Moving Towards Better Quantitative Data Analysis in FLL Research

Paul Collett
Shimonoseki City University

Reference Data:

Collett, P. (2017). Moving towards better quantitative data analysis in FLL research. In P. Clements, A. Krause, & H. Brown (Eds.), *Transformation in language education.* Tokyo: JALT.

Although null hypothesis statistical testing (NHST) is the dominant method for quantitative data analysis in foreign language learning research, it is a misused and inappropriate methodology. A brief outline of some of the problems with this methodology will be presented along with some suggestions on alternative data analysis methods that can either supplement or replace NHST and ideally contribute to better research practices in foreign language learning research.

外国語学習研究における量的データの解析方法として統計的仮説検定が広く用いられているが、この方法は誤用が多く適切とはいえない。外国語学習研究で統計的検定を用いる際の問題点について、この有意性検定の欠点を補う方法、或いは代用できるデータ解析手法を提案しながら、簡潔に論じる。この議論は、より良い外国語学習研究を実践していく上で有益となるであろう。

There is a recognition of the weaknesses and the misuse of the traditional null hypothesis statistical testing (NHST) methods that tend to dominate quantitative research in language learning research (as well as in education, psychology, and other related disciplines). Critiques of the orthodox quantitative research methodology have a long history within the psychology and education fields, from seminal papers such as Rozeboom (1960), Meehl (1967), and Cohen (1994) through to book-length expositions (Harlow, Mulaik, & Steiger, 1997; Kline, 2004). Discussion in the L2 or foreign language learning (FLL) fields has been somewhat more sporadic, however recently we have seen an ongoing dialogue about the need for reform in statistical methods used for research. Recommendations include augmenting or even replacing the NHST with confidence intervals (CIs) and effect sizes and a greater emphasis on the graphical presentation of data. Other issues relate to how the parametric tests that are often used to test data may not be appropriate in many instances, especially in studies with small sample sizes. Nonparametric tests (Turner, 2014); what Cumming (2012) refers to as the new statistics of CIs, effect sizes, and meta-analysis; and techniques such as bootstrapping (LaFlair, Egbert, & Plonsky, 2015) or robust statistics (Larson-Hall & Herrington, 2009) are all positioned as useful alternatives to NHST practices (Norris, 2015). One useful tool that may help researchers better move towards realising these recommendations is the R statistical environment for computing (R Core Team, 2017).

What Is the Problem?

For many years, quantitative research in the FLL/L2 field that was aimed at identifying how groups of learners respond to differing learning treatments has focused on NHST outcomes. There are many well-documented arguments against the NHST approach (Cohen, 1994; Harlow et al., 1997; Kline, 2004; Meehl, 1967; Rozeboom, 1960), and one of the dominant critiques is about how the outcomes of the tests are often interpreted. The standard approach to carrying out such tests is to set a null hypothesis of no difference between the means of the samples under study at a suitable level of probability—usually $p = .05$. Then, once the data are collected and have met the appropriate assumptions for the test to be used (usually, having been drawn from a random sample representative of the population under study and representing a normal distribution), the statistical test is run and based on the outcome of this test, a decision is made to either accept or reject the null hypothesis. However, what seems to be often misunderstood by many researchers reporting the outcomes of significance tests is that a statistically significant result means only that the test has quantified the probability of the observed data or of more extreme data given that the null hypothesis test is true (Fidler & Loftus, 2009; Gigerenzer, Krauss, & Vitouch, 2004; Norris, 2015). Nothing has been found out about the probability of the hypothesis itself, and therefore the results of the

hypothesis test do not allow the researcher to make comments on the probability of the null or alternative hypothesis. Gigerenzer et al. (2004) reported this as a commonly encountered misunderstanding amongst psychology researchers and it is likely that this same kind of misunderstanding persists in FLL/L2 research, as acknowledged by Norris (2015), who noted that "statistical significance testing is probably not well-conceived or accurately interpreted in much L2 quantitative research to date" (p. 106).

A second criticism levelled at the NHST is that the null hypothesis can never be true because no two samples are ever exactly the same, so given a large enough sample size, it is possible to reject the null hypothesis (Kirk, 2007; Meehl, 1990; Plonsky, 2015).

