button where listwise or pairwise is chosen. Record your choice

5

Report honestly

State the amount of missing data, the pattern and the method used in your methodology chapter

Choosing between the two commonest options

Listwise deletion

A case is dropped entirely if it is missing on any variable in the analysis. It keeps every result based on the same sample, but can shrink your sample severely.

Pairwise deletion

A case is used wherever it has the data required. It preserves sample size, but different parts of your output rest on different numbers of cases.

Mean substitution

Simple and quick, but it artificially reduces variance. Use it sparingly and disclose it.

From items to constructs: why we compute variables

A psychological construct such as self esteem is not measured by a single question. It is measured by a set of items. Before you can analyse self esteem, you must combine those items into one score. That combination is done in SPSS through Transform then Compute Variable.

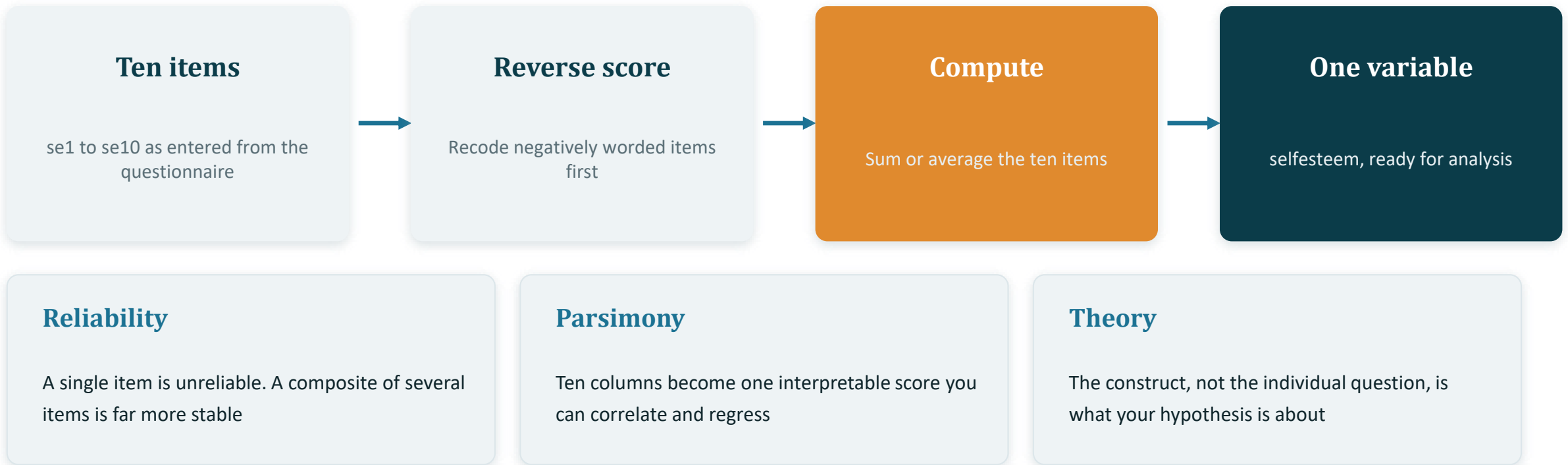


FIGURE 5

Transform then Compute Variable: the procedure

- 1 From the menu bar click Transform, then Compute Variable
- 2 In the Target Variable box on the left, type the name of the new variable, for example selfesteem
- 3 Click Type and Label and give it a full descriptive label
- 4 Build the expression in the Numeric Expression box on the right
- 5 Use the Function group list to find SUM, MEAN and other functions
- 6 Move variables into the expression using the arrow button
- 7 Click If to compute only for a subgroup, if that is what you need
- 8 Click OK. The new variable appears as the last column in Data View

The screenshot shows the 'Compute Variable' dialog box in SPSS. The title bar is dark blue with the text 'Compute Variable'. The dialog is divided into two main sections. On the left, under the heading 'Target Variable:', there is a text box containing 'selfesteem' and a larger list box below it containing the following variables: 'id', 'sex', 'age', 'faculty', 'level', 'se1', 'se2', and 'se3'. On the right, under the heading 'Numeric Expression:', there is an equals sign followed by a text box containing 'MEAN(se1 TO se10)'. Below this, there is a list box for 'Function group:' containing 'All', 'Arithmetic', and 'Statistical'. At the bottom right, there are two buttons: an orange 'If...' button and a blue 'OK' button.

Figure 5. Layout of the Compute Variable dialogue box.

Worked example: computing a self esteem score

Expression you type	What it produces	When to use it	Range for ten items scored 1 to 5
se1 + se2 + se3 + ... + se10	A simple total	Only when there is no missing data at all	10 to 50
SUM(se1 TO se10)	A total that ignores missing items	When you want a total score	10 to 50
SUM.10(se1 TO se10)	A total computed only if all ten items are present	When you want to enforce completeness	10 to 50
MEAN(se1 TO se10)	The average item response	When you want the score on the original item metric	1 to 5
MEAN.8(se1 TO se10)	The mean computed only if at least eight items are present	Recommended default for scale scoring	1 to 5

The TO shortcut

se1 TO se10 works only if the ten variables sit next to each other in that order in Variable View. This is why systematic naming and ordering pays off.

The dot notation

The number after the full stop sets the minimum count of valid items required. MEAN.8 means compute only if at least eight of the ten items were answered.

Before you compute: reverse score the negative items

If an item is worded negatively, a high score means low standing on the construct. Adding it directly to the others cancels out real variance. Reverse it first using Transform then Recode into Different Variables.

