Practice-Oriented Paper

# Conducting and Interpreting *T*-test and ANOVA Using JASP

Caroline J. Handley
*Asia University*

Many foreign language university educators conduct quantitative research; for some analysing and interpreting the data can be the most challenging aspect. A short introduction to data analysis is presented, focusing on why effect sizes and confidence intervals should be reported as well as the probability of a statistical difference between groups (*p* value). This is followed by an explanation of how to perform two commonly-used analyses in applied linguistics research, *t*-test and ANOVA, using JASP, an open source software package created at the University of Amsterdam. Finally, further consideration is given to effect size and confidence intervals, as related to statistical power, to highlight the importance of good study design. The overview provided should be useful for novice researchers who need to perform and report basic statistical analyses of quantitative data obtained through classroom research.

This paper provides an introduction to basic statistical analysis of quantitative data for novice researchers. Statistical analysis inevitably involves technical language; such terms that are used in this paper are explained and, for ease of reference, described in table format (Appendix A). A brief overview of some theoretical issues in understanding statistics is given first. Then, after a summary of the version of JASP[1](JASP Team, 2018; Version 0.9.0.1; downloaded at jasp-stats.org), a fictional data set is used to illustrate how to perform two frequently-used analyses in applied linguistics research using JASP software. First, I explain how to import data into JASP, then how to calculate and visualise the numbers that describe the data (descriptive statistics). Next, an example is given of an independent *t*-test and analysis of variance (ANOVA). These tests yield a probability (*p*) value which enables researchers to make predictions about the whole population based

on the sample used in one's experiment (inferential statistics). Finally, the data is used to discuss interpreting and supplementing *p* values and designing reliable experiments.

The experimental design described in this paper involves the comparison of students within a control group (i.e., a group which receives no experimental intervention) and one or more experimental groups. The intervention is the independent variable (the factor manipulated) and its effect is measured on the dependent variable (test scores) by means of an independent *t*-test and ANOVA (explained below). Both test the null hypothesis that any difference between the scores of each group is due to chance variation (random error). The tests give the probability of the results that were obtained if the null hypothesis were true, stated as a *p* value, which is typically declared statistically significant if it is less than .05 (if the results would be obtained due to chance in less than 5% of studies). This probability is based on the difference between the mean (average test score) of each group and the variance within each group (the difference between each student's score and their group mean), summarised as the standard deviation of each group. The tests do not test the research hypothesis (the alternative hypothesis) of a systematic effect of the intervention on test scores, therefore a significant *p* value does not indicate the research hypothesis is true (see Cohen, 1994 for a detailed explanation).

Statistical significance testing, although widely used, has long been criticised due to two further limitations (e.g., Carver, 1978). The *p* value obtained is largely dependent on the sample size (the number of students in each group) and does not provide any information about the magnitude of any difference between groups. Since 1994, APA guidelines have stated that effect sizes should always be reported alongside *p* values (American Psychological Association, 2009), although this is rare in applied linguistics research (Norris & Ortega, 2000; Plonsky, 2011). Effect size is a measure of the magnitude of the effect or difference between groups (or interventions), so it more directly diagnoses the importance of the factor(s) or variables being investigated (Plonsky, 2015). Benchmarks proposed by Cohen (1992) for the behavioural sciences are often used. For differences between the means of two groups (such as *t*-test),

he suggested that Cohen's $d$ = 0.2 should be considered a small effect, 0.5 is a medium effect, and any value over 0.8 indicates a large effect. Unlike when reporting $p$ values, the zero before the decimal point is needed as effect sizes larger than one are possible. However, Oswald and Plonsky (2010) tentatively proposed that in L2 research, Cohen's $d$ values of 0.4, 0.7, and 1.0 should be considered the benchmarks for small, medium, and large effects. For differences between the means of more than two groups (such as ANOVA), the standards for eta squared ($\eta^2$) or omega squared ($\omega^2$) are 0.01 for small, 0.06 for medium, and 0.14 for large effects (Field, 2017).

Confidence intervals (CIs) are another measure that should be reported (Cumming, 2012), as they show the uncertainty around the mean values obtained. They are based on the standard error, which is calculated from the standard deviation and sample size, meaning that confidence intervals decrease as sample or group size increases. The 95% confidence interval covers a range of values above and below the mean of each group, indicating 95% confidence that the true mean lies somewhere within this range. Small groups are more vulnerable to chance variation, so 95% CIs are likely to be long, reflecting low confidence in the accuracy of one's results.