Another problematic outcome of the statistical test(s) of data in this kind of situation is that the results present a pass–fail dichotomy based on a *p* statistic that implies research is an all-or-nothing endeavour wherein the aim is to break the elusive $p < .05$ barrier. In other words, the conclusions of the research are based on a statistical significance, which in many instances does not allow for extrapolation to substantive or practical significance. This may also lead to a bias in reporting practices that ignores results that do not reach the required *p* value even if these results have real-world significance. Furthermore, in many instances this kind of research is reported with little indication of considerations of the assumptions of the testing statistic being used, casting doubt on the veracity of the research report itself. There has been a promisingly growing call for changes to how quantitative research is assessed and reported in the L2 or FLL fields (see for example, Brown, 2015; Larson-Hall & Herrington, 2009; Norris, Plonsky, Ross, & Schoonen, 2015), mirroring similar recommendations in education and psychology research. These include

- a commitment to calculating and reporting effect sizes,
- a recognition of the importance of reporting CIs,
- a call for better reporting practices regarding descriptive statistics and more graphical presentation of data, and
- increased use of alternative statistical techniques such as robust statistics or non-parametric tests.

However, reform cannot happen without practice, and perhaps one barrier to reform is ongoing publication practices; another may be statistical illiteracy amongst L2/FLL researchers. What appears to be the dominant computer package used for data analysis in SLA and applied linguistics—SPSS (Loewen et al., 2014)—may also play a role here, as users may select default options when running an analysis using SPSS without thinking about—or understanding—exactly what they entail. One critique levelled at SPSS relates to its black box nature (Uprichard, Burrows, & Byrne, 2008); one can run a test on data without really understanding exactly what the test requires and if it is exactly what one should be doing with the data. Plonsky and Gonulal (2015) noted problems with the use of factor analysis in published L2 research, which they partly attribute to researchers likely choosing the default options of statistical packages in cases where these are not appropriate, but because they are set as the "default" it is assumed they are the ones to be used in the analysis (see also Uprichard et al., 2008). Another argument is that as SPSS made statistical data analysis "easier" there was a shift in the approach to teaching quantitative methods:

> In effect, prior to the arrival of SPSS on the PC, there had been a lot more concentration on both the philosophical nature of data—qualitative and quantitative—and with it, the role of the researcher in interpreting and constructing quantitative data. In contrast, post-1980s quantitative pedagogy places more time and focus on the output than the labours of the researcher. (Uprichard et al., 2008, p. 611).

Combined with low statistical literacy amongst language learning researchers (Loewen et al., 2014), this is likely to lead to a condition in which researchers will use what they are familiar with but may not be aware of alternative approaches or recommended best practices; reviewers of articles for publications may not be willing to question quantitative research methods due to a limited understanding of the techniques involved. So the status quo is perpetuated. Brown (2015) presented a number of reasons why language researchers should learn advanced techniques, and one step in this process is acquiring suitable tools to help with learning. One of his points was that learning about these statistical techniques may motivate researchers to learn how to use the R statistical environment, as this is probably the best tool for the process. I certainly agree with this perspective and would go further to argue that R can help researchers better understand their data, analyze it, and recognize what exactly they should report.

What is R?

R is an open-source statistical computing system. Being open-source means that, unlike commercial systems such as SPSS, R is free for anyone to download and use. Users and developers have contributed a large number of packages that can be plugged into the basic framework to extend its functionality as required. More recent development has also seen the introduction of packages for qualitative research, such as text-mining and corpus analysis. These open up the possibility that R may also be, or develop as, a viable alternative to commercial qualitative analytic tools such as Nivio.

R runs on all major computing platforms and requires nothing more than an Internet connection to download the required software and time to ensure that the researcher understands what he or she is doing when carrying out a particular statistical analysis. Because R requires entry of commands directly (i.e., it does not have a built-in graphical-based interface like SPSS, but packages do add this functionality), it does have a learning curve, but this is also useful in that it encourages users to think about their data and what exactly they want to get out of it. At least in my own experience, learning to use R has been very helpful in clarifying the stages of the data analysis and the outcomes obtained. These include

- exploring the data collected in research in the preliminary step,
- testing for assumptions required of traditional tests,
- recognizing the need to choose alternative options in cases when the assumptions are not met,
- understanding why these alternatives are appropriate, and finally
- generating more explanatory indices than just a measure of statistical significance (a *p* value) to help better explicate the outcomes of the research.

