Quantitative Research Methods and Data Analysis Workshop 2020

15 - 26 June 2020

T Probert[a], M Schaefer[b] and C Wilsenach[b]

Rhodes University[a] and University of South Africa[b]

CITATION

Probert, T., Schaefer, M. & Wilsenach, C. (2020). Quantitative Research Methods and Data Analysis Workshop 2020. DOI: *10.25399/UnisaData.12581483*

Please direct feedback and suggestions to Ms M Schaefer schaemn@unisa.ac.za

---

COURSE GUIDE

0.1. OUTCOMES

By the end of this workshop you will be able to:

- outline the steps and decisions involved in quantitative data analysis of linguistic data
- explain common statistical terminology (sample, mean, standard deviation, correlation, nominal, ordinal and scale data)
- perform common statistical tests using jamovi (e.g. t-test, correlation, anova, regression)
- interpret and report common statistical tests
- describe and choose from the various graphing options used to display data
- use jamovi to perform common statistical tests and graph results

Evaluation
You will use these skills and knowledge to complete the following activities for evaluation:

- analyse the data for your project and/or assignment (in part or in whole)
- plan the results section of your Honours project (where applicable)

0.2. INTRODUCTION

After completing these sessions, you should be able to independently analyse quantitative linguistic data, and report and interpret your findings. These sessions give you just enough guidance to find your feet. It's best to think of these sessions as your first step on an exciting journey of mastery that your instructors themselves are also on.

Course Structure

The course is set up to include both asynchronous and live sessions. Asynchronous sessions are completed in your own time. They take up the majority of the course, and should be completed before the live sessions. The asynchronous sessions will include prescribed reading (either in the course sheet or in the prescribed book), a concept video, and one or two demonstration videos. We also include activity sessions that you should complete in preparation for the live sessions to check your understanding of the content. We specify a minimum amount of time you should spend on each topic and give you plenty of opportunities for practice. We encourage you to watch each demonstration and then follow along on the second watch with the data. It is important to practice yourself so you can make sure you understand what to do. Finally, the end of course assessment will be used to determine if you have met the course objectives. Rhodes students should complete one of the two assignments prescribed for the Postgraduate Literacy module. Other students, should they want to receive a certificate of completion, can complete and need to pass an assignment which you can find in section 8 of this document.

The live sessions will take place on Friday 19 June 14.15 and Friday 26 June 14.15 and will be used to revise content and work through examples. We will be drawing on the activities in class, so please make sure that you have prepared these in advance.

Dates

The program is split over two weeks:

Week 1; 15 - 19 June; complete sections 1 - 4; Live session Friday 19 June at 14:15

Week 2; 22 - 26 June; complete sections 5 - 9; Live session Friday 26 June at 14:15

Support and Communication

We will also use Slack for collaborative note taking and question asking. Join our Slack channel [not available]. You may need to sign up for an account using your email address. Choose one of the channels below that best suits your purpose and post your comments/questions in the box

Use #general for collaborative note taking and dialogue with your classmates.

Use **#random** to post non-work related content, such as statistics memes and tiktoks. We will award points (that don't mean anything) to the funniest contribution.

Use **#statistics-workshop-2020** to post your questions on the content. Instructors will respond during the hours posted below, and participants are encouraged to help their classmates.

Instructor availability on Slack (please use Slack, not email)

Wednesdays 14:00 - 15:00

Thursdays 14:00 - 15:00

0.3. PREPARATION

Please complete the activities listed under preparation BEFORE the course starts on Monday 15 June 2020.

0.3.1. Activity: Pre-Workshop Questionnaire

0.3.2. Activity: Download and Install jamovi

0.3.3. Activity: Download Files

0.3.4. Activity: Required Reading

0.3.1. ACTIVITY: Pre-Workshop Questionnaire***

Please complete the course baseline questionnaire here.

Completion of this questionnaire is part of the pre-requirements for the Statistics Workshop.

Note that your responses are anonymous.

*** Not available. .

0.3.2. ACTIVITY: Download and Install JAMOVI (Windows)

View the video demonstration here.

JAMOVI is an open-source statistical software designed to be as simple to use as possible, but still allowing very sophisticated analyses.

To download and install it, let's go first of all to the site: https://www.jamovi.org/download.html

Click on the *recommended for most user's* option (to date, 1.1.9 solid), and download the installer (see Figure 0.3.1).

If you use another operating system (for example MacOS), click on the bottom of the word MacOS, solid (.dmg 1.1.9). It is NOT possible to install the software on a cellphone or TABLET (such as Ipad).

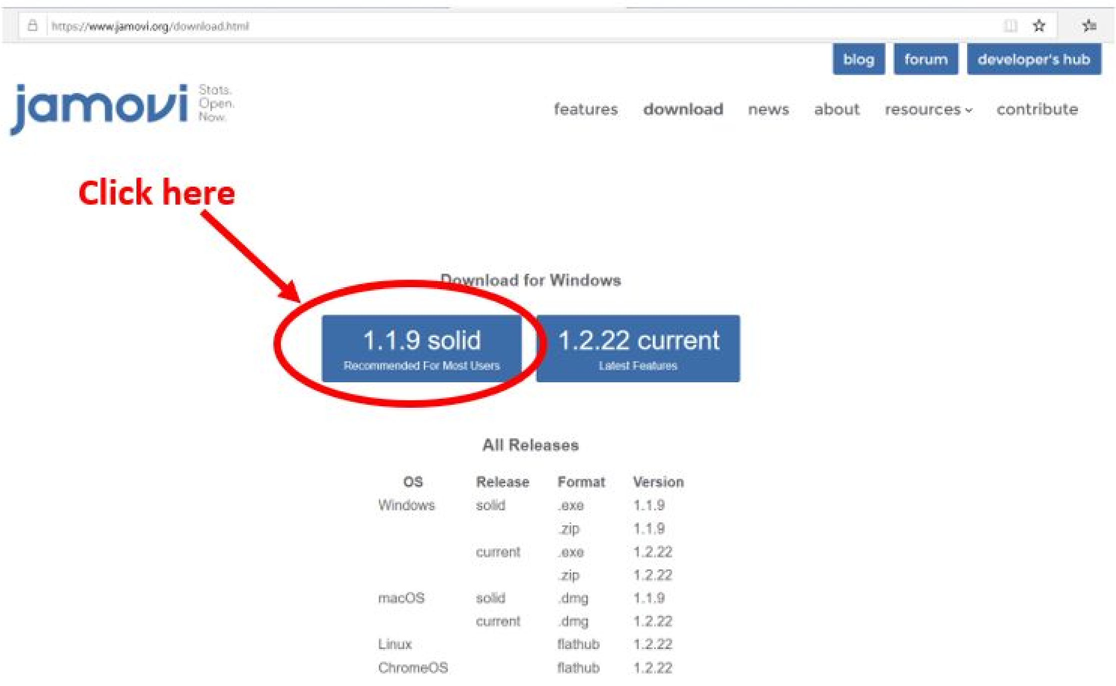


Figure 0.3.1. JAMOVI website

Once you have clicked on the blue block (1.1.9 solid), the installer should automatically begin downloading in the bottom left hand corner of your screen. See Figure 0.3.2 below.

Figure 0.3.2. Installer download window

Alternatively, a pop-up bar will appear at the bottom of your screen. See Figure 0.3.3. Click on the option 'Run.'

Figure 0.3.3. Installer pop-up window

Once the software has been downloaded, we click on the Installer (Figure 0.3.4), to start the Setup.

Figure 0.3.4. Installation file for Windows

A pop-up message may appear on your screen, "*Do you want to give this app from an unknown publisher permission to make changes on your device?*" Say yes.

Once the Installer has started, just click on "Install" and wait for the installation to complete (see Figure 0.3.5). At the end of the installation, click on "Finish".

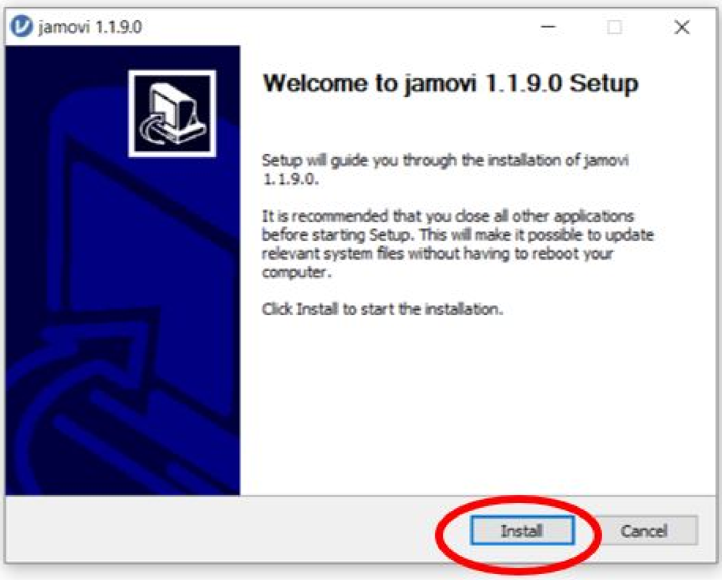


Figure 0.3.5 JAMOVI Installer Screen

Click on the JAMOVI logo (you can find it in "Start a search" -> "Jamovi") to start the software (see Figure 0.3.6).

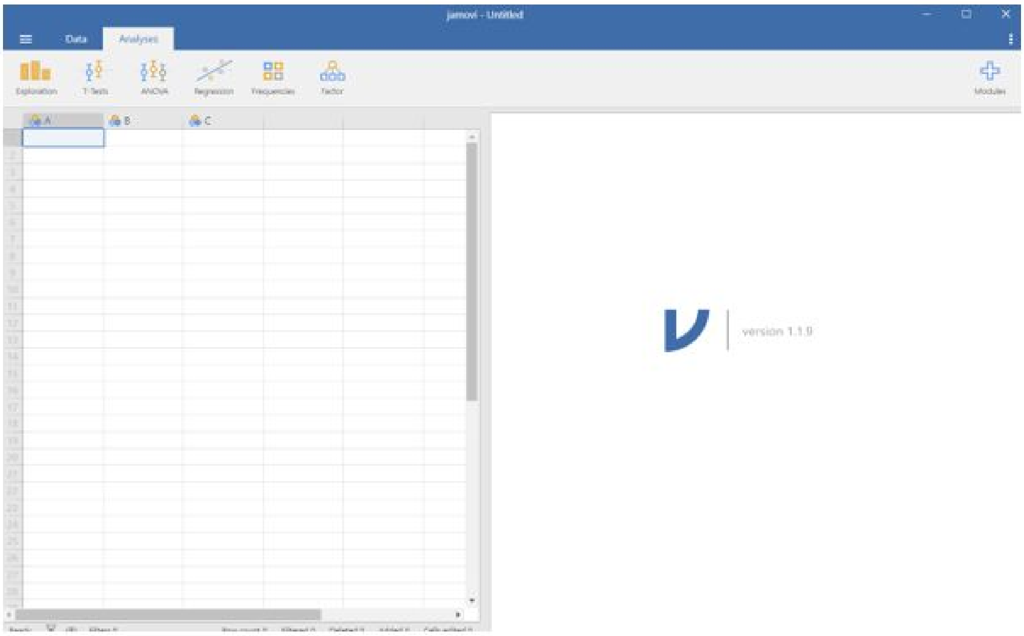


Figure 0.3.6. Jamovi's main screen

0.3.2.1. Cite jamovi

It is ethical practice to cite the data analysis software you use in your manuscript/project. The citation appears after you complete an analysis in jamovi in the Results tab. Remember to use the citation from your software if it differs from that below.

The jamovi project (2019). *jamovi.* (Version 1.1) [Computer Software]. Retrieved from https://www.jamovi.org.

0.3.3. ACTIVITY: Download Files (data, materials, articles)

We will be using datasets that have been made available using Open Science principles.

0.3.3.1. Download the demonstration dataset: MA and Comprehension

We will be using this dataset to demonstrate the analyses learned in this course.

James, E., Currie, N. K., Tong, S. X., & Cain, K. (2020, May 21). The relations between morphological awareness and reading comprehension in beginner readers through to young adolescents. Retrieved from https://osf.io/96542/

Follow this link to get to the project page. **Download** the dataset_postImp.csv and README_DataNotebook.pdf in the Data folder on OSF (Figure 0.3.3.1.1). To do this, click on the file. A new window opens. In the top right corner there is a blue Download button. Click Download, and then save each file in the folder you have set aside for the data analysis course. You may also download the Materials if you want a better sense of the test instruments, but these will not be referred to in the course.

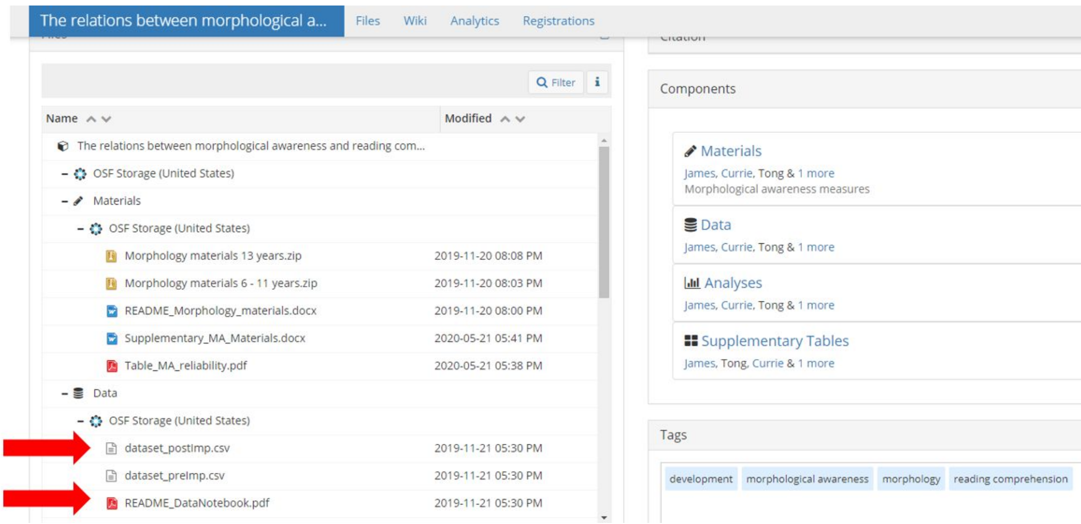


Figure 0.3.3.1.1. Where to find dataset_postImp.csv and README_DataNotebook.pdf on the author's OSF project page

The comma separated value file (dataset_postImp.csv) includes the data from the project after imputing missing values. The README_DataNotebook.pdf file contains information on each variable in the dataset. The dataset should open in Excel (we will later import it into jamovi) and the README file will open in your system's pdf viewer.