Effect sizes and CIs are explained not only to encourage better reporting of results, but also to highlight the importance of sample size in quantitative research. In the fictional experiment created to model data analysis, I used the average group size in second language research of 20 participants (Plonsky, 2013; Plonsky & Gass, 2011). This was deliberate to illustrate why this sample size is often too small to reliably detect a significant difference even when the research hypothesis is true. If the group size is too small, an experiment is underpowered, and chance largely determines whether the result will be statistically significant (Ioannidis, 2005). This issue in the analyses conducted is revisited at the end of this paper, and some further resources for understanding and conducting quantitative data analysis are provided in Appendix B.

## Data Analysis in JASP

JASP is used in this paper as it is free software, available for Windows, Mac, and LINUX. It is user-friendly for novices and includes guided analyses of various data sets, many taken from Field (2017). Results of statistical tests performed in JASP are displayed in tables and graphs that can be exported into documents in APA format. In addition, JASP is regularly being updated to expand its functionality.

JASP also offers advanced statistical analyses, such as structural equation modelling (SEM) and Bayesian versions of most inferential tests. Bayesian analysis is an alternative approach to probability testing which is being adopted in many scientific disciplines (e.g., Van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). It is not discussed in this paper, but a short introduction by the developers can be downloaded from the website (Wagenmakers, Gronau, Etz, & Matzke, 2017).

The analyses described below were performed using a deliberately simple fictional data set, simulating the scores of 20 participants per group on a vocabulary test with a maximum score of 20. The imaginary scenario was that students read texts containing a total of 20 unknown words, then were tested on their knowledge of those target words under one of two to three conditions: no explicit vocabulary instruction (control group), Intervention A, and, to describe ANOVA, an alternative Intervention B. The research hypothesis for both analyses was that the students who received the teaching intervention(s) would perform significantly better on the vocabulary test than the control group. Therefore, the null hypothesis being tested was that there would be no statistically significant difference in test scores between groups.

## Importing and Manipulating Data in JASP

JASP can open data files that are saved as .csv, .sav, or .ods files. Excel spreadsheets can be saved as .csv files by selecting that option from the dropdown menu on saving. When the file is open in JASP, any subsequent changes saved in the original spreadsheet are automatically updated. The data for each participant should be entered in separate rows, with appropriate column headings. In the example data sets used here, there are a control and one or two intervention groups. This

means that for both the *t*-test and ANOVA there are just three columns of data: participant ID, group type, and vocabulary test score.

As Figure 1 shows, JASP displays information in three panels. The data are on the left, the instructions options for the analyses are in the centre, and the results of the analyses performed are on the right. On loading the spreadsheet, JASP guesses the type of data in each column: nominal (non-numerical data, such as gender), ordinal (ranked numbers, such as the responses on a Likert scale), and scale (numbers that are separated by equal distances, such as test scores[2]). If JASP guesses incorrectly, click on the symbol next to the heading to change it before proceeding.

The available statistical tests are spread across the top bar, from which a sub-type is selected. Then the middle panel will appear with boxes for entering data into the analysis. Data can be moved by dragging and dropping or by highlighting and then clicking the arrow. Data can also be moved back out of the analysis with the same ease. Below these boxes there are expandable bars with extra options. This design keeps the screen uncluttered and allows users to add various additional elements to the main analysis as required.

## Descriptive Statistics

Inferential statistical tests such as *t*-test and ANOVA are based on two fundamental numbers from descriptive statistics: the mean (*M*) and standard deviation (*SD*) of each group. Thus, the mean and standard deviation of both or all groups should be calculated and visualised before performing such inferential statistical tests and reported alongside their results (Plonsky, 2015). Visualising the data enables one to see if the distribution of data points (in this case, test scores) looks similar to a bell-shaped curve, called a normal distribution. This is important as it is related to the *SD*, which is a measure of the variance within each group. Under a normal distribution, approximately 68% of all data points are within 1 *SD* of the mean and 95% lie within 2 *SD*s.