An Example Analysis

An example may help to demonstrate. For this purpose, a data set taken from Turner (2014, p. 162) will be utilized. This fictional data set is made up of test scores of children whose parents had received coaching on how to help with their homework and test scores of children whose parents had not received coaching. Because the data are from two independent groups (each with $n = 40$) and the comparison is focused on a single variable (test scores), the method of analysis selected by most researchers to test for any difference of means would likely be an independent *t* test. We want to test the null hypothesis of no difference in test scores between the two groups and set the alpha level at .05. This alpha level represents the likelihood of a type I error that one is willing to accept with the test—that is, the likelihood of making a claim about the outcome of the data being due to some intervention when in fact any difference between the samples is due to chance or sampling error. This alpha level needs to be decided prior to running the actual statistical analysis. Once the data have been entered into R (see Appendix), the *t* test can be performed as follows (the dependent variable is *score*, the independent variable is *coached*, and *data* is the data set containing the variables to be tested):

t test(score ~ coached, data = graphtxt).

This produces the result $t(71.63) = 3.22$, $p = .0196$. Recalling that this tells us the likelihood of our data given the null hypothesis, what we have found is an approximately 2% chance of getting the pattern of data observed in the sample under test if the pattern does not exist and the null hypothesis of no difference between the means of the sample is true (Norris, 2015). What we cannot say from the result is anything about the probability of the null hypothesis being true or false, or the probability of any alternative hypothesis being true.

The result of the statistical significance is of little use in showing if there is any difference between our two groups attributable to the different conditions. As well as being not particularly helpful, there are some other issues that need to be addressed. One is that the data have not been checked to confirm if they meet the assumptions required of a *t* test. Classical tests like *t* tests, ANOVA, and regression are based on the assumption that the data are from a normal distribution and are homogeneous. Violating these assumptions could result in the test in question failing to report results for which it was originally designed—for all intents, it may mean the researcher is carrying out an empty exercise that may or may not have any statistical significance, let alone any practical (or substantive) value. This is a well-discussed area in literature on research methodology, with some arguing that classical statistical methods used to analyse data are robust enough to deal with violations of normality and so forth. However, the argument is unresolved, and best practices would suggest that when choosing a particular statistical test for analysis, one would ensure that normality and homogeneity are addressed as part of the rationale for the choice of the test (Norris et al., 2015). This not only makes it easier for editors and reviewers to judge the appropriateness and validity of the chosen methodology but also for the intended audience to better draw informed conclusions from the research. Using an R package such as psych or pastecs allows for easy calculation and reporting of descriptive statistics, including skew and kurtosis, both of which are useful in indicating whether or not the data represent a normal distribution (Brown, 2016). Another way to check the normality and homogeneity of the data is with the Shapiro-Wilk test and Levene's test respectively (see, e.g., Field, Miles, & Field, 2012, or Turner, 2014), however there are some problems with these two tests when working with large samples (Field et al., 2012), as well as other issues that may influence their results (Erceg-Hurn & Mirosevich, 2008). Best practice would suggest not to rely solely on these methods when checking assumptions, but rather to augment them with plots of the data along with details of the skew and kurtosis of the sample. In the case of the *t*-test example, the pastecs package (Grosjean & Ibanez, 2014) can be used to provide descriptive statistics (see Appendix for the code for this, as well as subsequent routines). The key information from the resulting output is presented in Table 1. A careful reading of the data shows it is more informative than the results of the *t* test above, with higher mean and median scores and less variability for the coached group, suggesting that the coaching con-

dition may contribute to higher scores. For our immediate purposes, what should be noted are the columns labelled *skew.2SE* and *kurt.2SE*. If the values for these are above 1, it means the data are likely not normally distributed. Also look at the *norm test* columns, displaying the results of a Shapiro-Wilk test of normality—in this case there is a statistically significant result for the noncoached data that suggests the data are from a nonnormal population and therefore do not meet the assumptions required of a *t* test. Although the skew and kurtosis figures are both below 1, the kurtosis is quite high, suggesting the data are clustered around the lower or upper limits.

Table 1. Descriptive Statistics for the Example Data Set

Type	*n*	*M* (*SD*)	*Mdn*	95% CI	skew.2SE	kurt.2SE	norm test.W	norm test.p
Coached	40	85.25 (5.768)	85.00	[83.41, 87.1]	0.22	-0.57	0.97	.46
Non-coached	40	80.3 (7.845)	79.00	[77.79, 82.81]	0.28	-0.97	0.91	.004

Recalling the caution above about solely relying on the Shapiro-Wilks test to check assumptions, an additional check can be performed by plotting the data with a Q-Q plot or a histogram to see if it is normally distributed (see Appendix for code). The plots are given in Figure 1; it appears that the distribution is not normal, skewed instead towards the lower scores.