This dataset includes assessment data from three groups of children (6-8 year olds, 9-11 year olds, and 12-13 year olds) in the UK. The children completed assessments on reading accuracy, reading rate (fluency), reading comprehension, morphological awareness, vocabulary, phoneme awareness and IQ. **Although we do not have access to a proposal, preprint or published article we can see from the project title that the researchers are mostly interested in the relationship between morphological awareness and reading comprehension in different age groups.** The other variables are most probably included as control variables.

0.3.3.2. Download the student activity dataset and preprint: Trilingual Vocabulary

You will be using this **dataset** to practise the analyses learned in this course.

Côté, S. L., Gonzalez-Barrero, A. M., & Byers-Heinlein, K. (2020, February 18). Multilingual toddlers' vocabulary development in two languages: Comparing bilinguals and trilinguals. Retrieved from https://osf.io/us27h/

Follow [this link](https://osf.io/us27h/) to get to the project page. **Download** the matched_dataset.csv, trilingual_vocab.csv and Trilingual Project Codebook .csv in the Data folder on OSF (Figure 0.3.3.2.1). To do this, click on the file. A new window opens. In the top right corner there is a blue Download button. Click Download, and then save each file in the folder you have set aside for the data analysis course.

Trilingual Project Codebook .csv lists the variables and what values they take. matched_dataset.csv includes the data on the trilinguals exactly matched to the bilinguals. trilingual_vocab.csv includes data on all the children in the study.

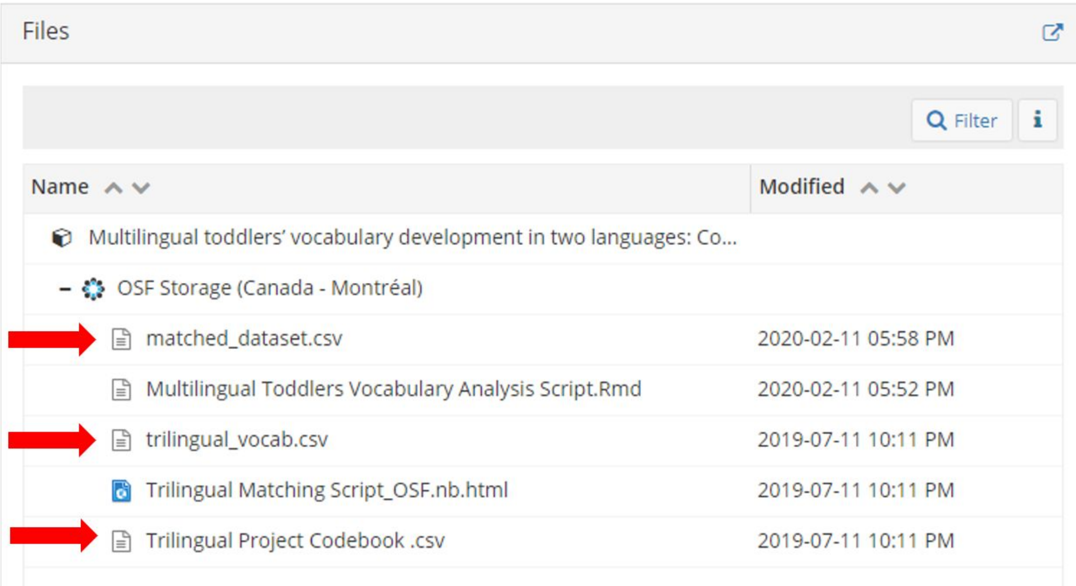


Figure 0.3.3.2.1. Where to find the data files on the author's OSF project page

This dataset has an accompanying **preprint article.** This is a manuscript which has been submitted for publication but has not yet been peer reviewed.

Côté, S. L., Gonzalez-Barrero, A. M., & Byers-Heinlein, K. (2020, February 18). Multilingual Toddlers' Vocabulary Development in Two Languages: Comparing Bilinguals and Trilinguals. https://doi.org/10.31234/osf.io/hpqez

Download the preprint from [this link](https://doi.org/10.31234/osf.io/hpqez). The download button is on the top right of the manuscript. Read the entire manuscript paying particular attention to: the abstract (acts as a summary), method (information on participants and how data was collected, and analysed) and results.

0.3.3.3. More datasets (for downloading later)

Here are additional datasets that you can use for practice. We recommend that you download the data later if you want to use the data for your own practice.

Hamalainen, J. (2020, April 30). Cognitive data for auditory processing and phonological development project. Retrieved from https://osf.io/rhmq5/

Note: use Aufon_cognitive_assessements.sav to preserve labels

Weng, Tsung-han. (2020). Survey data of foreign language learners' enjoyment and anxiety in the U.S.. *Data in Brief*, 30, 105221. https://doi.org/10.1016/j.dib.2020.105221

Note: download the data from the left hand side of the screen, under Multimedia component 1 and Multimedia component 2.

If you cannot open CSV files in your Microsoft Excel you can access the Workbooks here.

0.3.4. ACTIVITY: Required Reading

The prescribed book for this course is Navarro and Foxcroft (2019). It is an open-source (freely available) book written for introductory statistics courses, using examples from jamovi.

Navarro DJ and Foxcroft DR (2019). learning statistics with jamovi: a tutorial for psychology students and other beginners. (Version 0.70). DOI: 10.24384/hgc3-7p15

Before the course starts, read:

Chapter 1 (pp. 3-12) | Why do we learn statistics?

Chapter 2 (pp. 13 - 40) | A brief introduction to research design

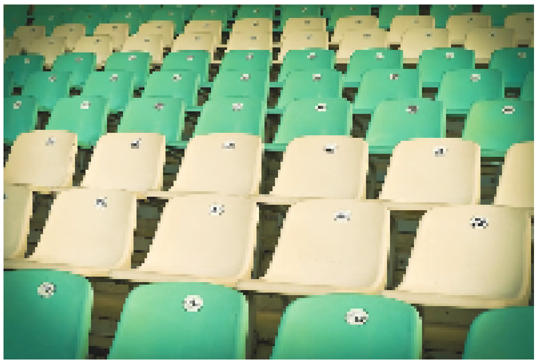
---

1.0 Spreadsheets in Microsoft Excel and jamovi

In this section, we take you through key points about understanding how spreadsheets work in Excel, and jamovi. We also look at how to edit the dataset in jamovi.

The objectives are met when the participant can:

- open a dataset in Excel
- manage workbooks, and sheets in Excel
- enter data into Excel
- identify rows and columns, observations and variables in Excel
- identify and justify a decision on levels of measurement
- open jamovi
- import data into jamovi
- save a jamovi file (.omv)
- edit column names in a jamovi file
- identify and edit levels of measurement in jamovi
- edit the levels of categorical/nominal variables in jamovi



Section 1 should take at least 1 hour to complete.

1.1. CONCEPT: Working with spreadsheets

1. Watch this video (11:22) to learn about how to work with spreadsheets in Microsoft Excel.
2. The video also introduces you to key terms; cases, variables, observations, and levels of measurement.
   Note: The word *observations* can have more than one meaning. Strictly, it refers to a single cell (as illustrated in the video). But, if you are speaking about a statistics programme, a case can also be referred to as an observation.
3. After watching the video once or twice, you are encouraged to familiarise yourself with using Excel.

Bonus: What's the difference between a spreadsheet saved as .xls and a spreadsheet saved as .csv?
Both formats store data. CSV files store the data in plain text with commas separating each value. Because it is a plain text file it can be opened by anyone, even those who do not have access to Microsoft Office. Excel uses proprietary software which means someone has to own the program to open the file. It's best practice to make your data available in .csv format so anyone can open it.

1.2. DEMONSTRATION: Working in jamovi

Tackle this session along with:

Navarro & Foxcroft (2019) Chapter 3 | Getting started with jamovi

1. Watch this video (5:20) to learn about how to navigate jamovi before data is imported.
2. Watch this video (6:09) to learn about variable types and labels. Note: *focus on the concepts, we will apply these concepts to our data in the next video.*
3. Watch this video (5:47) to learn about how to import data into jamovi, save data and edit columns and levels of measurement in jamovi. We use dataset_postImp.csv from the James et al. (2020) (MA and Comprehension) data for demonstration purposes. (See section 0.3.3.1)
4. After watching the video once or twice, you are encouraged to follow along on jamovi yourself before attempting the next activity with a different dataset. Close your jamovi session before moving to a new dataset.

1.3. ACTIVITY: Working in jamovi - Trilingual Vocabulary Data

Did you meet the objectives of this session? Let's see - follow the steps below and answer the questions.

1. In Excel, open Trilingual Project Codebook .csv (See section 0.3.3.2)
    - 1.1. How many variables are there in this dataset, according to the codebook?
    - 1.2. How many different **types**[1] of variables are there? e.g. are they all continuous variables?
    - 1.3. Which variable(s) uniquely identify the child's identity while keeping him/her anonymous?
    - 1.4. Which variable refers to whether a child is bi- or trilingual?
    - 1.5. What does TotalNW_Eng and TotalNW_Fr refer to?
    - 1.6. Refer to the preprint:
        - 1.6.1. How was LangExpRatio calculated?
            - 1.6.1.1. What does a value of 2 indicate?
            - 1.6.1.2. What does a value of 0.5 indicate?
        - 1.6.2. The authors explain that they added the French and English total words produced together. Which variable refers to the total words produced overall?
2. Open jamovi.
3. In jamovi, open matched_dataset.csv from Côté, Gonzalez-Barrero and Byers-Heinlein (2020) (See section 0.3.3.2)
    - 3.1. How many cases (observations) are there in this dataset?
    - 3.2. Some variables are not set at the correct level of measurement when compared to the levels of measurement specified in the codebook. Let's fix this.
        - 3.2.1. Change Sub_ID and Study_ID to ID variables
        - 3.2.2. Lang_Group and Dom_Lang are correctly specified as nominal variables
        - 3.2.3. Age_Days and Age_Months are not specified correctly. They are indicated as nominal variables, but we know that age varies from a value of 0 to a reasonably old age. Correct these variables' level of measurement.
        - 3.2.4. Is the level of measurement of Mat_Edu_Years specified correctly? *Hint: look at the codebook and Table 1 in the preprint to see how the authors dealt with this variable*
        - 3.2.5. Lang_Exp_Eng, Lang_Exp_Fr and Lang_Exp_Other seem to be correctly specified at continuous level
        - 3.2.6. Specify the correct level of measurement for: TotalNW_Eng, TotalNW_Fr, Total_Vocabulary and Total_Conceptual_Vocabulary
    - 3.3. Save your dataset as matched_dataset Note: *this save does not overwrite your matched_dataset.csv file as jamovi files save with the .omv extension -> matched_dataset.omv*

---

[1] Note: This dataset was created to be used with R so some of the terminology in Column B is different. In R, nominal/categorical variables are called factors. jamovi makes a distinction between the usual categorical variables (e.g. male and female) and categorical variables that specify unique numbers (e.g. participant IDs). In R, numeric variables are also continuous variables. In R, a string variable is one that contains characters/text.

2.0. Research Design

In this section, we take you through a very basic description of research designs and the thinking that goes behind creating the research design for a study.

The objectives are met when the participant can:

- list and define the steps involved in designing a research study
- define probability and non probability sampling methods
- identify the strengths and weakness of cross-sectional and longitudinal research designs
- identify the research design of a study

Before continuing with the session below, you need to read:

Navarro & Foxcroft (2019) Chapter 2 | A brief introduction to research design

Section 8.1. of Navarro & Foxcroft (2019) Chapter 8 | Estimating unknown quantities from a sample

We recommend that you first skim read this chapter and chapter section to get a sense of the main ideas, then go through the rest of sessions 2.1. - 2.3. Return to the book if things are still unclear.

Section 2 should take at least 2 hours to complete.

2.1. CONCEPT: Research design

Watch this video (27:34) to learn about

- what a research design is,
- what steps are involved in designing a research study,
- what the advantages and disadvantages of cross-sectional and longitudinal quantitative designs are, and
- to go through some examples.

2.2. DEMONSTRATION: Identifying research designs

Let's try to identify the possible research designs that were used to collect the data we have access to.

2.2.1. Demonstration Dataset: MA and Comprehension

James et al. (2020) collected data from three groups of children (6-8 year olds, 9-11 year olds, and 12-13 year olds) in the UK. In the 6-8 year old band, children were sampled from grades 2 and 3. For the 9-11 year old, children were sampled from grades 5 and 6. The 12 and 13 year olds were all in grade 8. Children were assessed only once on a range of tasks including vocabulary, phoneme elision, morphological awareness, reading fluency, reading accuracy and reading comprehension.

This study uses a **quantitative, cross-sectional, quasi-experimental** design. The study used primary data, i.e. data that is collected for the purposes of the study.

The study is **quantitative** because theoretical constructs (such as morphological awareness) are measured with a test (such as the morphological awareness inflection analogy test) that assigns a number (out of 24 in this case) which reflects how much of that construct (morphological awareness) the participant has.

The study is **cross-sectional** as the data was collected at one point in time (even though many tests were done). There was also no intervention as the authors relied on existing age differences in the groups.

The study is **quasi-experimental** because the authors use age group as an independent variable without necessarily manipulating any other variables.

2.2.2. Student Activity Dataset: Trilingual Vocabulary

Côté et al. (2020) analyse data that was already collected for other experiments (see pp. 6 -7 The Current Study). Thus, this study presents a secondary data analysis.

At the start of the Method section, the authors state:

> *Archival data from 215 multilingual toddlers (age range: 17 - 33 months) who had participated in studies at an infant research laboratory between January 2010 and May 2019 and met the trilingual or bilingual exposure criteria ... were available for the ... study* (Côté et al., 2020, p. 7).

Most children were assessed only once, but 24 children were assessed more than once at different time points. The authors analyse these children's different data points as if each data point corresponds to another observation. The authors use language group, language exposure, vocabulary and age as their variables. The authors compare the bilingual and trilingual children's performance.

This study uses a **quantitative, cross-sectional, quasi-experimental** design. The study used secondary data, i.e. data that has already been collected.

The study is **quantitative** because theoretical constructs (such as vocabulary) are measured with a test (such as the CDI) that assigns a number which reflects how much of that construct (vocabulary) the participant has.

The study is **cross-sectional.** Most participants were assessed only once, and the participants who were tested more than once are treated as if each assessment was its own observation.

There was also no intervention as the authors relied on existing bilingual/trilingual differences in the groups.

The study is **quasi-experimental** because the authors use multilingualism (whether a participant was bi or trilingual) as an independent variable. In some of their analyses, the authors attempt to match the trilingual participants to bilingual participants such that these participants were almost the same with regard to age, sex, socioeconomic status, maternal education and language exposure ration (Côté et al., 2020). However, because there is no intervention, and the participants were not randomly allocated the design remains quasi-experimental.