In the example analysis, in the instructions panel, Score is moved into the Variables box, then Split by Group (Figure 1). JASP provides the mean, standard deviation, and minimum and maximum test score for each group, shown in the

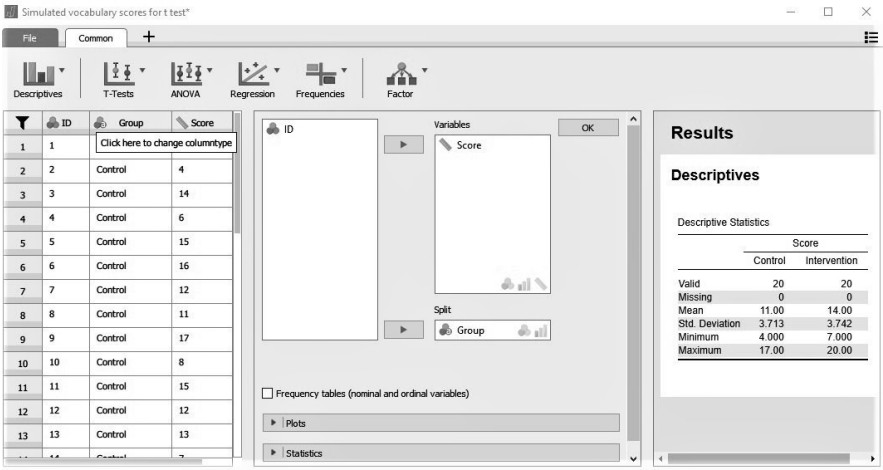


*Figure 1*. Descriptive statistics for the vocabulary scores of two groups as generated in JASP, with data on the left, instructions in the centre, and results on the right.

Results panel on the right. This information is visualised by clicking on Plots and selecting Boxplot. Although it might not be necessary to include the boxplot in a paper, it is a useful tool to understand one's data. As can be seen in the image on the left in Figure 2, a boxplot shows the mean (thick black horizontal line) surrounded by a grey box which represents 1 *SD* above and below it. The horizontal bars above and below represent 2 *SD*s. Any outliers are shown as dots above and below the bars. It is a good idea to check outliers in case they represent a mistake in the data (e.g., an extra digit). Colour, violin, and jitter options can be added to the boxplot by selecting the appropriate boxes. Violin and jitter effects are shown in the image on the right in Figure 2. The violin element shows the distribution curve, mirror-imaged and rotated 90°, and the jitter element adds small circles representing every individual data point. These effects show that test scores in both groups approximate a normal distribution and that there is considerable overlap between groups. To find out if the difference between the means of each group, given the variance in scores (*SD*) within both groups, is significantly improbable due to random sampling error, an independent *t*-test is conducted.

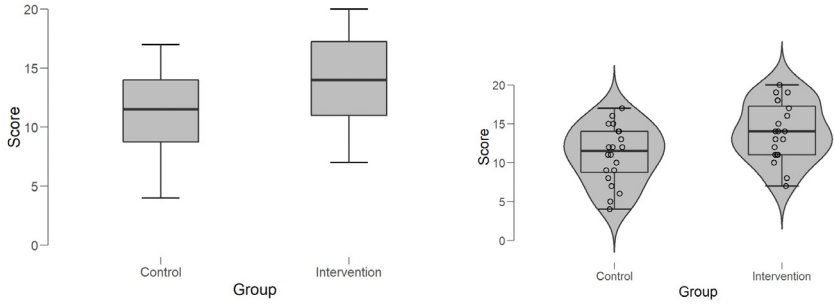


*Figure 2*. Boxplots of vocabulary test scores by group as generated in JASP.

## Independent *T*-Test

As each group contains different students, an independent samples *t*-test is needed. JASP offers three choices of *t*-test: Student, Welch, and Mann-Whitney. Student's and Welch's *t*-tests can be performed on scale data (they are called parametric tests and are based on the mean and the values of all data points). The Mann-Whitney *t*-test is used with ordinal data or if the distribution of data within the groups is not approximately normal (it is a non-parametric variant, based on the median and ranked data points). The Student's *t*-test also assumes the amount of variance within each group is similar. The Assumption Checks allow one to see if the assumptions of normal distributions and equal variance within each group are met and JASP explains how to interpret the results. Welch's *t*-test does not require equal variance; in statistical terms, it is more robust than the Student's *t*-test, making it a more reliable test which should be preferred (Delacre, Lakens, & Leys, 2017).

With the fictional data used here, the variance between groups is equal, so Student's and Welch's *t*-tests yield identical results (shown in the table in Figure 3). *T*-tests produce a *t* statistic (in this case -2.545), and the *p* value depends on that and the degrees of freedom (*df*), which is the number of observations (in this case students) minus the number of groups. With 40 students in two groups there are 38 *df*. The *p* value shows the probability of obtaining these results if there was only random variation between the two groups. Within social sciences, if this is less than 5% ($p = .05$) it is considered statistically significant.