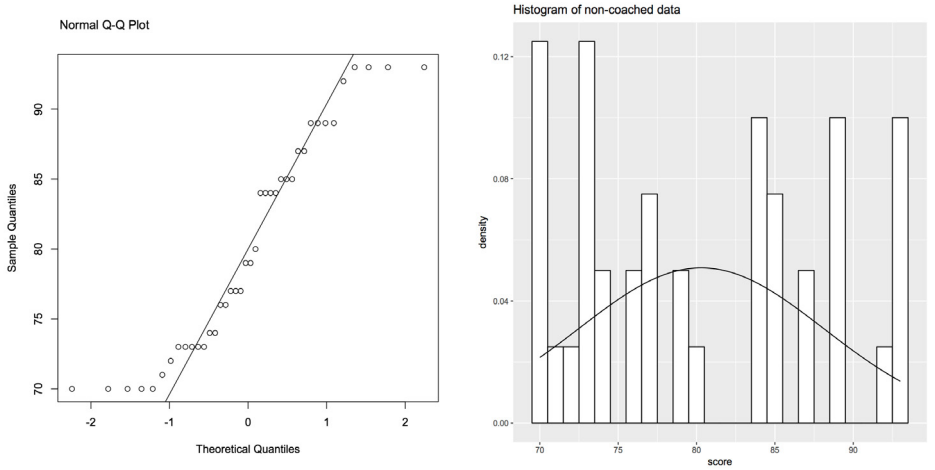


Figure 1. Q-Q plot and histogram for example data.

Levene's test can also be performed to test for the homogeneity of the data—or to see if the variances are equal. The results in Table 1 would suggest that they are not, as indicated by the difference in the standard deviations. To run Levene's test, the following code is executed:

```
> leveneTest(graphtxt$score, graphtxt$coached, center=median)
```

with the resulting $F = 7.705$, $p = .007$, which can be interpreted as showing a difference between the variances.

Overall, then, the evidence suggests that carrying out a standard *t* test is inappropriate. In this case, if the researcher wishes to carry out a test of statistical significance, the options are to use either a nonparametric test or to use robust statistics. Nonparametric tests, such as the Mann-Whitney test, the Wilcoxon rank-sum test for independent samples, or the Wilcoxon signed rank test for paired samples can be used as alternatives to the *t* test, and the Kruskal-Wallis and Friedman's test can be used in place of ANOVA (Field et al., 2012; Turner, 2014). As the two samples from the data in the example are independent, the Wilcoxon rank-sum test is appropriate; this test calculates the difference between the samples based on the medians rather than the means. The command to run in R is

```
> wilcox.test(score ~ coached, data = graphtxt, paired=FALSE, exact=FALSE,
correct=FALSE)
```

and the outcome shows that the coached students (*Mdn* = 85) had a significantly higher score than the noncoached (*Mdn* = 79), *W* = 1089, *p* = .005.

Robust statistics are an alternative to nonparametric tests (Field et al., 2012; Larson-Hall, 2015; Larson-Hall & Herrington, 2009); they are seen as preferable to nonparametric tests in a number of areas (Erceg-Hurn & Mirosevich, 2008; Mair & Wilcox, ca. 2016). In R, the WRS2 package is one option for running robust tests. For the example data, a robust version of the *t* test based on Yuen's method can be used (Mair, Schoenbrodt, & Wilcox, 2016). The code should be familiar:

```
> yuen(score ~ coached, data = graphtxt)
```

In this case, a version of the test with a 20% trimmed mean is being performed, and the outcome is *t*(39.3) = 2.55, *p* = .015.

As with the previous two versions of the *t* test, the results provide the probability of the data given the null hypothesis. However, this still does not provide any helpful information regarding the difference between the two test samples. To better understand whatever differences may exist between the two experimental groups, we can look at effect sizes, confidence intervals (CIs), and visual plots of the data. In regards to the latter, one particular strength of the R statistical package is its exceptional plotting functionality, extremely useful in helping with exploration and analysis of data.