2.2.3. A South African Example: The Second Early Grade Reading Study

Schaefer and Kotze (2019) explain that the Second Early Grade Reading Study (EGRS II) is designed as an randomised control trial (RCT), which aims to measure and compare the effects of two teacher-support interventions to the situation among a comparison group of learners who receive regular teaching. The interventions include structured lesson plans, reading coaches and additional material and focus on strengthening teachers' instructional practices for EFAL. The study began in 2017 in 180 no-fee primary schools in the Mpumalanga province of South Africa and will continue until 2019.

At the start of the study, 50 schools were randomly allocated to Intervention 1 (coaching), 50 schools were randomly allocated to Intervention 2 (virtual coaching) and 80 Schools were randomly allocated to the control group (business as usual). Within each school, 20 children were randomly selected in Grade 1 to be assessed on a range of preliteracy, literacy and language skills in their Home Language (isiZulu or Siswati) or English (the additional language). The same group of children were reassessed at the end of each grade on similar literacy and language assessments. See Figure 2.2.3.1.

In each year, other aspects of the project were also evaluated. 60 teachers' lessons for English as First Additional Language was observed, and teachers completed structured interviews with fieldworkers. In another study conducted each year, a researcher completes qualitative case studies with six teachers in six schools.

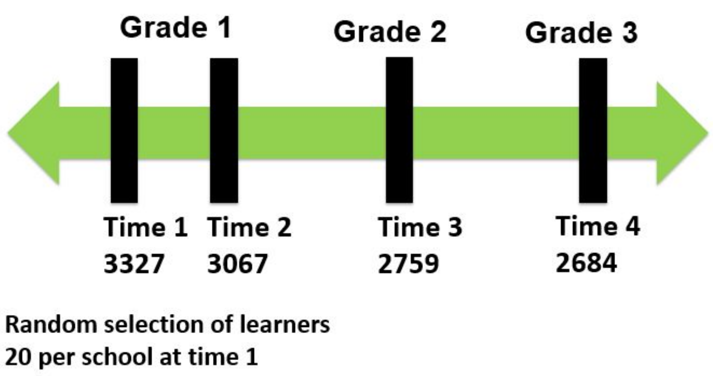


Figure 2.2.3.1. Number of participants assessed at each time point in the Second Early Grade Reading Study

This study uses a **mixed-methods, longitudinal, experimental** design. The study used primary data, i.e. data that is collected for the purposes of the study.

The study uses **mixed-methods**. There are quantitative aspects to the study, such as the assessment of children's language and literacy skills and using a form to record data from classroom observations. There are also qualitative aspects to the data collection, such as in the use of in-depth case studies, and interviews.

The study is **longitudinal**. The **same** participants were assessed at the start of Grade 1, at the end of Grade 1, at the end of Grade 2, and at the end of Grade 3. There is also a plan to assess the same learners at the end of Grade 4.

The study is **experimental** because the authors randomly allocated schools and the participants within them to different treatment groups. The intervention groups received a coaching and virtual coaching intervention, and the control group continued with teaching as usual. Because the schools, and the participants within the schools, were randomly selected, the researchers were able to establish that the groups were similar to each other. Thus, any changes in language and literacy skills later in the study can be attributed to the interventions.

Sources

*Kotze, J., Fleisch, B., & Taylor, S. (2019). Alternative forms of early grade instructional coaching: Emerging evidence from field experiments in South Africa. *International Journal of Educational Development*, 66(September 2018), 203–213. https://doi.org/10.1016/j.ijedudev.2018.09.004

Schaefer, M. & Kotzé, J. (2019). Early reading skills related to Grade 1 English Second Language literacy in rural South African schools. *South African Journal of Childhood Education*, 9(1), a644. https://doi.org/10.4102/sajce.v9i1.644

*This study is unfortunately not published in open-access format.

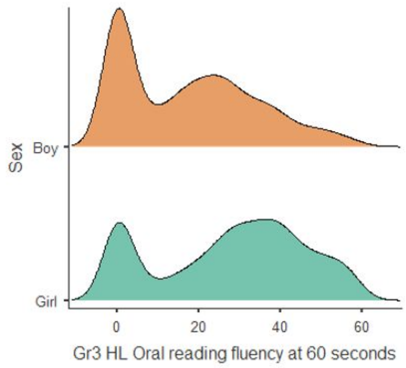
***The EGRS II builds on the work of the first Early Grade Reading Study. You can find the reports and data collection instruments here.

2.3. ACTIVITY: Identifying research designs

Similar to DEMONSTRATION 2.2. identify the research design of any ONE open-access study/dataset you have access to.

You may use any of the datasets presented in section 0.3.3.3., the open-access dataset for your assignment, or an open-access academic preprint or academic article. This means that the article should be downloadable by everyone. Articles which are open access usually say "This work is licensed under CC Attribution 4.0" (e.g. the article written by Probert, 2019).

3.0. Descriptive Statistics



In this section, we take you through exploring your data in numerical and graphical form. First, we provide some background information on data exploration in the CONCEPT sessions then we demonstrate how to generate numerical and graphical summaries of your data using jamovi in the DEMONSTRATION session. This is quite simple in jamovi as all the relevant menus are stored in the Exploration tab. Finally, we ask you to test your own understanding by asking you to complete your own exploratory data analysis in the ACTIVITY session.

The objectives are met when the participant can:

- explain common statistical terminology (sample, mean, standard deviation, median, percentile, box plot, violin plot)
- summarise data numerically appropriately for the data
- summarise data using graphs appropriately for the data
- explain what the numerical and graphical data summaries show

Before continuing with the session below, you need to read:

Navarro & Foxcroft (2019) Chapter 4	**|**	**Descriptive statistics**
Navarro & Foxcroft (2019) Chapter 5	**|**	**Drawing graphs**

We recommend that you first skim read these chapters to get a sense of the main ideas, then go through the rest of sessions 3.1 - 3.6. Return to the book if things are still unclear.

Section 3 should take at least 2 hours to complete.

3.1. CONCEPT: Inspecting data using numbers

Watch this video (9:18) for an introduction to basic statistical terms used when describing data numerically.

Remember these concepts are addressed in chapter 4 of Navarro and Foxcroft (2019).

3.2. CONCEPT: Inspecting data using graphs

Watch this video (8:24) for an introduction to graphs used to summarise data.

Remember these concepts are addressed in chapter 5 of Navarro and Foxcroft (2019).

We've only touched on a very small number of possible graphs you can use to summarise your data, partly because we want to keep the content manageable, and because we want to show you what's possible in jamovi (for now).

Need some inspiration and motivation to learn R? Have a look at the R Graph Gallery to expand your horizons. To pique your interest, why not find out what correlogram and ridgeline plots are? We're sure you're very keen to try out **ggplot2**.

3.3. DEMONSTRATION: Inspecting data using numbers and graphs

Navarro and Foxcroft (2019) go through how to run descriptive statistics in jamovi (see Chapters 3 and 4). Reading through these chapters will give you enough information to complete descriptive statistics and create graphs. If you are limited in your access to data to access videos, you will be able to complete the session objectives by only reading the chapters and following along in jamovi. The videos below provide visual assistance and use a linguistic dataset.

For the video demonstrations, we will use the James et al. (2020) data to demonstrate how to run descriptive statistics and generate graphs so we can better understand the data. We will use only some of the variables for the purposes of this demonstration. We can think of these variables as fitting into two types (Table 3.3.1). First, our independent / predictor variables are those that we "use to do the explaining" (Navarro & Foxcroft, 2019, p. 23). Independent / outcome variables are the x in an equation. Second, the dependent or outcome variables are those that are being explained. Dependent variables are y in an equation.

Table 3.3.1. "The terminology used to distinguish between different roles that a variable can play when analysing a dataset." (Navarro & Foxcroft 2019, p. 24)

role of the variable	classical name	modern name
"to be explained"	dependent variable (DV)	outcome
"to do the explaining"	independent variable (IV)	predictor

Source: Navarro & Foxcroft 2019, p 24

What you will learn through your research, is that it is rare that variables are only dependent or independent, which is why outcome and predictor are preferred terms. Table 3.3.2 lists the variables that we will consider as our predictor and outcome variables for the demonstration. The original owners of this data can use a different set of predictor and outcome variables.

Table 3.3.2. Variables of interest for our demonstrations (James et al. 2020)

Predictor Variables	**Outcome Variables**
ageGroup - age group that children fall into **schoolYear** - UK school year/grade **ageMonths** - age in months **gender** - gender of participants **vocab** - Raw score from receptive vocabulary test (/36) **CTOPP_PE_raw** - CTOPP phoneme elision raw score **MA_IA -** Morphological awareness inflection analogy score (/24) **MA_DA** - Morphological awareness derivation analogy score (/20) **MA_CA** - Morphological awareness compound analogy score (/24) **MA_IJ** - Morphological awareness inflection judgement score (/24) **MA_DJ** - Morphological awareness derivation judgement score (/20) **MA_CJ** -Morphological awareness compound judgement score (/24) **MA_factor** - Factor score from PCA on all morphological awareness tasks	**YARC_comp_ab** - YARC reading comprehension ability score

In this video (14:03), we use our saved jamovi dataset dataset_postImp.omv (see section 1.2.) to demonstrate how to run descriptive statistics and generate graphs. We will also touch on how to interpret the results.

After watching the video once or twice, you are encouraged to follow along on jamovi yourself before attempting the next activity with a different dataset.

Additional Resources
Note: watching these videos is not compulsory. View the videos if there are still some topics you need a different perspective on, but remember that the concepts are also addressed in chapters 3 and 4 of Navarro and Foxcroft (2019)

- descriptive statistics (5:26)
- histograms (4:25)
- density plots (3:22)
- boxplots (3:25)
- violin plots (2:37)
- dot plots (3:06)
- barplots (3:27)

3.4. ACTIVITY: Inspecting data using numbers and graphs: Trilingual Vocabulary Data

Did you meet the objectives of this session? Let's see - follow the steps below and answer the questions.

1. Open **Trilingual Project Codebook .csv** in Excel (See section 0.3.3.2) and the **author's preprint**.
   1.1. According to the authors' preprint, what are the **predictor** variables in the study? How do you know?
   1.2. According to the authors' preprint, what are the **outcome** variables in the study? How do you know?
2. In jamovi, open matched_dataset.omv that you created in section 1.3.
   2.1. Create frequency tables for Lang_Group and Dom_Lang.
      2.1.1. Are there the same number of bi and tri linguals in this sample?
      2.1.2. What language is dominant for the most children?
   2.2. Leave Dom_Lang in the Variables box and remove Lang_Group. Split the sample by Lang_Group. Generate a frequency table.
      2.2.1. To what extent is the proportion of English dominant speakers similar for bilinguals and trilinguals in this sample?
   2.3. Using the matched_dataset.omv recreate Table 1, leaving Sex out, of Cote et al. (2020, p. 9) using the tools available in the Exploration tab in jamovi. i.e. generate the M and SD for the variables listed in the table except for Sex, and split by Lang_Group.
      2.3.1. Why are your results different from the authors'?
   2.4. Generate the mean, SD, and median of TotalNW_Eng, TotalNW_Fr and Total_Vocabulary, with histograms for the **whole sample (i.e. not split by** Lang_Group**).**
      2.4.1. How similar are the mean and median values for these variables?
      2.4.2. Is it a fair representation of your data to report the mean without a general comment on the distribution of the data?
      2.4.3. Do the patterns change if you split the sample by Lang_Group?
   2.5. How would you determine if there were outliers on the vocabulary measures for the bi and trilingual group? Mention one way to do this using a graph, and one way to do this using the numerical information.

3.5. DEMONSTRATION: Presenting a table of descriptive statistics

In this section, you'll learn about where the descriptive statistics table goes in a report/manuscript, why it's important and how to format the table for a manuscript/report.

- Why do we include a table of descriptive statistics?
    - I think you know the answer. In case you need a reminder, we need to present the descriptive statistics so that the readers of our work can also get a sense of what our data looks like.
- Where does the table of descriptive statistics go?
    - Usually at the start of the results section.
- What goes into the table of descriptive statistics?
    - Usually all the data for the variables you will use in your study.
- What about participant characteristics?
    - If you collected a lot of information about your participants, you can include a table of participant characteristics in the Participants section of your Method section (e.g. Table 1 in Schaefer & Kotze, 2019). However, most often, you can report the participant characteristics in the text only (e.g. see Participants section in Wilsenach, 2015).
- What do you say about the descriptive statistics in the text?
    - Because all the information is in the table, you only need to highlight any trends you want the reader to observe.

The tables that jamovi produces are very beautiful, but unfortunately need some revision before being pasted into your document. This video (8:53) shows how to format a table for your manuscript. Below is a summary of the steps to follow:

Step 1: Make your descriptive statistics table in jamovi.

Step 2: Copy the table from jamovi into a blank excel workbook. Select the option *match destination formatting* when you paste.

Step 3: Transpose your table by copying and selecting the paste option *transpose*.

Step 4: Relabel the table appropriately so that it has the group categories (where applicable) on the top row, followed by the sample size (N), means and standard deviations.

Step 5: Ensure that the mean and standard deviation for each group are together. Do this by cutting and inserting the standard deviation next to the corresponding mean. You can relabel standard deviation as SD.

Step 6: Merge your group categories and sample size columns.

Step 7: Delete all unnecessary labelling, information, and blank rows

Step 8: Format numbers to 2-decimal places

Step 9: Copy your table to your word document. Paste using the option *keep source formatting*

Step 10: Format the table - ensure columns are evenly spread and variables have descriptive labels, and include a bottom and top border to your table.

Step 11: Give your table a label by adding a caption. Table labels always go above the table.

Step 12: Add a note (key) to the bottom of your table if there is any information included that needs to be explained, such as opaque variable names.