Original value	Becomes	Rule for a five point scale
1	5	New value = 6 minus old value
2	4	New value = 6 minus old value
3	3	The midpoint stays the same
4	2	New value = 6 minus old value
5	1	New value = 6 minus old value

Table 3. Reverse scoring map for a five point Likert item.

For a scale scored 1 to 5 the constant is 6. For a scale scored 1 to 7 it is 8.
The rule is simply the maximum plus the minimum.

The procedure

- 1 Transform, then Recode into Different Variables
- 2 Move the negative item, for example se4, into the box
- 3 Under Output Variable type a new name such as se4r and click Change
- 4 Click Old and New Values and enter each pair: 1 becomes 5, 2 becomes 4 and so on
- 5 Click Add after every pair, then Continue, then OK
- 6 Use se4r, not se4, in your Compute expression

Always recode into a NEW variable, never over the original. Your raw data must remain recoverable.

Descriptive statistics: summarising what you have

Descriptive statistics organise and summarise a dataset without generalising beyond it. They are the first analysis you run, both to describe your sample and to check that your data are sound enough for inferential work.

Frequency	Central tendency	Dispersion	Shape
• Counts • Percentages • Valid percentages • Cumulative percentages	• Mean • Median • Mode	• Range • Minimum and maximum • Variance • Standard deviation	• Skewness • Kurtosis • Histogram with normal curve

Match the statistic to the scale: report frequencies and the mode for nominal variables, the median and range for ordinal variables, and the mean with the standard deviation for interval and ratio variables. Reporting a mean for faculty is meaningless, however neatly SPSS prints it.

TABLE 4

Which descriptive procedure should you use?

Menu path	What it gives you	Best suited to
Analyze, Descriptive Statistics, Frequencies	Frequency tables, percentages, plus optional mean, median, mode, standard deviation, skewness and charts	Nominal and ordinal variables. Also the first cleaning check for every variable
Analyze, Descriptive Statistics, Descriptives	A compact table of N, minimum, maximum, mean and standard deviation. Can also save z scores	Continuous variables, especially when you have many of them
Analyze, Descriptive Statistics, Explore	Detailed summary by group, with confidence intervals, trimmed mean, boxplots, stem and leaf plots and normality tests	Checking normality and comparing subgroups before inferential tests
Analyze, Descriptive Statistics, Crosstabs	A two way table of counts and percentages, with chi square available	Relationships between two categorical variables

Work in this order: Frequencies on every variable first, as a cleaning check, then Descriptives on your continuous variables and computed scores, and Explore only when you need to test normality.

TABLE 5

Reading the frequency output

Analyze, Descriptive Statistics, Frequencies. Move sex into the Variables box, tick Display frequency tables, click OK.

Sex of respondent	Frequency	Per cent	Valid per cent	Cumulative per cent
Valid Male	86	43.0	44.1	44.1
Female	109	54.5	55.9	100.0
Total	195	97.5	100.0	
Missing No response	5	2.5		
Total	200	100.0		

Table 5. Frequency distribution of respondents by sex.

Per cent

Calculated on all 200 cases, including the missing ones

Valid per cent

Calculated only on the 195 who actually answered.
This is what you report

Cumulative per cent

A running total. Meaningful for ordinal variables, not for nominal ones

TABLE 6

Reading the descriptives output and reporting it

Variable	N	Minimum	Maximum	Mean	Std. deviation	Skewness	Kurtosis
Age in completed years	196	18	39	22.84	3.117	0.642	0.318
Total self esteem score	193	1.40	4.90	3.512	0.684	-0.271	-0.405
Valid N (listwise)	190						

Table 6. Descriptive statistics for age and total self esteem score.

Reporting it in APA style

Participants ranged in age from 18 to 39 years, with a mean age of 22.84 years and a standard deviation of 3.12. The mean total self esteem score was 3.51 with a standard deviation of 0.68, indicating a moderately high level of self esteem in this sample.

Checking normality from this table

As a common working guide, skewness and kurtosis values between minus two and plus two suggest an acceptably normal distribution for most parametric tests. Both variables here fall comfortably inside that range.

For a formal test, use Explore and read the Shapiro Wilk statistic.

Ten errors that cost postgraduate students the most time

1 Entering data before defining variables

2 Leaving variables as VAR00001 with no label

3 Typing words instead of numeric codes

4 Failing to declare missing value codes

5 Setting Measure to Scale for nominal variables

6 Computing a scale score before reverse scoring

7 Using plain SUM when items are missing

8 Deleting rows instead of correcting cells

9 Reporting per cent rather than valid per cent

10 Never saving, or saving over the raw file

Every one of these is preventable in the first fifteen minutes of working with a dataset.

Your take home task and what comes next

Take home task

- Take twenty completed questionnaires from your own study
- Define every variable fully in Variable View and save the file
- Enter all twenty cases into Data View
- Declare your missing codes and run Frequencies on every variable to check for impossible values
- Reverse score any negatively worded items into new variables
- Compute at least one scale score using MEAN with dot notation
- Run Descriptives on your computed score and bring the output to the next session

Coming in the next sessions

- Reliability analysis and Cronbach alpha
- Testing assumptions: normality, linearity, homogeneity of variance
- Correlation and simple regression
- Independent samples t test and ANOVA
- Chi square for categorical relationships
- Multiple regression, mediation and moderation
- Exporting SPSS output into an APA formatted results chapter

Questions and discussion

Define carefully, enter honestly, transform deliberately, describe accurately. Everything else in quantitative analysis is built on those four habits.

Facilitator: Assoc. Prof. Enoch Oladunmoye
College of Humanities and Social Sciences
Kampala International University
Email: enoch.oladunmoye@kiu.ac.ug