Graphing in R: Ggplot

Although R has good basic graphing and plotting functionality, the ggplot2 package (Wickham, 2009) extends this substantially, allowing the user to build explanatory-rich plots of data. Graphical representation of data is especially useful for displaying large amounts of information in a way that is immediately and easily understandable (Hudson, 2015). Although a number of quite complex graphs and diagrams provide the potential to transmit detailed information, the main types of graphs used in the major L2 journals are the relatively simplistic line graphs and grouped bar charts (Hudson, 2015) perhaps because these are the easiest to produce in SPSS or Excel. Of course, if used appropriately these kinds of graphs can be very informative, especially if CIs are included. Figure 2 provides an example of a simple line graph with 95% CIs made using the example data set that we have been working with. It should be clear from Figure 2 that there is a difference between the means of the two groups, and the length of the CIs indicates greater variability in the noncoached group. The fact that the CIs do not cross, which can also be seen in Table 1, points to a statistically significant result.

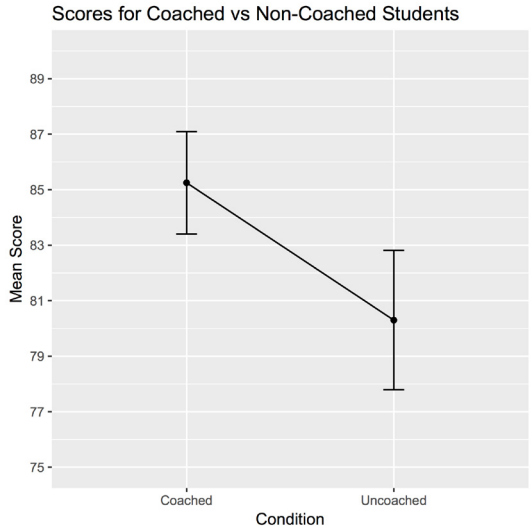
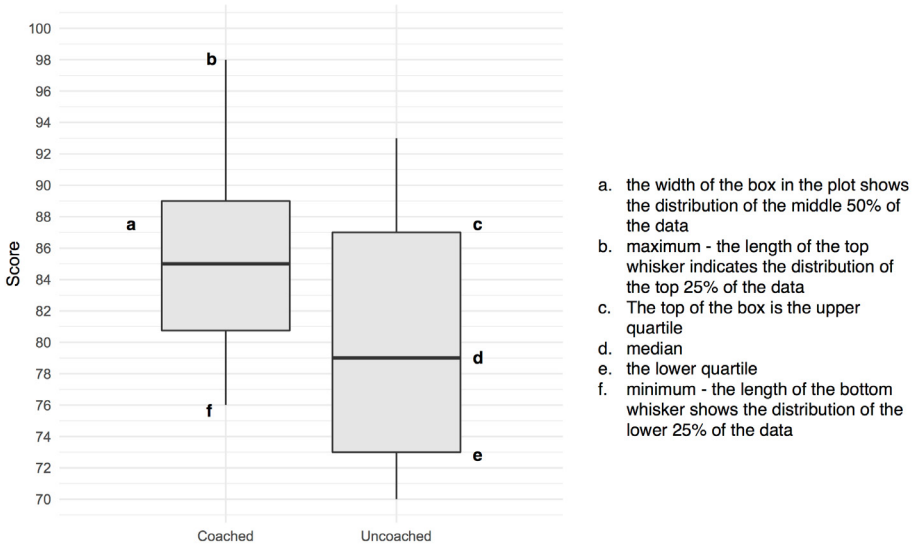


Figure 2. Example line graph showing error bars for confidence intervals.

A more informative way to plot and display the data is with a box plot (or box and whisker plot). A similar graph rarely seen in the language learning literature is the violin plot, this shows the same information as a box plot but also includes information on probability densities of the data. Both are quite straightforward to produce using R; the necessary code can be found in the Appendix. Examples of the output produced using the same data set *graphtxt* from Figure 2 grouped by the *coached* condition are presented in Figures 3 and 4. In the violin plot, the left graph includes each individual response, but the right-side plot incorporates a box plot demonstrating the kinds of options available with this versatile kind of plot. All of these graphs present a much clearer picture of the differences between the two treatment groups than a *p* value provides, and augmenting the graphs with a table of descriptive statistics and an effect size measurement would probably be a better way to report results.



a. the width of the box in the plot shows the distribution of the middle 50% of the data
b. maximum - the length of the top whisker indicates the distribution of the top 25% of the data
c. The top of the box is the upper quartile
d. median
e. the lower quartile
f. minimum - the length of the bottom whisker shows the distribution of the lower 25% of the data

Figure 3. Box plot example and code.