Example 3.5.1. shows you how you would report the descriptive statistics in your manuscript/report. Example 3.5.1. shows that:

- the descriptive statistics table includes relevant predictor and outcome variables that will be addressed in the study
- the descriptive statistics table occurs early in the results section
- the table is referenced in the text, and has its own title which goes at the top of the table.
- the title of the table can stand alone and clearly tells the reader what is in the table
- the table is much easier to read than the one from jamovi; the variables are in rows and the statistics of interest per group are in the columns. It is useful to always report N in your table when working with data in groups.
- the variable names have been rewritten for the **reader** to better understand the tale. Where the variable names are too long for the table, the abbreviations are written out in a note at the bottom of the table.

~~~~~~

Example 3.5.1 - writing about descriptive statistics in your report/manuscript

Results

The means and standard deviations for each task and age group is presented in Table 1. Participants in the 6-8 year age group appear to have lower scores than the other age groups on all the variables. The 9-11 and 12 - 13 year olds appear to perform quite similarly on most variables. These group differences are further explored using a one way analysis of variance (ANOVA).

Table 1. Means and standard deviations for vocabulary, phoneme elision, morphological awareness and comprehension by age group

	6-8 years		**9-11 years**		**12-13 years**	
Variables	**N = 128**		**N = 146**		**N = 147**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Vocabulary	22.70	4.22	27.50	5.30	28.40	3.89
Phoneme Elision	21.40	6.26	26.90	4.99	28.30	3.89
MA_IA	15.30	4.47	18.50	3.23	18.70	3.51
MA_DA	9.62	3.99	13.00	3.00	14.20	2.91
MA_CA	18.40	5.18	21.80	2.74	22.50	2.77
MA_IJ	17.10	3.97	20.80	2.99	20.90	2.58
MA_DJ	10.70	4.01	15.50	2.84	15.70	2.32
MA_CJ	17.10	3.13	19.90	2.79	21.50	2.12
Comprehension	49.10	11.40	60.10	10.10	61.00	8.93

Note: MA_IA - Morphological awareness inflection analogy score; MA_DA - Morphological awareness derivation analogy score; MA_CA - Morphological awareness compound analogy score; MA_IJ - Morphological awareness inflection judgement score; MA_DJ - Morphological awareness derivation judgement score; MA_CJ - Morphological awareness compound judgement score

~~~~~~

3.6. ACTIVITY: Presenting a table of descriptive statistics: Trilingual Data

Present a table of descriptive statistics for the matched_dataset.omv file. Include separate columns for the Bilingual and Trilingual children and remember to specify the sample size. Present the Mean, and SD of the following variables for each group of children:

- Lang_Exp_Eng
- Lang_Exp_Fr
- Lang_Exp_Other
- Total_NW_Eng
- Total_NW_Fr
- Total_Vocabulary

Similar to example 3.5.1 (section 3.5.) present your table of descriptive statistics with a suitable title and write a sentence or two as if you were including the table in your report/manuscript.

> **Want to push yourself?** Include the 25th, 50th and 75th percentiles in your table as well. Make sure to update your one/two sentence description in the text if necessary.

4.0 Checking in: Live Session Friday 19 June 2020

Join us via Zoom for the live session [not available].

You may begin logging into the session at 14.00. We will start promptly at 14.15.

14.15-15.00 Activities

You need to have prepared the activities (1.3, 2.3, 3.3 and 3.6) for this week before the live session. Bring along your answers.

You also need to have skim read the following articles, paying particular attention to the presentation of the descriptive statistics.

Pretorius, E.J. & Stoffelsma L. (2017). How is their word knowledge growing? Exploring Grade 3 vocabulary in South African township schools. *South African Journal of Childhood Education*, 7(1), a553. https://doi.org/10.4102/sajce.v7i1.553

Schaefer, M. & Kotzé, J. (2019). Early reading skills related to Grade 1 English Second Language literacy in rural South African schools. *South African Journal of Childhood Education*, *9*(1), a644. https://doi.org/10.4102/sajce.v9i1.644

15.00 - 15.15 Break

15:15 - 16.00 Demonstration and practical activity

5.0. Inferential data analysis

In this section, we take you briefly through some statistical tests used for Null Hypothesis Significance Testing (NHST). NHST is briefly explained in this video (3:57).

The objectives are met when the participant can:

- explain common statistical terminology (null and alternative hypothesis, p value, effect size, statistical significance etc)
- choose appropriate inferential statistics for a research question
- run inferential statistics (correlation, regression, t-test and ANOVA) in jamovi
- interpret the results of correlation, regression, t-test and ANOVA
- report the results of correlation, regression, t-test and ANOVA

This section can seem overly technical. Remain calm.

Focus on understanding what the inferential statistics are **used for**, **how to run them in jamovi**, and **how to report and interpret the results.** There's no need to memorise any of the underlying mathematics.

Notes on this section:

- **Read through the course guide and watch all the prescribed videos** (we let you know when it's additional work). Just by doing this you should be able to
    - explain what a particular inferential test is used for,
    - run the inferential test in jamovi and
    - report and interpret the results.
- We list the relevant sections in Navarro and Foxcroft (2019) as an additional resource. **We do not expect you to read each chapter.** We include this information so you can turn to the book as a reference. The authors do a good job of including both the statistical theory and how the tests work in jamovi. If I were you, I would skip over anything I don't understand, and only look at how the author uses jamovi for each test.
- Additional videos are also included as an extra resource. You do not have to watch them.
- Throughout this section, we try to highlight new words (in yellow) that you'll come across. We try to include a lot of repetition so hopefully you will understand them a bit better after the course without having to explicitly study them. If you need to use any of these statistical tests for your project, you'll probably need to spend some extra work on understanding the terminology.

Section 5 should take at least 10 hours to complete.

5.1. CONCEPT: Inferential Statistics

The topic of inferential statistics is introduced here (10:41). The following topics are introduced:

- a description of what inferential statistics is
- hypothesis testing
- confidence intervals
- brief overview of correlation and regression
- parametric and nonparametric tests
- introduction to inferential statistical tests

Additional resources (not compulsory)

These concepts are addressed in:

Navarro & Foxcroft (2019) Chapter 8 | Estimating unknown quantities from a sample

Navarro & Foxcroft (2019) Chapter 9 | Hypothesis testing

5.2. T-tests

This section introduces you to t-tests which are used to compare two means. They include the one sample t-test, the paired samples t-test and the independent samples t-test, although we will focus on the latter two.

T-tests can be used to answer the research question,

To what extent does the height of giraffes and zebras differ?

Although our world knowledge tells us that giraffes will always be taller than zebras, a t-test allows us to be objective.

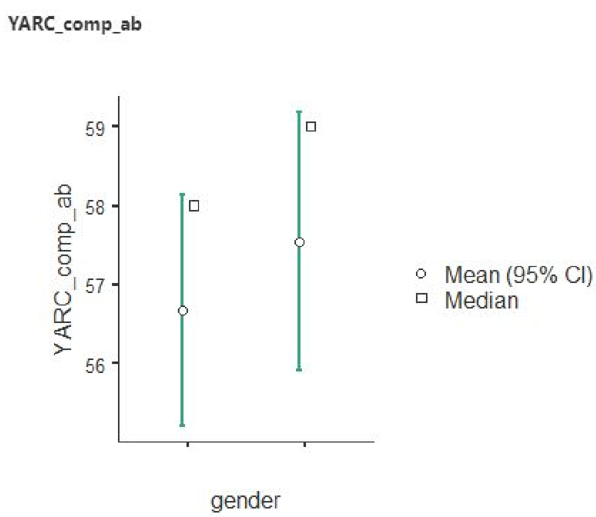
5.2.1. CONCEPT: T-tests

This [video](#) (8:03) explains what independent and paired samples t-tests are.

Additional resource (not compulsory)
These concepts are addressed in:

Navarro & Foxcroft (2019) Chapter 11 | **Comparing two means**

5.2.2. DEMONSTRATION: T-tests



Independent samples t-test (Student's or Welch's t-test): comparing the means of <u>two</u> different groups

In this demonstration we use an independent samples t-test to answer this research question:

> *To what extent does the reading comprehension raw score differ between boys and girls?*

From the graph (left) we can see that scores for boys and girls overlap a lot. But is this a significant difference (according to the principles of NHST)? If yes, is this a meaningful difference?

This demonstration video (4:24) shows how we run an independent t-test in jamovi to answer this research question. An independent samples t-test is used because the two groups of children (boys and girls) are independent from one another. First, we check the assumptions of an independent samples t-test. According to Navarro and Foxcroft (2019, p. 260) these are:

- normality of the distribution of data in each sample
- independent observations
- homogeneity of variance (homoscedasticity) i.e. the samples have the same SD → This assumption does not need to be met for Welch's test

In our previous videos we saw that comprehension follows a mostly normal distribution. We include the Descriptives to see the SD (to check the homogeneity of variance assumption). The SDs are very similar (girls = 11.2; boys = 11.6) and Levene's test shows us that the equality of variance assumption is met. This means we satisfy the assumptions for a Student's t-test. We run both types of independent t-tests for the demonstration.

Reporting the results (Navarro & Foxcroft, 2019, p. 258)
In the results section:

> The mean comprehension raw score for girls was 56.7 (SD = 11.2) and for boys was 57.5 (SD = 11.6). A Student's independent samples t-test showed that the mean difference of -.88 between girls and boys was not significant, t(419) = -.79, p = .432, 95%CI = [-3.07; 1.32], *d* = -.08.

#################

Independent samples t-test (Mann-Whitney U t-test): comparing the means of two different groups (non-parametric)

We will not provide a demonstration for the non-parametric Mann-Whitney U t-test for ordinal/ranked data. Consult Navarro and Foxcroft (2019, p. 277). You will have to self-study this topic if you come across a need to run the test.

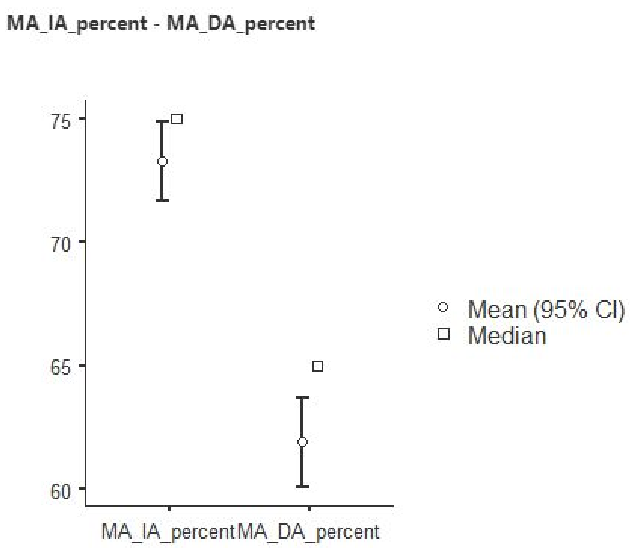
#################

Paired samples t-test: comparing the means of two variables for one group

In this demonstration we use a paired samples t-test to answer this research question:

> *To what extent do participants differ in their awareness of inflectional and derivational morphology?*

We use the analogy tasks to measure inflectional and derivational morphology. From the graph (left) we can see that scores for inflectional morphological awareness (MA) are higher than scores for derivational MA. But is this a significant difference (according to the principles of NHST)? If yes, is this a meaningful difference?



This demonstration video (4:01) shows how we run a paired samples t-test in jamovi to answer this research question. A paired samples t-test is used because the same group of participants completed the two tasks, MA_IA and MA_DA. Remember that the two tasks have different maximum scores so we converted the scores to percentages (watch this video (0:55) for how to compute scores yourself). The variables now have scores that can be compared. Each score is now out of 100%.

First, we check the assumptions of paired samples t-test. The paired samples t-test seems to only have the assumption that the **difference** in scores (between variable 1 and variable 2) are normally distributed (each variable does not need to be normally distributed per se). The Shapiro-Wilk test and the Q-Q plot show that the difference in scores between the inflectional and derivational tasks is normally distributed. We can therefore go ahead with the Student's paired samples t-test.

Reporting the results (Navarro & Foxcroft, 2019, p. 258)

> The mean for inflectional morphology was 73.3% (SD = 16.8%) and for derivational morphology was 61.9% (SD = 19%). A Student's paired samples t-test showed that the mean difference of 11.4% between these tasks was significant and represented a large effect size, t(420) = 15.8, p = < .001, 95%CI = [9.98; 12.8], d = .77.

#################

Paired samples (Wilcoxon rank) t-test: comparing the means of two variables for one group

We will not provide a demonstration for the non-parametric Wilcoxon rank t-test. Consult Navarro and Foxcroft (2019, p. 278). You will have to self-study this topic if you come across a need to run the test.

#################

One sample t-tests

We will not provide a demonstration for one sample t-test. Consult Navarro and Foxcroft (2019, chapter 11) and this video (7:07). You will have to self-study this topic if you come across a need to run the test.

#################

Additional Resources
Note: watching these videos is not compulsory. View the videos if there are still some topics you need a different perspective on, but remember that the concepts are also addressed in chapter 11 of Navarro and Foxcroft (2019).

- Step-by-step instructions to run an independent samples t-test
- video (6:39): independent samples t-test
- Step-by-step instructions to run a paired samples t-test
- video (5:12): paired samples t-test
- Step-by-step instructions to run a one sample t-test

5.2.3. ACTIVITY: T-tests

Use the data in matched_dataset.omv that you created in section 1.3. Below are some research questions. For each research question:

1. identify the variables of interest
2. identify the appropriate t-test to run to help answer the research question
3. check the assumptions of the t-test selected above and revise your choice if necessary
4. provide at least one reason for why you chose the t-test you did
5. run the appropriate t-test
6. report the results in a short paragraph

Question 1
Do the mothers of boys and girls have similar education levels?

Question 2
To what extent does English and French vocabulary size differ for the whole sample (bi- and trilingual participants together)?

Question 3
To what extent do bi- and trilingual children in this sample differ in their French vocabulary size?

5.3. Correlation

This section introduces you to correlation which is used to determine the **strength** and **direction** of a relationship between two variables. Correlations include Pearson's (most common - for linear relationships), Spearman's (rank order correlations) and Kendall's tau b but we focus on the first two in this section.

5.3.1. CONCEPT: correlation

This video (2:18) explains correlations: what they are, how they are interpreted and provides an example of a correlation matrix.

For fun

Ah! beware the spurious correlation (e.g. Figure 5.3.1.) - when two variables seem to be associated with one another, but their relationship is due to chance or the influence of another unaccounted for variable. See more spurious correlations here.