*Figure 3*. Welch's *t*-test and descriptives plot as generated in JASP.

Exact values of *p* are reported between .05 and .01. At lower values of *p*, it is often reported as being less than a certain value, for example, $p < .01$ or $p < .001$. If *p* is less than .001, the difference between groups is considered highly significant. The table created in JASP contains all the information needed to report the results, after stating the mean and SD of each group. For this fictional data set, the results could be reported as follows: "An independent *t*-test showed that the control group recalled significantly fewer words than the group which received the vocabulary teaching intervention, $t(38) = -2.545$, $p = .015$."

The effect size should also be reported, displayed by checking the Effect size box under Additional Statistics. This returns the value for Cohen's *d*. The value obtained for the fictional data is $d = -0.805$, which is considered a large effect size (Cohen, 1992), although Oswald & Plonsky (2010) suggest it should be interpreted as a medium-sized effect within second language research. Finally, a Descriptives plot is created, which shows the mean score for each group, surrounded by bars which represent the 95% confidence interval (CI) around each mean. Due to the small group sizes, the 95% CIs are very long, showing that there is a lot of uncertainty as to the value of the true mean for the population to which each group belongs. If this were a real experiment, the results should be reported with caution, as it is uncertain whether similar results would be

obtained with different groups of students.

## ANOVA

Similar to the *t*-test, ANOVA is also used to test for a significant difference between groups, when there are more than two groups. A simple ANOVA design is described here, in which the only difference from the *t*-test above is the addition of an extra group of 20 students who experienced Intervention B. The process for analysing the data in JASP is basically the same as described above for *t*-test. Once the data file has been loaded and descriptive statistics have been produced, the ANOVA test is selected (Figure 4). Score is moved to the Dependent Variable box, and Group is entered as the Fixed Factor. In the *t*-test the default order, in which the control group was compared to the intervention, was reported (which is why the *t* statistic and effect size were negative, as control group scores were lower). This time, by clicking on the spreadsheet headers in the data panel in JASP and selecting the control group then the down arrow, the order is reversed, so that the group receiving Intervention A is compared with Intervention B and then with the control group. The result is again significant, although slightly less so, at $p = .023$.

As the result is statistically significant, post hoc tests are performed to identify which groups are significantly different. These are similar to conducting multiple *t*-tests, but they control the Type I error rate (multiple comparisons multiply the error rate). In JASP, expand the Post Hoc Tests bar in the centre panel and move Group into the box on the right. Tukey is the default option, although other tests, such as Bonferroni, can be selected. Tukey requires equal group sizes; Bonferroni is more generalisable if one is unsure which test is most appropriate. Although there is an effect size box, Cohen's *d* is not appropriate for multiple comparisons. Instead, the correct effect size estimates are under the Additional Options bar. Omega squared ($\omega^2$) was selected, as this has been shown to perform better with small group sizes than the default option, eta squared ($\eta^2$), and equally well with larger groups (Okada, 2013).

The results can be reported in a similar way to the *t*-test above. However, in this case the *F* statistic is calculated and there are two values for degrees of


*Figure 4*. ANOVA, post hoc tests, and descriptives plot as generated in JASP.

freedom (*df*). The first is the number of groups minus one (3 – 1 = 2) and the second is the sum of *df* for each group (three groups of 20 minus one, therefore 3(20 - 1) = 57). To report this, one could write:

> An independent one-way ANOVA showed there was a significant, medium-sized effect of teaching intervention on vocabulary test scores; $F(2, 57) = 4.017$, $p = .023$, $\omega^2 = 0.091$. Planned post hoc tests using Tukey test showed that students who received Intervention A recalled significantly more words than the control group ($p = .023$). However, there was no significant difference in test scores between the control group and the group that received Intervention B ($p = .776$), and the difference between Intervention A and B, on average 2.25 words, was marginally non-significant ($p = .111$).

Note that although there was no significant difference between Intervention B and either of the other two groups, the *p* values are quite different. One would expect the observed difference in test scores under Intervention A and B in 11% of studies due to random sampling error, a higher Type I error rate than acceptable. However, this does not indicate there was no difference between the two groups.