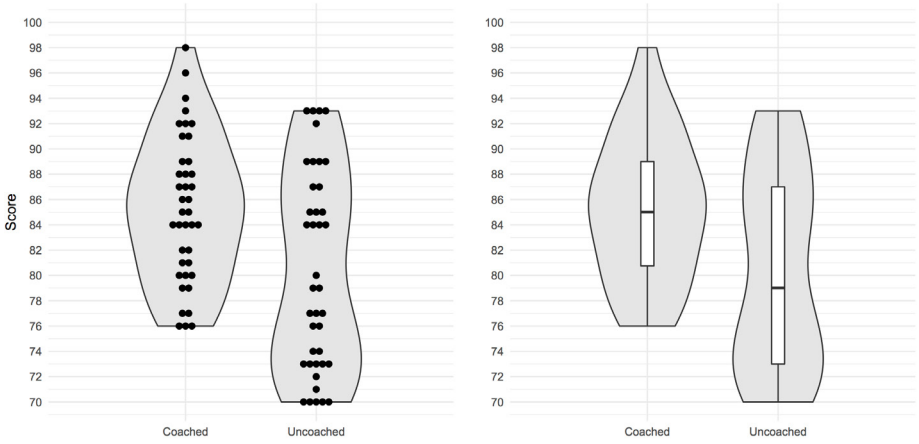


Figure 4. Violin plot examples and code.

Other packages in R also offer graphing or charting functionality to supplement their general functions, for example correlograms provided by the corrplot package or pairs panels from the psych package, which can be helpful when exploring the data.

Effect Sizes

Rather than the "either-or" mindset encouraged by reports of *p* values, effect sizes can give a broader understanding of the possible magnitude of the effect under study. Cohen (1988) suggested that effect sizes should be classified as small, medium, and large for his index of effect *d* (*d* = 0.2, *d* = 0.5, *d* = 0.8 respectively), which is one of the standard measures. Although he presented guidelines to determine these delineations that are commonly used today, he cautioned that the guidelines will vary depending on the substantive (practical) significance of the research outcome and researchers should ideally determine the import of an effect based on their own understanding of the conditions under study. The important point though is that calculating an effect size from a sample gives a better indication of what the researcher is trying to test than just reporting a *p* value. If a researcher finds a statistically significant result at $p < .05$ but a small effect size, it would suggest that the results were not strongly in support of the actual effect of the intervention, something not evident from just reporting statistical significance. On the other hand, a researcher may find a nonstatistically significant result from a test but still get a large effect size, suggesting further research into the effect under study is in order, something that may not be apparent if the *p* value is the sole arbitrator of a study's outcome.

Returning to the example data set, effect sizes can be calculated for each of the outcomes from the three different tests—the standard *t* test, the nonparametric Wilcoxon rank-sum test, and the robust *t* test. First, for the *t* test, effect size is not reported as part of the output in R, but it can be easily calculated. Using the compute.es package (Del Re, 2013), the code is

```
> tes(3.2151, 40, 40, level=0.05)
```

and the result is *d* = 0.72, a medium to large effect.

For the nonparametric test, the effect size also has to be calculated separately. This is slightly more complex, using a function from Field et al. (2012) in R, which gives the effect size *r* = -0.31. Finally, for the robust version of the *t* test, different options are available. One is to compute a robust version of Cohen's *d*. In this case, if variance between groups is not equal, it is possible to get effect sizes for the two groups and check if they are approximately equal. Otherwise, an alternative explanatory measure of effect size, ξ ,

can be used (Mair & Wilcox, ca. 2016). Given the unequal variance found in the example data, the following code can be run (as with the robust *t* test this is done with the WRS2 package):

```
> akp.effect(score ~ coached, data = graphtxt, EQVAR = FALSE)
```

The result is 0.80, 0.52, which gives different information about the size of the effect. As this is not helpful, ξ can be calculated instead:

```
> yuen.effect.ci(score ~ coached, data = graphtxt)
```

with an outcome of ξ = 0.46, a medium effect size.

As can be seen, each test gives a different kind of effect size, which is one challenge to calculating effect sizes—which one to calculate and how do they differ? Although there is not the space to go into this discussion here, the reader is directed to Ferguson (2009) and Kirk (2007) for guidance.