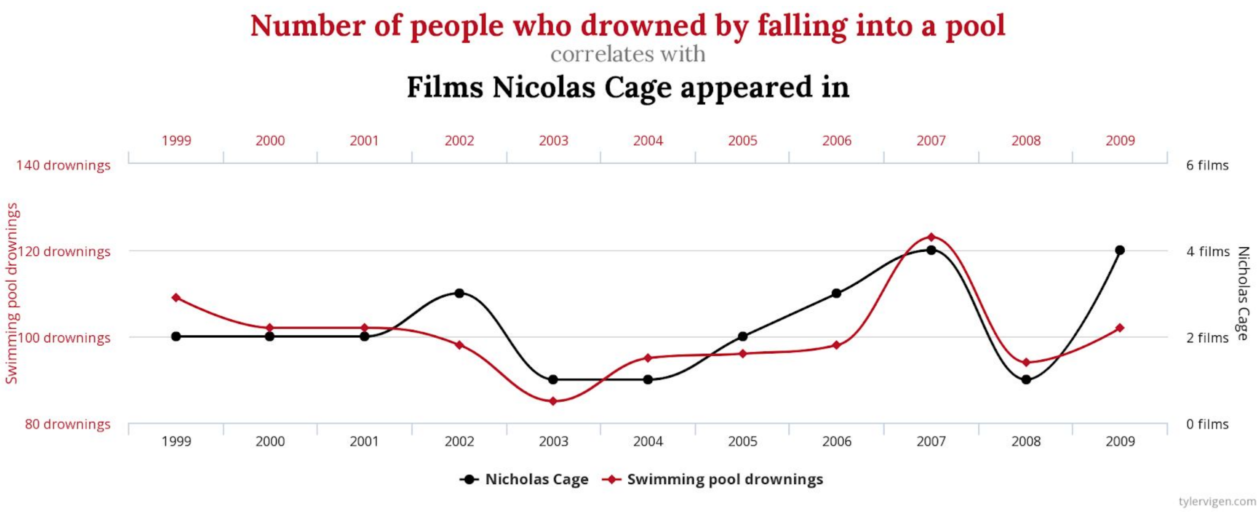


Figure 5.3.1. The correlation between the number of people who drowned by falling into a pool and the number of films NIcholas Cage has appeared in
Source: tylervigen.com

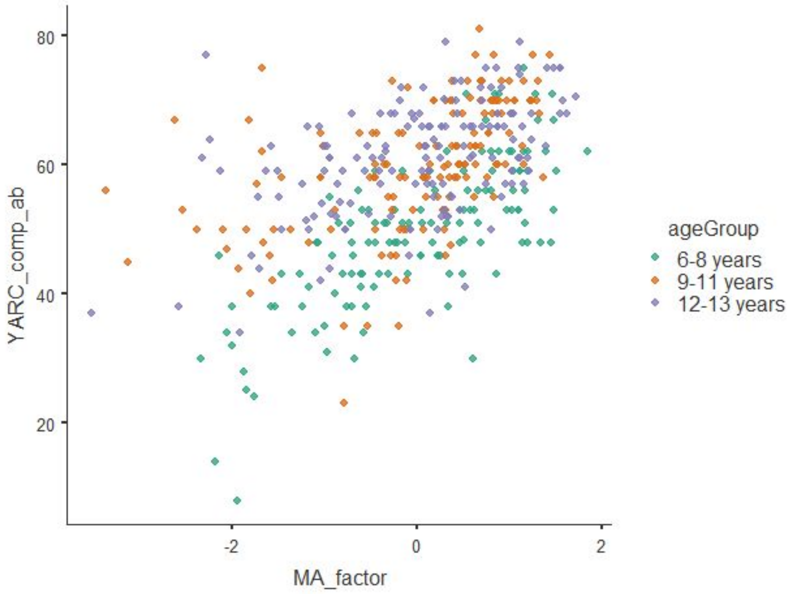
Additional resources (not compulsory)

This topic is addressed in:

Navarro & Foxcroft (2019) Chapter 12 | Correlation and linear regression

---

5.3.2. DEMONSTRATION: correlation



In most research reports or manuscripts, researchers are interested in the relationship between variables. **Researchers use correlation to determine the strength and direction of the relationship between two variables.** This relationship can be represented using a scatterplot (see figure left).

Strength and direction of a linear relationship: Pearson's r

In the James et al. (2020) dataset we are primarily interested in the relationship between morphological awareness and comprehension. In the figure above, we visualise the data using a scatterplot to determine if there is a linear relationship between the two variables. Usually, the predictor variable is placed on the x axis (horizontal axis) and the outcome variable is placed on the y axis (vertical axis). The figure shows us that there is a linear relationship between the variables so it is acceptable to use a linear correlation. Usually, researchers report the correlations between all their variables in one table (called a correlation matrix), and then mention some of the significant trends in the text. We demonstrate how to run the correlations for the variables of interest in James et al. (2020) here (3:03).

This video (5.08) shows how to format your correlation matrix for your manuscript.

Reporting the results

Table 5.3.1 shows that comprehension has a weak to moderate relationship with all other variables, r (421) = .387 - .567, p <.001.

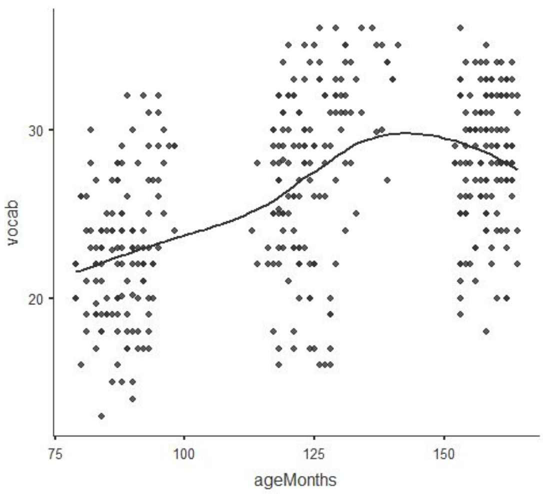
Table 5.3.1. Pearsons Correlation Coefficients for vocabulary, phoneme elision, morphological awareness and comprehension

	1	2	3	4	5	6	7	8	9
1. Vocabulary	—								
2. Phoneme Elision	0.441*	—							
3. MA_IA	0.445*	0.541*	—						
4. MA_DA	0.505*	0.604*	0.662*	—					
5. MA_CA	0.442*	0.524*	0.623*	0.613*	—				
6. MA_IJ	0.57*	0.578*	0.618*	0.599*	0.588*	—			
7. MA_DJ	0.648*	0.595*	0.606*	0.607*	0.552*	0.745*	—		
8. MA_CJ	0.583*	0.525*	0.524*	0.537*	0.535*	0.619*	0.635*	—	
9. MA_factor	0.465*	0.469*	0.734*	0.621*	0.626*	0.7*	0.608*	0.566*	—
10. Comprehension	0.472*	0.33*	0.435*	0.387*	0.409*	0.358*	0.351*	0.371*	0.567*

Note: All correlations are significant at p < .001
MA_IA - Morphological awareness inflection analogy score; MA_DA - Morphological awareness derivation analogy score; MA_CA - Morphological awareness compound analogy score; MA_IJ - Morphological awareness inflection judgement score; MA_DJ - Morphological awareness derivation judgement score; MA_CJ - Morphological awareness compound judgement score

*Did you notice how we didn't report an effect size? Well, actually, we did. It's just that the effect size in this case is Pearson's r.

#################



Strength and direction of a rank ordered relationship: Spearman's ρ

When the data are related (as one variable increases, the other increases or decreases) but do not have a linear relationship, one can use Spearman's rank order correlation. The test statistic we report is ρ (pronounced rho). The statistic can be interpreted the same as Pearson's r (Navarro & Foxcroft, 2019, p. 288).

To the left is a scatterplot of age and vocabulary, with a smoothed non-linear trend line. There's a general trend that as age increases, vocabulary increases. The relationship does not look linear, however. Because the data is split up into the three age groups, it is probably best to use a Spearman correlation to determine the relationship between age and the other variables. If age has a significant and meaningful relationship with the other variables we know we need to control for age in our other analyses.

In jamovi, run a Spearman's rank order correlation by going to Analyses > Regression > Correlation Matrix. Check the box next to Spearman. Include age, and our other variables of interest (vocabulary, phoneme elision, morphological awareness and comprehension). For this walk-through, I also clicked Pearson's correlation to see if there would be any differences.

Reporting the results

The results show that age has a moderate positive relationship with all the variables (all $\rho > .4$, $p < .001$), except for the inflectional analogy task where the relationship is weak ($\rho = .34$, $p < .001$) (Table 5.3.2.1).

Table 5.3.2.1. Spearman's rank order correlations between age and the other variables

	Age
Vocabulary	0.479
Comprehension	0.469
Phoneme Elision	0.435
MA_IA	0.338
MA_DA	0.460
MA_CA	0.438
MA_IJ	0.403
MA_DJ	0.492
MA_CJ	0.581

Note: all correlations significant at the p < .001 level
MA_IA - Morphological awareness inflection analogy score; MA_DA - Morphological awareness derivation analogy score; MA_CA - Morphological awareness compound analogy score; MA_IJ - Morphological awareness inflection judgement score; MA_DJ - Morphological awareness derivation judgement score; MA_CJ - Morphological awareness compound judgement score

Additional Resources

Note: using these resources is not compulsory. View the videos if there are still some topics you need a different perspective on, but remember that the concepts are also addressed in chapter 12 of Navarro and Foxcroft (2019).

- video (7:31): Pearson's correlation in jamovi
- video (6:53): correlation coefficient explained in three steps

5.3.3. ACTIVITY: correlation

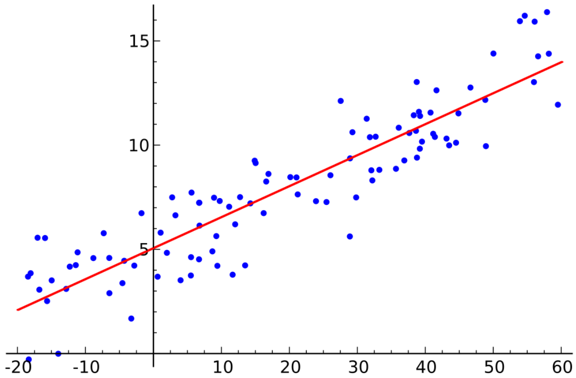
1. **Create a correlation matrix** using Pearson's r for the variables of interest in matched_dataset.omv Include the following variables for the whole sample:
    - Lang_Exp_Eng
    - Lang_Exp_Fr
    - Lang_Exp_Other
    - Total_NW_Eng
    - Total_NW_Fr
    - Total_Vocabulary
2. Format the table for a report and paste it in your Word document. Add an appropriate title to the table.
3. Write a sentence or two summarising the main trends in the correlation matrix, also referring to the table in your text. *Remember to look at things such as the strength and direction of the relationship between variables.*

#################

For fun

Test how good you are at estimating the Pearson r correlation coefficient from looking at a scatterplot in the fun game of Guess the Correlation. Let us know your highest score in the #random channel on Slack.

5.4. Linear Regression



This section introduces you to *linear regression* (Ordinary Least Squares Regression or OLS) which is used to determine the **strength** and **direction** of a **linear** relationship between two variables while also accounting for the influence of other variables. Linear regression models a line of best fit in your data. The model is interpreted by comparing the predicted values from the model, to the actual values in your dataset. Models can usually never be perfect, so we use a range of tools to determine if the model is good enough.

5.4.1. CONCEPT: Linear Regression

This video (3:59) addresses the concepts related to linear regression analysis. These concepts include:

- outcome (dependent) and predictor (independent) variables
- using regression for prediction and explanation
- terms used in linear regression
    - model fit (R^2)
    - model coefficients (standardised estimate β)
    - covariates
    - model assumptions

These concepts will become clearer in the demonstrations.

Additional resource (not compulsory)

Linear regression is addressed in:

Navarro & Foxcroft (2019) Chapter 12 | Correlation and linear regression

---

5.4.2. DEMONSTRATION: Linear Regression

Using the James et al. (2020) dataset, we want to answer the following research question:

What is the relationship between morphological awareness and reading comprehension?

From earlier in our course we know that in this research question, morphological awareness (MA) is the predictor (or independent) variable and comprehension is the outcome (or dependent) variable.

If you think about it, we've answered this question before, when we did a correlation analysis in section 5.3.2. There was a significant positive and moderate relationship between MA and comprehension ($r = .52$, $p < .001$). Correlation just tells us that the variables are associated. Perhaps the relationship between MA and comprehension is due to both variables being influenced by a third variable? For example, vocabulary and comprehension were moderately and positively associated with one another too ($r = .651$, $p < .001$), and so was vocabulary and MA ($r = .465$, $p < .001$). Clearly, we need a way to be more precise in our understanding of the relationship between MA and comprehension. One way to account for the influence of other variables, is to use linear regression (also called ordinary least squares regression) to control for the influence of other variables.

We'll keep the demonstration simple and allow you to expand your own knowledge on linear regression. We recommend you watch the videos first and then read steps 1 - 3 for reference when completing the activities.

- Watch the demonstration video (3:44) that shows you where to find the linear regression options, and which options to select for our example. This video combines information in steps 1 and 2.
- We have included a separate video (6:13) specifically talking about assumption checking for this model (in step 2).
- This video (6:33) explains how to report and interpret the results of the linear regression (step 3).

Step 1 - think about a theoretically suitable model

The first step is to **use your theoretical knowledge to identify a suitable model**. After all, we are linguists using statistics as a tool to solve problems with our discipline knowledge.

In this demonstration, we use YARC_comp_stand (standardised score) as the outcome (dependent) variable. So far we have been using YARC_comp_ab (a raw score) in all our examples. We switch to the standardized score because:

- age is positively correlated with all variables and should therefore be accounted for in the regression. Instead of adding age as a predictor we can use the comprehension scores standardized by age for all participants. This allows us to control for age without making the model more complex.
- the raw score led to the residuals of the model not being normally distributed. When we use the standardised score our residuals become normally distributed, meeting the assumptions of the test

We use vocab as a control variable. It's possible that the relationship between MA and comprehension is explained by vocabulary levels so we want to make sure we can say something about this relationship.

We use the MA_factor as the morphological awareness score. This score derives from a Principal Components Analysis of all six MA tasks. The shared variance of all six tasks is captured in this one score. This allows us to keep the model simple by including only one variable that best describes the MA ability of participants.

#################

Step 2 - set up the linear regression in jamovi

The second step is to **run the linear regression in jamovi choosing the relevant options for output**. In this case, we actually have a multiple linear regression because there is more than one predictor (called a covariate in jamovi). I'll go over each option offered in jamovi, let you know what each option is for, and tell you what to select for this demonstration.

--- **Model Builder**

Used to specify the blocks of a hierarchical regression. See Navarro and Foxcroft (2019, p. 323). Leave as is for the demonstration.

--- **Reference Levels**

Used to specify which level of a nominal variable should be the default (reference) in a comparison. For example, when including sex in the model, you can specify whether boys should be the default and then the coefficient refers to how much better or worse the girls do. We do not have nominal variables in our model so leave as is.

--- **Assumption Checks**

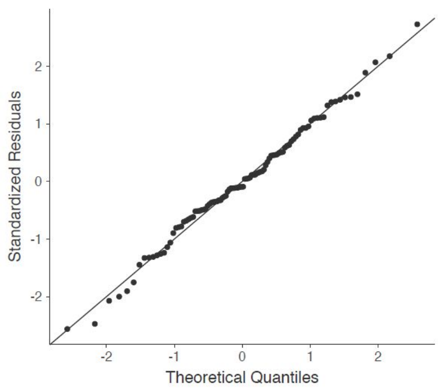
For this demonstration, select all the options except for the Autocorrelation test.