## Discussion

The fictional data presented above, with the small group size of twenty people, was designed to produce a significant difference between test scores of students in the control group and those who received Intervention A, but not between Intervention A and Intervention B. Despite this, the Intervention B scores were, on average, only slightly higher than those of the control group. A common mistake is to interpret such results as indicating no difference between Intervention A and B (Amrhein, Greenland, & McShane, 2019). The accurate interpretation is that the results are inconclusive. Intervention A may be more effective than Intervention B, but no conclusion can be drawn without replicating the experiment with larger groups of participants. This does not mean repeating the experiment would necessarily be worthwhile. The effect size could be inflated beyond the true value, and the null hypothesis of no difference between the two interventions could be true.

However, it is intended as a caution against conducting research with group sizes that are too small to have a good chance of detecting an effect even when there really is one (an underpowered study), which, unfortunately, is common in applied linguistics (Plonsky, 2013; Plonsky & Gass, 2011). To accompany his book, Cumming (2012) created interactive spreadsheets using Exploratory Software for Confidence Intervals (ESCI; https://thenewstatistics.com/itns/esci/esci-for-utns/) to illustrate the issues of underpowered studies and random sampling error. The basic message is that with small groups the chance of a type II error (not rejecting the null hypothesis when it is false) is often high and, concomitantly, so is the chance that a published significant result in an experiment with small groups is a type I error (rejecting the null hypothesis when it is true). This latter phenomenon has led to the so-called crisis in psychology, in which the effect found in famous published studies has not been replicated in large-scale experiments (Open Science Collaboration, 2015). There are several reasons for this, but the most important is probably Abelson's (2012) first law of statistics, "chance is lumpy" (p. 19). In other words, with small group sizes random sampling error (chance effects) can be big enough to create a spurious significant difference between groups.

One way to enable better interpretation of the significance of the data is to report confidence intervals (CIs). The 95% CI covers a range of values around the mean of each group (Figures 3 and 4); there is 95% probability the true mean is within these values. In the simulated experiments above, the bars are quite large, showing the uncertainty of the results due to random sampling variation. This means the results should be interpreted and reported with caution. Although there is less than a 5% probability of obtaining these results if there was no systematic difference between Intervention A and the control group, the true group mean scores could be much smaller or much larger than that found with the simulated participants. In this experiment Intervention A was effective, compared with no intervention. However, due to the small sample size, it remains uncertain whether this result would be replicated with different student groups. In other words, it is unclear whether this result would be generalisable, which is the reason for conducting inferential statistical tests (for a fuller explanation, see Cumming, 2012). Confidence intervals can also be calculated for effect size, providing the range of values within which the true magnitude of the effect lies (Kelly & Rausch, 2006).

Effect size and CIs illustrate why it is important to design experiments with large enough groups or enough power to reliably detect an effect if it does exist. The free software G*Power 3 (http://gpower.hhu.de), described in Faul, Erdfelder, Lang, & Buchner (2007), enables researchers to calculate the power of an experiment, before or after it has been conducted. As an example, in an experiment with two independent groups, to have an 80% chance of detecting a medium effect size of 0.5 (80% power), with the type I error rate set at the standard alpha ($\alpha$) = 0.05, each group would need 64 participants. With a medium effect size and groups of 20 participants, an experiment is so underpowered that even when there is a true effect there is less than a 50% chance of detecting it as a statistically significant difference.

A potential solution for the experiment described above would be to design it such that the participants acted as their own control, by participating in both control and intervention conditions, learning two different sets of words. Due to space constraints, such a design has not been explained in this paper, but

analysis would require a paired samples *t*-test or repeated measures ANOVA. Comparing students with themselves under two conditions removes variability between individuals (ability, motivation, etc.). Although there would still be within individual variation (sleep levels, mood differences, etc.), this is much smaller, so measurement of the dependent variable (test scores) would be more reliable, provided the sets of words were of similar difficulty. Repeated measures designs are closer to comparing the target comparison factors or variables than independent designs (Hand, 1994). Whereas 128 participants are needed to achieve 80% power with two independent groups and a medium effect size, with a repeated measures design the same power is reached with just 34 participants.

Alternatively, only one group is needed to investigate the relationship between variables (i.e., the relationship between time of test and test scores), which is analysed by correlation (in JASP, regression: correlation matrix). Another option is to conduct action research (Burns, 2005), rather than attempting to use inferential statistics to generalise beyond one's own teaching context. In sum, a basic understanding of experimental design can prevent a lot of wasted effort as well as increasing the reliability of reported research results.