One other important role of effect sizes is to help with power analysis of a statistical study. The power of a statistical test is the probability that the test will correctly lead to the rejection of the null hypothesis (Cohen, 1988). It has been shown that many studies utilising NHST in other disciplines are underpowered and in fact may have sample sizes that are too small to adequately detect any differences between the samples (Gigerenzer et al.; Kline, 2004). It is unlikely that FLL/L2 research is any better here (Norris, 2015), so it is recommended that power analysis be carried out either prior to doing research or at least once research is completed. Here, knowing the sample size, desired *p* value for the significance test, and effect size allows one to calculate power using, for example, an appropriate R package or the tables provided by Cohen (1988). A conventional view is that aiming for a power of 0.8 is good practice (Cohen, 1988), but *p* values of .05 or .01 are usually used in power calculations. For the *t*-test example, the effect size, calculated with the R compute.es package (Del Re, 2013), is $d = 0.72$. Using the R pwr package (Champely, 2017), with $p = .05$, we find that the power of the test is .89. In other words, our *t* test has an 89% chance of correctly rejecting the null hypothesis.

It should be noted that in the case of a nonparametric test, if the data are not normally distributed, it is not possible to calculate the power of the test (Field et al., 2012), so in this case we cannot compare the power of the result of the Wilcoxon test with that from the *t* test. Similarly, it seems there is no solution available to calculate power for a bootstrapped version of the *t* test. Of course, if one is not reporting *p* values due to their limitations, then power is no longer a concern. However, overall the evidence seems to suggest that tests done with robust statistics have higher power than a standard parametric test (e.g., *t* test, chi squared, ANOVA, or other variation of the general linear model; Erceg-Hurn & Mirosevich, 2008), but a nonparametric test carried out on normally distributed data will have lower power than its parametric equivalent (Field et al., 2012).

Ideally, when possible one would carry out a power analysis prior to doing research, using effect sizes from previous research as a guide, or if this is not available, setting an effect size the researcher believes is acceptable. Aiming for increased power means that a larger sample size is required; similarly accepting a smaller *p* value or a smaller effect size requires increasing the sample size to achieve a suitable level of power. In the *t*-test example above, if the desired power was set at .9 with $p = .01$ and the effect size $d = 0.72$, a sample size of 59 in each group would be necessary to carry out the analysis; but if a smaller effect size was expected (say 0.4), a sample of 188 in each group would be required. The point that needs to be made here is that in research with small to medium effect sizes, large samples are required to correctly reject the null hypothesis. Many of the types of questions researched in FLL/L2 are likely to uncover only a small to medium effect, which brings up issues about the value of this research when the sample sizes are small.

Effect sizes are also useful as a standardised measure, as they allow for comparisons of research findings in meta-analytic studies, as well as making replications of research easier to carry out and report on. Replication is a key to progress in knowledge, a necessary step to show that an effect uncovered in one study may be applicable to more than one situation. Finding similar effects across differing conditions gives greater support for the substantive value of the effect than simply relying on *p* values. In terms of meta-analysis, having a standardised measure makes it easier to summarise previous research and show potentially promising trends in analysis. Not reporting effect sizes can mean that the research may not be useful in helping contribute to the aggregation of useful knowledge.

Although the American Psychological Association task force on statistical inference has stated, "always provide some effect-size estimate when reporting a *p* value" (Wilkinson, 1999, p. 599), one of the problems with reporting effect sizes is that it is often not really clear for the practitioner how to calculate and report effect sizes. SPSS, for example, does not make it obvious how to do so, and in fact earlier versions do not have the option to calculate some effect sizes, meaning the researcher needs to calculate them by hand. Although not so difficult to do, this could be seen as troublesome by some, and if the necessity for an effect size is not part of a journal's publication requirements, it may mean this particular measure is left out of the report. As can be seen from the examples used here, R does allow for relatively easy calculation of effect sizes. Alternative options exist; for example, Lebowitz (2015) suggested using the MBESS package to aid with effect size analysis. Other packages such as compute.es offer similar functionality.