Navarro and Foxcroft (2019, pp. 309 - 310) list the six assumptions of linear regressions which are listed in Table 5.4.2.1. I have reordered the assumptions to more closely follow the order in which you would check them. The Table also mentions how you would check each assumption in jamovi. In addition to Navarro and Foxcroft (2019), I also used the helpful guide by Statistics Solutions (2020) to fill in the table. Don't be alarmed if you feel overwhelmed. Things will become clearer to you the more you practice and work in the program.

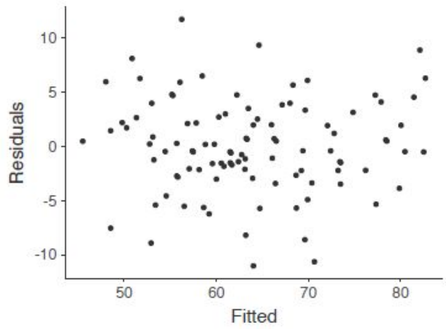
Table 5.4.2.1. Checking the assumptions of a linear regression model in jamovi

Assumption of linear regression	**Where to find this information in jamovi**	**What are we looking for?**
1. linearity of the the relationship with outcome variable	Regression > Correlation Matrix > Plot > Correlation matrix OR Add scatr module; Exploration > Scatterplot	the predictor variables should be linearly related to the outcome variable i.e. you should be able to draw a straight line through the data

<table>
<tr><td></td><td>Regression > Correlation Matrix > Correlation Coefficients > Pearson</td><td>relationship between predictor and outcome variables is significant and at least weak; direction of relationship affects interpretation only
<table>
<tr><th>Correlation</th><th>Strength</th><th>Direction</th></tr>
<tr><td>-1.0 to -0.9</td><td>Very strong</td><td>Negative</td></tr>
<tr><td>-0.9 to -0.7</td><td>Strong</td><td>Negative</td></tr>
<tr><td>-0.7 to -0.4</td><td>Moderate</td><td>Negative</td></tr>
<tr><td>-0.4 to -0.2</td><td>Weak</td><td>Negative</td></tr>
<tr><td>-0.2 to 0</td><td>Negligible</td><td>Negative</td></tr>
<tr><td>0 to 0.2</td><td>Negligible</td><td>Positive</td></tr>
<tr><td>0.2 to 0.4</td><td>Weak</td><td>Positive</td></tr>
<tr><td>0.4 to 0.7</td><td>Moderate</td><td>Positive</td></tr>
<tr><td>0.7 to 0.9</td><td>Strong</td><td>Positive</td></tr>
<tr><td>0.9 to 1.0</td><td>Very strong</td><td>Positive</td></tr>
</table>
(Navarro & Foxcroft, 2019, p. 288)</td></tr>
<tr><td rowspan="3">2. uncorrelated predictors

(for multiple regression only i.e. more than one predictor variable)</td><td>Regression > Correlation Matrix > Plot > Correlation matrix</td><td>avoid including predictors with r >= .8 (Statistics Solutions, 2020)</td></tr>
<tr><td>Regression > Linear Regression > Assumption Checks > Assumption Checks > Collinearity statistics</td><td>variance inflation factor values (VIF) should be less than 10 but preferably less than 5 (Statistics Solutions, 2020)</td></tr>
<tr><td>Regression > Linear Regression > Assumption Checks > Assumption Checks > Autocorrelation test</td><td>not addressed by Navarro and Foxcroft (2019) so we will leave this up to self-study. You should be able to assess multicollinearity using the previous two approaches</td></tr>
<tr><td rowspan="2">3. normality of residuals</td><td>Regression > Linear Regression > Assumption Checks > Assumption Checks > Shapiro-Wilk</td><td>normally distributed if Shapiro-Wilk p value > .05</td></tr>
<tr><td>Regression > Linear Regression > Assumption Checks > Assumption Checks > Q-Q plot of residuals</td><td>you want values close to the normal distribution indicated by the straight line


(Navarro & Foxcroft, 2019, p. 317)</td></tr>
</table>

4. homogeneity of variance (homoscedasticity)	Regression > Linear Regression > Assumption Checks > Assumption Checks > Residual plots	in the fitted vs residuals plot, and others generated from this option, there should be no pattern or bunching of the dots i.e. the dots should look randomly placed in the graph (like in the graph below)  (Navarro & Foxcroft, 2019, p. 318)
5. residuals independent of one another		
6. no influencer data points i.e. outliers with high leverage	Regression > Linear Regression > Assumption Checks > Data Summary > Cook's distance	aim for Cook's distance < 1 values > 1 indicate influential data points that need to be explored further (Navarro & Foxcroft, 2019, p. 315)

--- **Model Fit**

This section allows us to select what statistics to include to assess the model fit i.e. how well our equation explains the real data.

Select R, R^2, Adjusted R^2 and F test and leave the other options unchecked.

R^2 (coefficient of determination) tells you "the *proportion* of the variance in the outcome variables that can be accounted for by the predictor" (Navarro & Foxcroft, 2019, p. 301). It tells you how well the model fits the data and is the effect size for linear regression. R^2 ranges from 0 to 1. 0 indicates that the predictor explains no variance in the outcome variable. 1 indicates that the predictor explains all the variance in the outcome variable. Because R^2 is a proportion, we can also speak about this statistic as a percentage, i.e. the predictor explains x percentage of variance in the outcome variable. These values should be interpreted within your discipline.

Adjusted R^2 is used in multiple regression to correct for the inclusion of additional variables in your model. Because of how the regression works, adding more variables will always increase R^2, and this can give you the false impression that your added variable explains more variance. Adjusted R^2 does not always have a simple interpretation as R^2 so usually just report R^2.

The F-test tests the whole model to see if it "[performs] better than chance (Navarro & Foxcroft, 2019, p. 304). **The null hypothesis for this test is that there is no relationship between the variables.** If you get a large F value, and p < .05 (i.e. significant) then you have evidence to reject the null hypothesis. This means that your F test shows that there **is** a relationship between the variables.

AIC, BIC, and RMSE are other measures of the goodness of fit of a model. We will not explore these now.

--- **Model Coefficients**

Under Estimate, select Confidence Interval (leave at 95%). Under Standardized Estimate, select Standardized estimate and Confidence Interval (leave at 95%).

The output for each predictor is a regression coefficient (Estimate in the table called Model Coefficients), e.g. the model coefficient for MA_factor is 4.9. This means that keeping all other values the same, a one unit increase in MA_factor, leads to 4.9 unit increases in the standardized comprehension score. The model coefficient is interpreted in its units, e.g. if the predictor coefficient is 2, and if your predictor is measured in percentages, and your outcome is measured in rands then you interpret the coefficient as one percentage increase in [predictor] leads to a R2 increase in [outcome].

This coefficient is tested with a t-test (interpret in same way as section 5.2.). The t-test tests the hypothesis that the coefficient is zero. An extreme t value with $p < .05$ tells you that the predictor does influence the outcome variable.

The standardized estimate (also referred to as β or beta) helps us interpret the model coefficients without referring to the original units of the variables. Navarro and Foxcroft (2019) explain how jamovi does this. What's important for you to know is that the standardized estimate reports standard deviation units:

> *Regardless of what the original variables were, a β value of 1 means that an increase in the predictor of 1 standard deviation will produce a corresponding 1 standard deviation increase in the outcome variable. Therefore, if variable A has a larger absolute value of than variable B, it is deemed to have a stronger relationship with the outcome. Or at least that's the idea. It's worth being a little cautious here, since this does rely very heavily on the assumption that 'a 1 standard deviation change' is fundamentally the same kind of thing for all variables. It's not always obvious that this is true. when the units of the variables are different.* (Navarro & Foxcroft, 2019, p. 308).

In our output, we see that morphological awareness ($\beta = .44$) makes almost double the contribution to reading comprehension than vocabulary ($\beta = .27$).

--- **Estimated Marginal Means**

This section allows you to estimate the mean of a variable when controlling for other variables. We will not be doing this in our demonstration so leave it as is.

#################

Step 3 - one way to report the results

There are different ways to report the results of a regression analysis. You can always follow the format of another author. Here we give you one option to report your results.

What information should be reported from a linear regression? Depending on the complexity of your model, some of the information can be reported in the text, and some of the information can be reported in the table. The details to include in your report and where to include them are as follows:

- the effect size (R^2) (table or text)
- the model coefficients (unstandardised, standard error of the unstandardized coefficient, and the standardized coefficient) (usually in a table)
- the F statistic and p value for the whole model (table or text)
- the t-statistic and p value for the coefficients (table or text, but usually at least p values in the table)

Example 5.4.2.1. demonstrates how to report the results of the linear regression in text and in a table. We provide the interpretation as part of the video. The interpretation of the results is included in the Discussion section of your report or manuscript. The interpretation requires you to integrate the results with what has been reported in the literature on the subject.

~~~~~~

Example 5.4.2.1 Reporting the results of a linear regression

Results of the multiple linear regression (Table 5.4.2.2.) indicated that vocabulary and morphological awareness had a significant effect on reading comprehension, F(2, 418) = 127, p < .001. Together they explained 37.7% of the variance. Both vocabulary (t= 6.11, p < .001) and morphological awareness (t = 10.17, p < .001) were significant predictors. Morphological awareness (β = .358) made a larger contribution than vocabulary (β = .266).

Table 5.4.2.2. Model coefficients for vocabulary and morphological awareness predicting reading comprehension

				95% Confidence Interval			
Predictor	B	SE	β	Lower	Upper	t	p
Intercept	88.026	2.5318				34.77	< .001
Vocabulary	0.578	0.0947	0.266	0.181	0.352	6.11	< .001
Morphological awareness	4.948	0.4863	0.444	0.358	0.529	10.17	< .001

################

Another way to report the results - graph of model coefficients

If many predictors are included in the model, a better way to report the coefficients is through the use of a graph. Figure 5.4.3.1. (Murray et al., 2020) presents the model coefficients of a linear regression used to determine the predictors of student evaluations of lecturers. The authors find that being attractive, being male, having a course interesting to students, having a Humanities course, being a full professor, and being white were positively associated with a course rating i.e. the higher these factors, the better the course rating. We can see this because the model coefficients are all greater than zero.
~~~~~~

Having a difficult course, being a teaching assistant, and having an accent were negatively associated with course ratings i.e. these factors lead to lower evaluations after controlling for other effects. We can see this because the model coefficients are all lower than zero. The authors found no relationship between the quality of teaching and research performance.

The study provides evidence of continued bias in lecturer evaluations, highlighting the challenges that female academics and academics of colour continue to face in their professional careers.

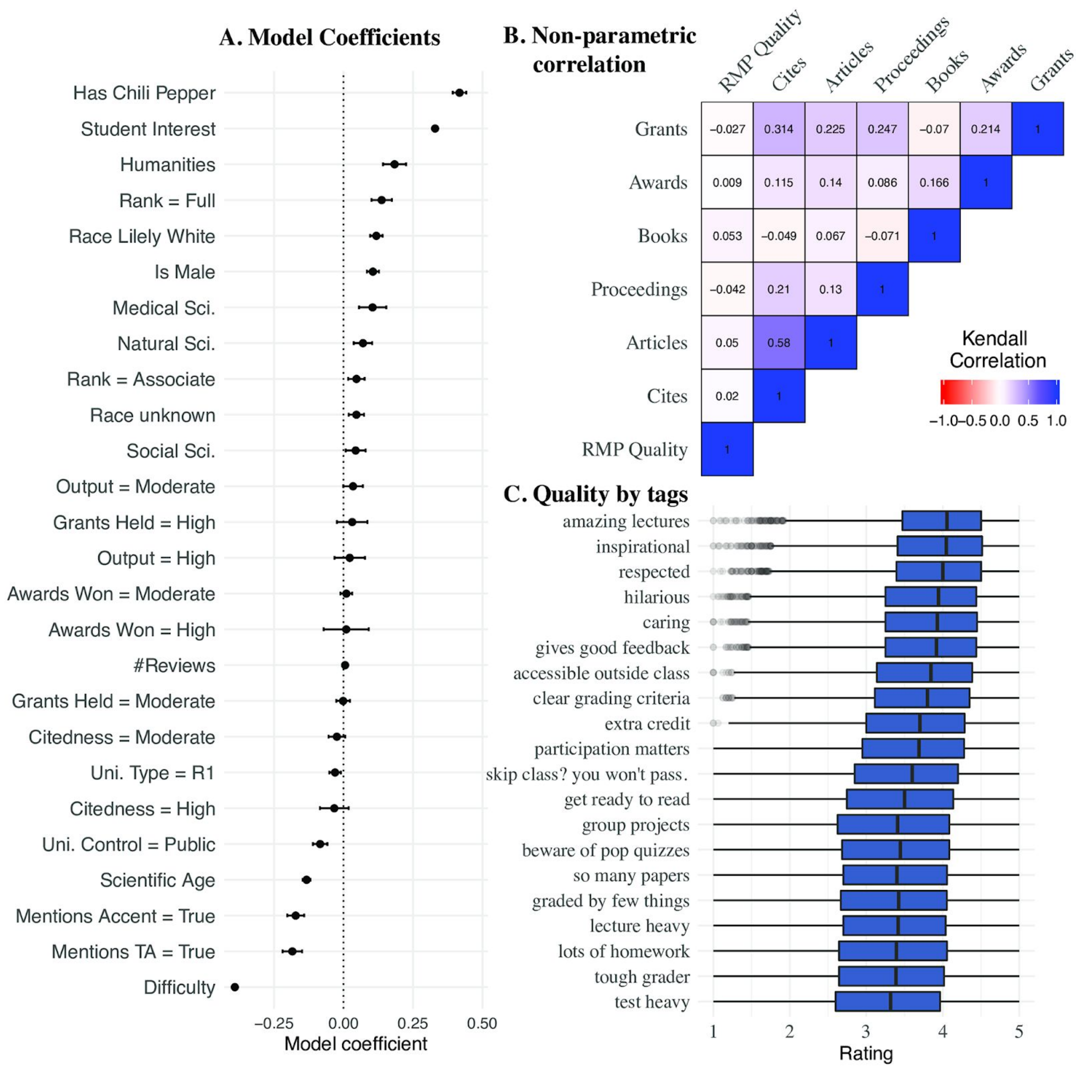


Figure 5.4.3.1. Individual, classroom, university and research characteristics associated with overall teaching quality.