## Conclusion

The design and aim of the simulated experiment were very simple, as the focus was on how to conduct two common analyses in applied linguistics research, *t*-test and ANOVA, using JASP. More complicated research questions or longitudinal designs often necessitate independent intervention and control groups, requiring such analyses. The description given here should guide novice researchers in conducting basic quantitative data analysis and interpreting the results.

In addition, the need for good experimental design was highlighted, including the reporting of effect size and CIs. The fictional experiment was underpowered due to small group size; a repeated measures design would have generated more reliable results. When conducting quantitative research, access to sufficient participants to have at least an 80% chance of achieving the research goals should be ensured. Otherwise, spending a large amount of time collecting interesting information about students and their learning may be meaningless.

## Notes

[1] Newer versions of JASP have since been released with additional capabilities.

[2] In classical test theory in which a linear distance between scores is assumed. However, as highlighted in the Rasch model, in reality distances between raw scores are not equal due to differences in item difficulty.

## References

Abelson, R. P. (2012). *Statistics as principled argument*. (Kindle version). Retrieved from Amazon.co.uk.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Amrhein, V., Greenland, S., & McShane, B. (2019, March 20). Scientists rise up against statistical significance. *Nature, 567*(7748), 305-307, https://doi.org/10.1038/d41586-019-00857-9.

Burns, A. (2005). Action research: An evolving paradigm? *Language Teaching*, *38*(2),57-74. https://doi.org/10.1017/S0261444805002661

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378-399. https://doi.org/10.17763/haer.48.3.t490261645281841

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159. https://dx.doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997-1003. https://dx.doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* (Kindle version). Retrieved from Amazon.co.uk.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology*, *30*(1), 92-101. https://doi.org/10.5334/irsp.82

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and

biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Field, A. P. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). London, England: Sage.

Goss-Sampson, M. A. (2018). *Statistical analysis in JASP: A guide for students*. Retrieved from: https://jasp-stats.org/jasp-materials/

Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 317-356. https://doi.org/10.2307/2983526

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kelly, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363-385. https://dx.doi.org/10.1037/1082-989X.11.4.363

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417-528. https://doi.org/10.1111/0023-8333.00136

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, *40*(2), 129-147. https://doi.org/10.2333/bhmk.40.129

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 1-8. https://doi.org/10.1126/science.aac4716

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85-110. https://doi.org/10.1017/S0267190510000115

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, *61*(4), 993-1038. https://dx.doi.org/10.1111/j.1467-9922.2011.00663.x

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second*

*Language Acquisition*, *35*(4), 655-687. https://doi.org/10.1017/S0272263113000399

Plonsky, L. (2015). Statistical power, *p* values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research*. (Kindle version). Retrieved from Amazon.co.uk.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*(2), 325-366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Wagenmakers, E. J., Gronau, Q. F., Etz, A., & Matzke, D. (2017, May 4). *The JASP Book*. Retrieved from https://osf.io/3bdqp

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217-239. https://dx.doi.org/10.1037/met0000100

## Author bio

Caroline Handley *is an English lecturer at Asia University in Tokyo. She is currently pursuing a PhD at Swansea University in Wales, where she is using word association tasks to research the relation between conceptual and linguistic knowledge in lexical processing. handley.caro@gmail.com*

Received: November 23, 2018
Accepted: April 23, 2019

# Appendix A

## Technical Terminology

The table below summarises the technical terms related to quantitative data analysis used in this paper. The terms are listed so that terms related to similar concepts are adjacent.

| Technical Term | Description |
|---|---|
| Descriptive statistics | The numbers that describe the data collected, such as the mean (M) and standard deviation (SD). |
| Inferential statistics | Tests performed to make predictions about the whole population based on the sample used in one's experiment. |
| Mean (M) | This is the average of all the data points. (It should not be confused with the median, which is the middle number in a ranked data set.) |
| Standard deviation (SD) | This is a measure of the amount of variance within a group. It summarises the distance of each data point from the mean. A small SD indicates most data points are close to the mean, a larger SD shows the data points are more dispersed. |
| Normal distribution | This is a bell-shaped curve which covers the distribution of data points around the mean. Most data points cluster around the mean, with few points in the tail ends of the curve. |
| Null hypothesis (H0) | The hypothesis that any difference between groups is due purely to chance (random noise). |
| Alternative hypothesis (H1) | The research hypothesis or theory being tested which proposes there is a systematic difference between groups due to some factor. |
| Independent variable | The variable (or factor) that is changed or controlled in an experiment in the prediction that it will have an effect on another variable. |
| Dependent variable | The variable that is measured in an experiment (such as test score), to test the null hypothesis that it will not be systematically altered by the effect of the independent variable. |