A. Estimates of linear regression model using the overall teaching quality (continuous, 1-5) as the response and all variables from Table 1 as the predictor variables. The x-axis corresponds to the estimate for each covariate, which are listed along the y-axis. For binary variables, "false" is always used as the reference level. For Gender, "female" is used as the reference. For race, "Non-White" is used as the reference. For "Rank", "Assistant" is used as the reference. For Discipline, "Engineering" is set as the reference. For Uni. Control, "Private" is used as the reference. For Uni. Type, "Not R1" is used as the reference. For all research indicators, "Low" is used as the reference. Error bars surrounding each point correspond to the 95th percentile confidence intervals. Results are also shown in S4 Table. B. The non-parametric Kendall Rank Tau between research indicators and overall teaching quality. Values map to the correlation between 1 (correlated) and -1 (inversely correlated). Raw values for this test can be found in S5 Table. C. The distribution of overall teaching quality ratings for faculty possessing each of the pre-defined "tags" listed on their RateMyProfessor.com profile.

Source: [Murray et al. (2020, p. 8)](). CC 4.0.

Additional resource (not compulsory)

- video (6:14): linear regression
- video (6:20): assumption checks of linear regression
- video (7:15): hierarchical regression analysis
- step-by-step guide to run a linear regression in jamovi

5.4.3. ACTIVITY: Linear Regression

1. Use a **linear regression analysis** to answer the following research question:

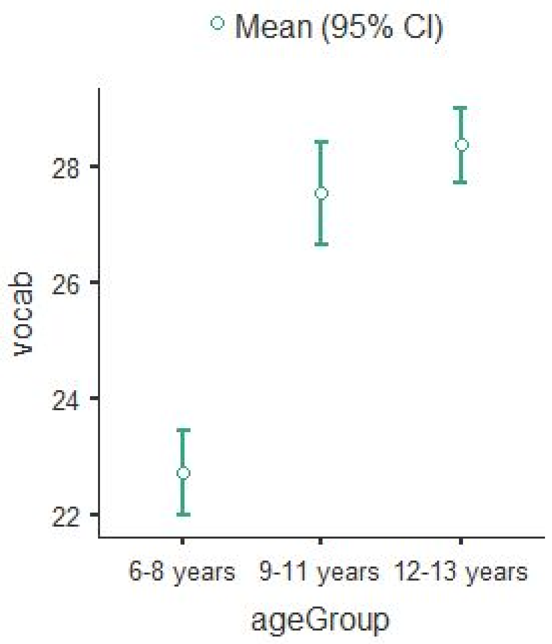
    To what extent is the english vocabulary size of multilingual children influenced by french vocabulary size, after taking language exposure in both languages into account?

2. **Report** the results of the multiple linear regression using text and a table.
3. Provide an **interpretation** of the results under a heading called Discussion

Further instructions and hints:

- Use matched_dataset.omv
- before beginning, identify your theoretical model (step 1). What are the predictor variables? What is the outcome variable?
- Specify the necessary output (step 2)
- check the assumptions of the model (step 2). (You already addressed some of the assumptions in earlier activities). If the assumptions are met, move to step 3.
- report the results (step 3)
- **Push yourself:** are there other variables with an influence on the predictor and outcome variables you should consider including in the model? How does the inclusion of another covariate change the results? Do the assumptions still hold?

5.5. Analysis of Variance (ANOVA)



This section introduces you to analysis of variance (ANOVA) which is used to compare several means. You'll find that there is some overlap between a t-test and ANOVA. The t-test is used to compare the means of two groups, and the ANOVA is used to compare the means of more than two groups.

ANOVA helps us test the null hypothesis that the group means are identical. A significant result shows us that at least one group is different from the others. Post hoc tests are used to see which group or groups has a statistically significantly different mean from each of the other groups..

5.5.1. CONCEPT: ANOVA

This video explains what an ANOVA is. The video also mentions different types of analysis of variance including repeated measures ANOVA, analysis of covariance (ANCOVA) and multiple analysis of variance (MANOVA). In this course, we address only one-way ANOVA and encourage you to self-study the other options using Navarro and Foxcroft (2019).

Additional Resource (not compulsory)

One-way ANOVA is addressed in:

> **Navarro & Foxcroft (2019) Chapter 13 | Comparing several means (one-way ANOVA)**

5.5.2. DEMONSTRATION: ANOVA

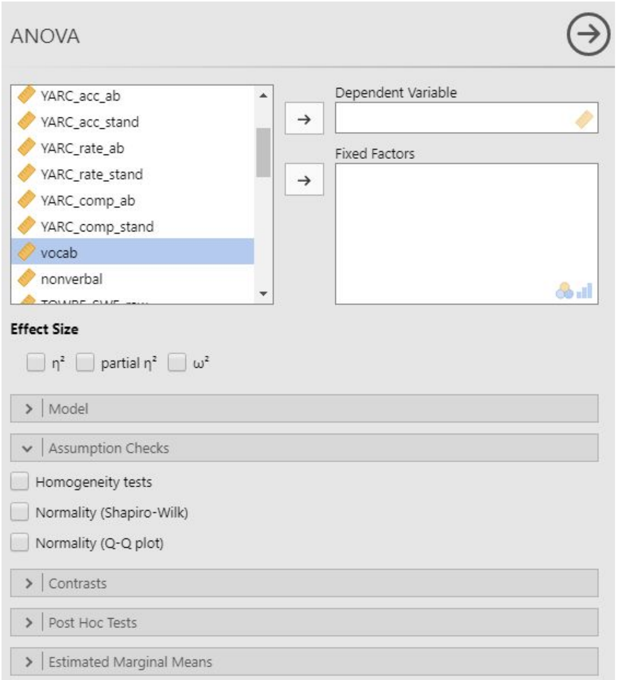
Using the James et al. (2020) data, we want to answer this research question:

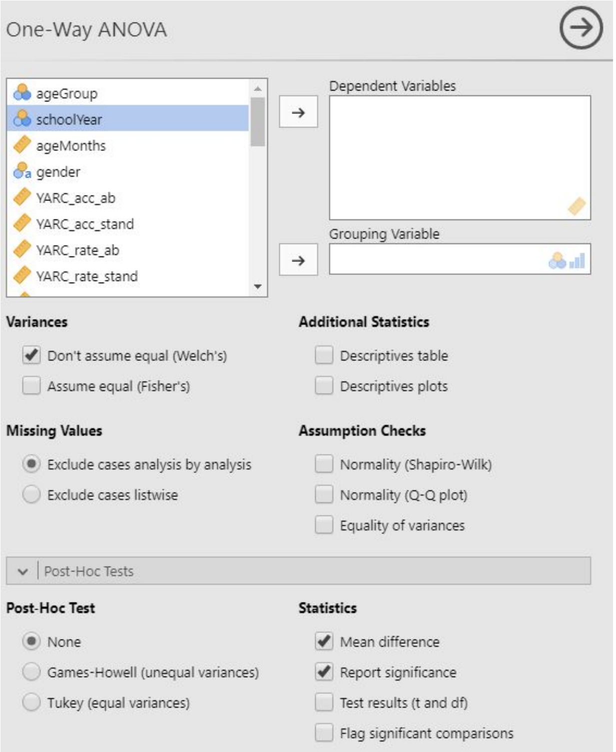
> *To what extent does reading comprehension differ by age group?*

We demonstrate how to set up the ANOVA analysis in jamovi in this video. We use YARC_comp_ab as the outcome variable. ageGroup is the predictor variable.

In jamovi, there are two ways to run a one-way ANOVA. Table 5.5.2.1. outlines the two approaches and their major advantages and disadvantages. You will most likely need to use both options to find the relevant information you need to report.

Table 5.5.2.1. Where to find a one-way ANOVA in jamovi and why you should select one option over the other

Menu in jamovi	**Why you would and would not use this option**
ANOVA > ANOVA 	The benefits of this option are that you can: ● determine effect size (eta squared, partial eta squared and omega squared) ● include more than one nominal variable ● specify more post-hoc tests (no correction, Tukey, Scheffe, Bonferroni and Holm) However, this option assumes that the homogeneity of variance assumption is met i.e. you cannot specify Welch's ANOVA.
ANOVA > One-Way ANOVA	The benefits of this option are that you can: ● specify Welch's ANOVA if the homogeneity of variance assumption is violated ● create graphs of the group differences (descriptives plots) ● specify whether the t-tests should or should not assume equal variances The major drawback here is that you cannot get the effect size. You can also only use one predictor variable.



To start with, we will use the ANOVA > ANOVA approach because it provides an effect size. Below I outline the options in ANOVA > ANOVA and specify what you should select for the demonstration. Watch the demonstration video (5:43) here.

Move the variables into the relevant blocks. YARC_comp_ab must be placed in the Dependent Variable block. ageGroup should be placed in the Fixed Factors block.

--- **Effect Size**

Select the effect sizes eta squared (η^2) and omega squared (ω^2). These tell you how much variance in the outcome variable is accounted for by the predictor variable. Because we only have one predictor, we do not need to select partial eta squared (partial η^2).

--- **Model**

Here you can specify the interaction terms if you have more than one factor in the model. Leave as is because we only have one predictor.

--- **Assumption Checks**

Select all options.

--- **Contrasts**

Leave as is. You can self-study this topic.

--- **Post-Hoc Tests**

We tell the program which variable it should use to examine group differences. Move ageGroup into the block on the right. Select No Correction, Bonferroni and Holm. Under Effect size, check the block for Cohen's d.

--- **Estimated Marginal Means**

This option allows you to see the group means after controlling for another variable. Leave as is because we only have one predictor.

As always, before interpreting the model results, you need to check if the assumptions of the test have been met (Table 5.5.2.2.).

Table 5.5.2.2. Assumption checks for the ANOVA

<table>
<tr><th>Assumption of linear regression</th><th>Where to find this information in jamovi</th><th>What are we looking for?</th></tr>
<tr><td>1. homogeneity of variance: population SD same for all groups</td><td>ANOVA > One-Way ANOVA > Assumption Checks > Equality of variances

or

ANOVA > ANOVA > Assumption Checks > Homogeneity tests</td><td>variance is homogenous if Levene's test p > .05

if p < .05 then use Welch's ANOVA (specified in ANOVA > One-Way ANOVA > Variances > Don't assume equal (Welch's)</td></tr>
<tr><td rowspan="2">2. normality of the residuals</td><td>ANOVA > One-Way ANOVA > Assumption Checks > Normality (Shapiro-Wilk)

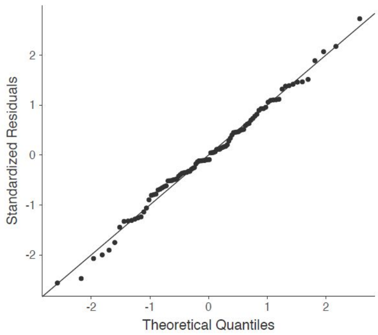
or

ANOVA > ANOVA > Assumption Checks > Normality (Shapiro-Wilk)</td><td>normally distributed if Shapiro-Wilk p value > .05

if not normally distributed use Kruskal-Wallis ANOVA test</td></tr>
<tr><td>ANOVA > One-Way ANOVA > Assumption Checks > Normality (Q-Q plot)

or

ANOVA > ANOVA > Assumption</td><td>you want values close to the normal distribution indicated by the straight line</td></tr>
</table>

<table>
<tr><td></td><td>Checks > Normality (Q-Q plot)</td><td>

(Navarro & Foxcroft, 2019, p. 317)</td></tr>
<tr><td>3. independence of observations i.e. you should only have one observation per participant</td><td>you do not check this in jamovi</td><td>you can only run a one-way ANOVA if you have one observation per participant. If this is not the case, choose another test.</td></tr>
</table>

I check the assumptions and see that the assumption of homogeneity of variance is met (Levene's test is not significant). The Shapiro-Wilk statistic is significant, but given that the statistic can be significant just from the sample size, I turn to the Q-Q plot. The distribution is mostly normal, except for a problem at the left end. I think this is "good enough" for the demonstration. In real life however, you may want to run a non-parametric test (i.e. Kruskal-Wallis ANOVA) which does not assume a normal distribution of the residuals. You can find it in ANOVA > One-Way ANOVA (Non-parametric). If Levene's test had been significant (so the homogeneity of variance assumption is violated) you can run Welch's ANOVA in the ANOVA > One-Way ANOVA option.

Reporting the results of the ANOVA in the text

Navarro and Foxcroft (2019) give examples on how to report the results of an ANOVA on page 339. On page 344 they explain how to report the results of the post hoc test. The results are presented in example 5.5.2.1. ANOVA results are usually only reported in the text. The interpretation of some of the results (e.g. post hoc tests) relies on the inclusion of the descriptive statistics earlier in your report.

Remember that ANOVA tests the hypothesis that the group means are identical. If the test is significant, it means that **at least one** group mean is different from the others. To see which group mean(s) is/are different, we use post-hoc tests, correcting for multiple comparisons using the Holm correction (Navarro and Foxcroft (2019) recommend it) or the Bonferroni correction (widely used).

~~~~~~

Example 5.5.2.1.

One-way ANOVA showed a significant effect of age group on reading comprehension (F(2, 418) = 57, p < .001, $\eta^2 = .214$). Post hoc tests (using the Holm correction to adjust p) indicated that the 6-8 year old group had lower comprehension scores than the 9-11 year old group (t(418) = -8.92,
~~~~~~

p < .001, d = .02) and the 12-13 year old group (t (418) = -9.72, p < .001, d = .03). The older groups did not differ significantly from one another (t (418) = -.82, p = .41, d = .002). The very small effect sizes in the form of Cohen's d indicate that the differences between groups may be of little practical significance.

Additional resources (not compulsory)

- video (7:03): one-way ANOVA
- step-by-step guide for one-way ANOVA
- step-by-step guide for factorial ANOVA
- video (9:49): repeated measures ANOVA
- step-by-step guide for repeated measures ANOVA
- step-by-step guide for mixed ANOVA
- video (6:40): ANCOVA
- video (7:12): MANCOVA
- video (5:59): Kruskal-Wallis test
- video (5:29): Friedman test
- **Navarro & Foxcroft (2019) Chapter 13 | Comparing several means (one-way ANOVA)**
- **Navarro & Foxcroft (2019) Chapter 14 | Factorial ANOVA**

5.5.3. ACTIVITY: ANOVA

1. Use dataset_postImp.omv. Run a one-way ANOVA to answer the following research question:

    To what extent does vocabulary differ by age group?

    1.1. Report the results of the ANOVA and relevant post hoc tests in a paragraph assuming the assumptions of the test were met.
    1.2. The assumptions of ANOVA in fact are not met. What assumptions are violated?
    1.3. What statistical test should you run instead to overcome the violations of the assumptions? (hint: see Table 5.5.2.2.)

2. Why do you think we did not use the matched_dataset.omv ?

Push yourself:

Run the analysis that you suggest in 1.3. and report the results in a paragraph.

5.6. WARNING - P hacking

Don’t be a p-hacker; be responsible.

Head et al. (2015) explain that,

> “There are two widely recognized types of researcher-driven publication bias: selection (also known as the “file drawer effect”, where studies with nonsignificant results have lower publication rates [7]) and inflation [12]. Inflation bias, also known as “p-hacking” or “selective reporting,” is the misreporting of true effect sizes in published studies (Box 1). It occurs when researchers try out several statistical analyses and/or data eligibility specifications and then selectively report those that produce significant results [12–15]. Common practices that lead to p-hacking include: conducting analyses midway through experiments to decide whether to continue collecting data [15,16]; recording many response variables and deciding which to report postanalysis [16,17], deciding whether to include or drop outliers postanalyses [16], excluding, combining, or splitting treatment groups postanalysis [2], including or excluding covariates postanalysis [14], and stopping data exploration if an analysis yields a significant p-value [18,19].”

We may have misled you by running so many analyses on one data set. You’ll forgive us, I hope; because we all knew it was a training exercise. Going forward, you should think carefully about your analyses before conducting your study to avoid p hacking, a pitfall of NHST.

You are encouraged to preregister your analysis plan. Nosek et al. (2018) explain why pre registration is important and how to do it. We encourage you to read their article. There are many places to pre-register your analysis including the Open Science Framework (from which we downloaded data) and AsPredicted.

Do you want to p hack without the problems and feelings of guilt? Try this app to MAKE SURE you find a significant difference between men and women on the nerdy personality scale.

6.0 Checking in: Live Session Friday 26 June 2020

Join us on Zoom for our live session [not available]. We will start at 14:15 on Friday 26 June.

You may switch your video off if you have limited connectivity. Please mute your microphone until you need to contribute. You may post questions you want us to answer on Slack. Please prepare the activities below before the session.

Activities

1. On Friday morning we will allocate each student an activity to talk about in the live session in the afternoon. You will need to email any relevant formatted tables and text by 12:00 on the day. Students are encouraged to complete all activities in Section 5 in preparation.

   The activities include:

   - 5.2.3. t-tests (select appropriate test, provide reasons, and report results)
     - question 1
     - question 2
     - question 3
   - 5.3.3. correlation
     - correlation matrix (formatted in word) and reporting the results of the correlation
   - 5.4.3. linear regression
     - run regression and report results in table and text
     - push yourself activity report back
   - 5.5.3. ANOVA
     - run ANOVA and report results in text
     - push yourself activity

2. James et al. (2020) have released the preprint of their article accepted in Journal of Research in Reading. Read the abstract here and view and interpret Figure 6.1. in preparation for discussion in the live session.
   - List the steps the authors took to analyse their data. How similar/different are these steps to what you have learned in the course?
   - What is the research question you can infer from the abstract?
   - What type of inferential statistics did the authors use to answer the research question?

- What were the authors' results? To what extent do they differ from our results generated in the course? If there are differences, why are there differences?
- Comment on what you have learned through the use of data and materials made available in an Open Science format. To what extent will you make your own data and material available to others? Will your decision to work with collaborators be affected by their data sharing policy?

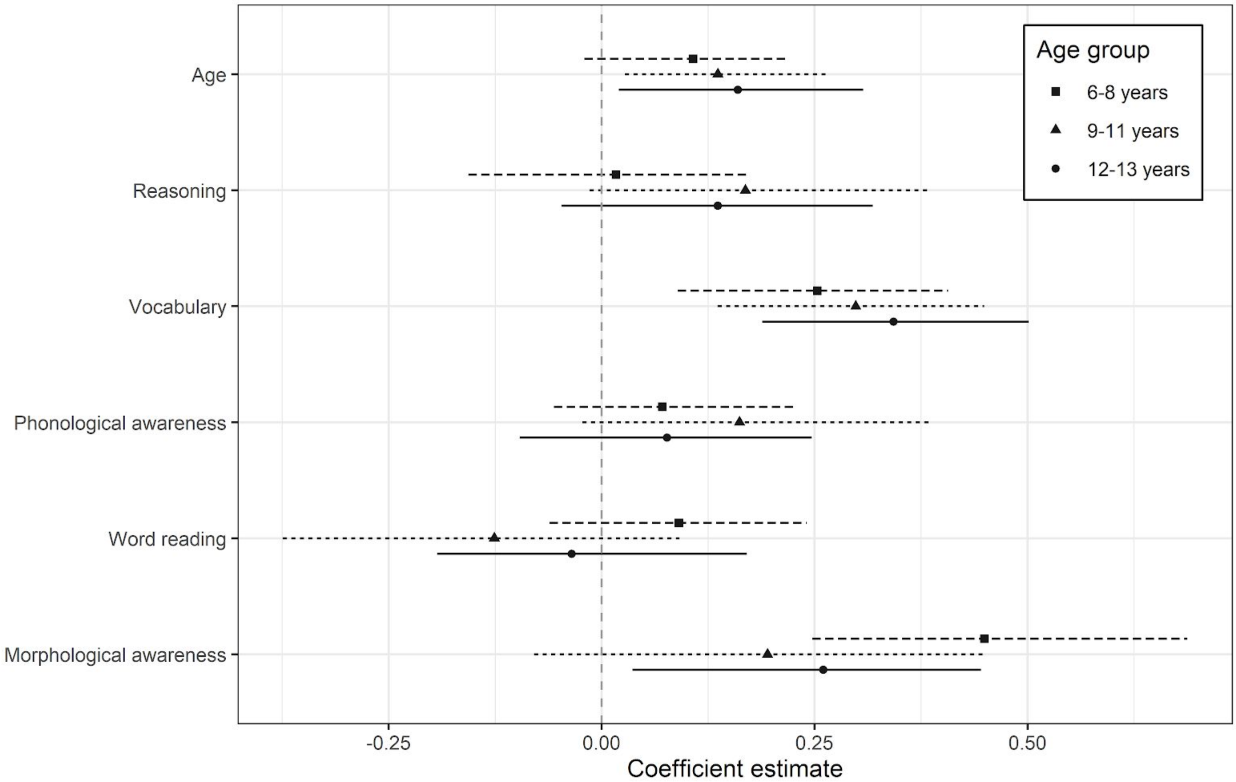


Figure 6.1. Model coefficients with 95% confidence interval for age, reasoning, vocabulary, phonological awareness, word reading and morphological awareness predicting reading comprehension for each age group

Source: James et al. (2020). CC 4.0.

Program

14:15 Greeting and check in

14:20 Answer questions from Slack

14:40 Review of activities

15:00 TEA TIME!

15:15 Commenting on James et al. 2020

15:55 Post course evaluation

16:00 Close

7.0 ACTIVITY: Course evaluation

Please complete the post-workshop evaluation **after the live session.** The link will be shared with you in the live session **[not available].**

Please answer the post-workshop evaluation as part of the requirements of the Statistics Workshop. Reflect on your experience doing the online Statistics Workshop. Answer the questions as truthfully as possible. Your answers are anonymous. Your answers will help us to improve the workshop in future. Thank you for your participation.

8.0 EVALUATION: course assignment for certification

If you completed this course, you can complete the assignment here for possible certification **(this is not an accredited certificate. If you pass, the certificate will state that you completed the course and have met the objectives, and is not aligned with any institution or qualifications authority).**

If your assignment demonstrates that you have achieved the course objectives you will receive a certificate listing the skills you have demonstrated.

Note: if you would like to complete the assignment, you need to submit an application to Ms M Schaefer (schaemn@unisa.ac.za) 6 weeks before you intend to submit. In your application of no more than 500 words, you need to provide some information about who you are and why you would like to complete the module. Ms Schaefer reserves the right to decline or accept applications and submissions as time allows.

9.0 REFERENCES

The reference list is split into different sections according to the type of reference.

Abstracts, Articles and Preprints

Côté, S. L., Gonzalez-Barrero, A. M., & Byers-Heinlein, K. (2020, February 18). Multilingual Toddlers' Vocabulary Development in Two Languages: Comparing Bilinguals and Trilinguals. https://doi.org/10.31234/osf.io/hpqez

Head, ML., Holman, L., Lanfear,. R., Kahn, AT., & Jennions, MD. (2015). The Extent and Consequences of P-Hacking in Science. *PLoS Biol 13*(3): e1002106. https://doi.org/10.1371/journal.pbio.1002106

James, E., Currie, N., Tong, S. & Cain, K. (2020). The relations between morphological awareness and reading comprehension in beginner readers through to young adolescents. *Journal of Research in Reading*. ISSN 0141-0423 (In Press). Abstract accessed from http://www.research.lancs.ac.uk/portal/en/publications/the-relations-between-morphological-awareness-and-reading-comprehension-in-beginner-readers-through-to-young-adolescents(a1853ec1-2602-4e11-8aff-18ade5a079eb).html

Kotze, J., Fleisch, B., & Taylor, S. (2019). Alternative forms of early grade instructional coaching: Emerging evidence from field experiments in South Africa. *International Journal of Educational Development*, 66(September 2018), 203–213. https://doi.org/10.1016/j.ijedudev.2018.09.004

Murray, D., Boothby, C., Zhao, H., Minik, V., Bérubé, N., Larivière, V., & Sugimoto, C.R. (2020). Exploring the personal and professional factors associated with student evaluations of tenure-track faculty. *PLoS ONE 15*(6): e0233515. https://doi.org/10.1371/journal.pone.0233515

Nosek, BA., Ebersole, CR., DEHaven, A., & Mellor, DT. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 115(11). 2600-2606. https://doi.org/10.1073/pnas.1708274114

Pretorius, E.J. & Stoffelsma L. (2017). How is their word knowledge growing? Exploring Grade 3 vocabulary in South African township schools. *South African Journal of Childhood Education*, 7(1), a553. https://doi.org/10.4102/sajce.v7i1.553

Probert, T.N. (2019). A comparison of the early reading strategies of isiXhosa and Setswana first language learners. *South African Journal of Childhood Education*, 9(1), a643. https://doi.org/10.4102/sajce.v9i1.643

Schaefer, M. & Kotzé, J. (2019). Early reading skills related to Grade 1 English Second Language literacy in rural South African schools. *South African Journal of Childhood Education*, 9(1), a644. https://doi.org/10.4102/sajce.v9i1.644

Wilsenach, C., (2015). Receptive vocabulary and early literacy skills in emergent bilingual Northern Sotho-English children. *Reading & Writing*, 6(1), Art. #77, 11 pages.

Books

Navarro, DJ. & Foxcroft, DR. (2019). learning statistics with jamovi: a tutorial for psychology students and other beginners. (Version 0.70). DOI: 10.24384/hgc3-7p15

Datasets

Côté, S. L., Gonzalez-Barrero, A. M., & Byers-Heinlein, K. (2020, February 18). Multilingual toddlers' vocabulary development in two languages: Comparing bilinguals and trilinguals. https://osf.io/us27h/

Hamalainen, J. (2020, April 30). Cognitive data for auditory processing and phonological development project. https://osf.io/rhmq5/

James, E., Currie, N. K., Tong, S. X., & Cain, K. (2020, May 21). The relations between morphological awareness and reading comprehension in beginner readers through to young adolescents. https://osf.io/96542/

Weng, T. (2020). Survey data of foreign language learners' enjoyment and anxiety in the U.S.. *Data in Brief*, 30, 105221. https://doi.org/10.1016/j.dib.2020.105221

Images and figures (in order of presentation)

Page 1. Globe: qimono, 2016. Pixabay. https://pixabay.com/photos/world-globe-day-night-sun-stars-1582347/

Page 4. Footprints on beach: ddzphoto, 2019. Pixabay. https://pixabay.com/photos/footsteps-beach-sand-sea-ocean-3938563/

Page 14. Stadium chairs: MichaelGaida, 2016. Pixabay. https://pixabay.com/photos/stadium-sit-auditorium-grandstand-1750794/

Page 17. Crumpled paper and pencil: congerdesign, 2016. Pixabay. https://pixabay.com/photos/pencil-notes-chewed-paper-ball-1891732/

Page 28. EKG coffee: GDJ, 2019. Pixabay. https://pixabay.com/vectors/ekg-coffee-health-drink-beverage-4388872/

Page 29. Stone build rock. OyeHaHa, 2017. Pixabay. https://pixabay.com/photos/stone-build-rock-nature-sky-2798007/

Page 31. Zebra and giraffe: Ellethyia, 2018. Pixabay https://pixabay.com/photos/giraffes-pair-zebra-herd-3529651/

Page 35. Figure 5.3.1.: Vigen, T. (nd.). Number of people who drowned by falling into a pool correlates with Films Nicholas Cahe appeared in. Spurious Correlations. Accessed June 2020 from https://www.tylervigen.com/spurious-correlations

Page 39. Linear regression: Sewaqu, 2010. https://en.wikipedia.org/wiki/Linear_regression#/media/File:Linear_regression.svg

Page 48. Figure 5.4.3.1. Individual, classroom, university and research characteristics associated with overall teaching quality. Murray, D., Boothby, C., Zhao, H., Minik, V., Bérubé, N., Larivière, V., & Sugimoto, C.R. (2020). Exploring the personal and professional factors associated with student evaluations of tenure-track faculty. *PLoS ONE 15*(6): e0233515. https://doi.org/10.1371/journal.pone.0233515

Page 56. anonymous face: janjf93, 2016. Pixabay. https://pixabay.com/vectors/mask-anonymous-face-panel-internet-1587566/

Page 58. Figure 6.1. Model coefficients. James, E., Currie, N. K., Tong, S. X., & Cain, K. (2020, May 21). The relations between morphological awareness and reading comprehension in beginner readers through to young adolescents. https://osf.io/96542/

Software

The jamovi project (2019). *jamovi*. (Version 1.1) [Computer Software]. Retrieved from https://www.jamovi.org.

Videos

We created most of the videos for the course. Unless otherwise indicated, other hyperlinked videos were created by Poulson and retrieved from the jamovi playlist of datalabcc on Youtube.

Poulson, B, (2019). Jamovi. Retrieved June 2020 from Datalab.cc.from https://datalab.cc/jamovi/

Sprightly Pedagogue, 2016. Null Hypothesis Significance Testing. https://youtu.be/B2_UJGs0hLI

Websites

Alagöz, Ömer Emre Can & Nuijten, Michèle, (n.d.). P-Hack the Personality. https://emrealagoz.shinyapps.io/phack-thepersonality/

Centre for Open Science, 2020. OSFHome. https://osf.io/

Holtz, Y., (2018). The R graph gallery. https://www.r-graph-gallery.com/

Rafi, J., (2020). The jamovi quickstart guide. https://www.jamoviguide.com/index.html

Statistics Solutions, (2020). Testing Assumptions of Linear Regression in SPSS. https://www.statisticssolutions.com/testing-assumptions-of-linear-regression-in-spss/

Simonsohn, U., Simmons, J, & Nelson, L., (nd). As Predicted. https://aspredicted.org/

Wagih, O., (n.d.). Guess the Correlation. http://guessthecorrelation.com/