| Technical Term | Description |
|---|---|
| Control group | Scientific design involves the random assignment of individuals to one of two groups. The control group receives no treatment or intervention but provides a comparison for the experimental group. |
| Experimental group | This group receives the treatment or intervention (the independent variable). The alternative hypothesis is that this will produce a statistically significant difference in the dependent variable compared to the control group indicating a systematic effect. In repeated measures designs the experimental group acts as its own control. |
| Type I error | Rejecting the null hypothesis when it is true. By standard, the chance of making this error is kept to less than 5% (p. = .05). |
| Type II error | Not rejecting the null hypothesis when it is false. The chance of making this error is related to effect size and sample size. The smaller the systematic effect of the independent variable on the dependent variable the larger the sample size needed to reduce the chance of making this error. |
| Statistical significance | Results are said to be statistically significant if the observed difference between groups would be expected to occur due to random sampling error in 5% or less of studies. |
| Parametric tests | These are inferential statistical tests based on the mean and individual data points. Such tests assume the data collected is normally distributed. |
| Non-parametric tests | These are inferential statistical tests based on the median and ranked data points. Such tests make no assumptions about the distribution of the data collected. |
| Independent t-test | This statistical test is used to compare the means of the control and experimental group and indicates the probability that the two groups come from the same population (are identical), given the variance within each group. |

| Technical Term | Description |
|---|---|
| ANOVA | This statistical test is used to compare the means of the control and two or more experimental groups and indicates the probability that all groups come from the same population, given the variance within each group. |
| Repeated measures ANOVA | This statistical test is used to compare the means of one group under three or more different conditions (including a control condition) and indicates the probability that all conditions are equal, given the variance within each condition. The equivalent t-test is called a paired samples t-test. |
| Post-hoc tests | These tests are conducted after an ANOVA that yields a statistically significant result to find out where the differences are. They are similar to conducting multiple t-tests but prevent inflation of the Type I error rate. The most common tests are Tukey's HSD and Bonferroni. |
| Effect size | The effect size is a calculation of the magnitude of the difference between two or more groups. There are different measures of effect, such as Cohen's d, each with standardised benchmarks for small, medium, and large effects. |
| Confidence interval (CI) | This is an estimate of the range of values within which the true population mean is expected to lie. A 95% CI is typically calculated, meaning that there is a 95% probability that the mean is within this range. Larger sample sizes enable greater accuracy of measurement and smaller CIs. |
| Power | Statistical power is related to Type II errors. It is the probability that an effect will be detected as significant if there is a true effect. Power is increased by making better hypotheses, effect size, and sample size. Many studies are underpowered as the sample size is too small. Power can be calculated before or after conducting an experiment. |

# Appendix B

## Suggested Resources

### Books

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge (Kindle version). Retrieved from Amazon.co.uk.

Goss-Sampson, M. A. (2018). *Statistical analysis in JASP: A guide for students*. Retrieved from: https://jasp-stats.org/jasp-materials/.

Plonsky, L. (2015). *Advancing quantitative methods in second language research*. Routledge (Kindle version). Retrieved from Amazon.co.uk.

### Free Online Resources

- Coursera: an intermediate-level course by Daniël Lakens about interpreting statistics, including Bayesian statistics (coursera.org/learn/statistical-inferences)
- JALT Testing and Evaluation SIG: produces the journal *Shiken*, which includes articles on statistical issues, typically in relation to classroom assessment (teval.jalt.org)
- JASP Statistics: a collection of videos demonstrating how to perform various data analyses using JASP (youtube.com/channel/UCSulowI4mXFyBkw3bmp7pXg)
- Khan Academy: a beginner-level course on statistics and probability (khanacademy.org/math/statistics-probability)
- Penn State Eberly College of Science Statistics course: a text-only elementary statistics course which is also useful as a reference (onlinecourses.science.psu.edu/stat200/home)
- Statistics of DOOM: a large collection of videos showing how to perform various data analyses using JASP, Excel, SPSS, R, and Python (youtube.com/channel/UCMdihazndR0f9XBoSXWqnYg)