Confidence Intervals

Already touched on in the discussion of plotting data, CIs are another important measure, as they give not just a guide to the statistical significance of one's findings, but also to the degree of confidence one has in these findings. CIs give a range around the sample mean within which there is a 95% probability (or whatever one wishes to set, though 95% tends to be standard) that the true population mean is likely to fall. In other words, they show an interval within which one can be relatively confident that the population mean exists for the effect one is testing for. They can also give an indication of the variability of the data—the wider the CI, the more variable the data, which may suggest problems with the sample—and help researchers design tests with greater precision or power (Coulson, Healey, Fidler, & Cumming, 2010). CIs can also serve the same purpose as a statistical significance test when comparing two or more point estimates. Look at the upper and lower boundaries of CIs level of significance (e.g., 95% CIs for an analysis in which the initial hypothesis has set *p* at .05). If the CIs do not cross, there is a statistically significant difference in the data, but if the CIs cross, it is likely (but not always the case, when they cross only slightly one may need to investigate further) there is a nonstatistically significant finding. This should be evident in Figure 2, where the lower limit of the CI for the coached sample ends at a higher point than the upper limit of the noncoached sample results. Of course, if the point is to use CIs to move away from the kind of dichotomous thinking that statistical significance testing is said to encourage, this interpretation of CIs would not be encouraged. One other advantage of CIs is that they can also help contribute to meta-analytic thinking—if studies are giving large CIs, "appreciating the large extent of uncertainty should encourage researchers to focus on cumulation of evidence over studies" (Coulson et al., 2010, p 2). To go back to an earlier point, here graphical representations of data are particularly useful, as presenting graphs of a distribution with CIs provides an easy way for the reader to see what a *p* value may not make apparent.

Overall Recommendations

In a real research situation, a final decision on the results of the data analysis would draw on more than just the outcomes of statistical tests, however it is hoped that the preceding discussion has shown that more than the calculation of a *p* value is necessary if statistical tests are to be used. However, it should be stressed that the logic behind CIs and effect sizes is based on similar reasoning to that used in the calculation of *p* values. If there is a problem with the data that may bias the *p*-value outcome, this will also adversely influence the effect size and CI calculations. If the initial exploration of data to be tested has shown that a classical parametric test is appropriate, the researcher should proceed with the appropriate analysis. The results should include a report of effect sizes and CIs, along with descriptive statistics, and as much as possible, explanatory plots of the data. If parametric tests are not appropriate, one choice may be to focus solely on descriptive statistics and plots of the data, as these can go a long way in informing readers about the import of ones findings, more so than a *p* value. The other option is to use nonparametric or, preferably, robust statistics and report effect sizes and CIs calculated for the particular test adopted. Whatever the decision, it is further hoped that the utility of the R system for the analysis has been demonstrated.

Where to Begin With R

With the aid of a good reference book, as well as guides on the Internet, it should not be too difficult to learn how to use R to carry out quite complex data analysis. Field et al. (2012) is recommended if the aim is to develop a better understanding of both statistical methods and the R system. Whilst written mainly as a textbook aimed at psychology students, it is nonetheless very comprehensive and lays out a clear framework for many different data analysis scenarios, including nonparametric and robust alternatives to standard classical tests. It does have a somewhat irreverent tongue-in-cheek tone that may be a bit off-putting to some, but that does not distract from its overall utility. Two other recent R guidebooks are both written with the language researcher in mind. Turner (2014) focuses on nonparametric tests for small-scale research such as classroom research or action research projects. Larson-Hall (2015) is a recent guide to using both SPSS and R, with a particular emphasis on robust methods. There is also an earlier publication available that focuses solely on R (Larson-Hall, n. d.). Finally, as with any popular open-source application, R has numerous online guides and discussion forums, providing details on how to carry out different kinds of procedures.

Conclusion

It must be stressed, any system is open to errors or abuse, and just because researchers do their data analyses in R does not mean they will automatically get the "right" results. However, reiterating Brown (2015), R is probably the best tool for the process. It is readily available and provided the researcher is willing to move beyond simple reports of *p* values, it offers a powerful framework to explore and analyse the data collected. When taking into consideration the commitment of time and resources involved in running much research, common sense suggests the researcher will want to uncover as well as possible the patterns and trends in the data. Statistical significance testing is not much

help here, but it is hoped that the information presented in this paper will be an aid for those looking to move towards better quantitative analysis.

Bio Data

Paul Collett works at Shimonoseki City University. He is interested in research epistemology and methodology and the psychology of language learners. <collett@shimonoseki-cu.ac.jp>

References

Brown, J. D. (2015). Why bother learning advanced quantitative methods in L2 research? In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 9-20). New York, NY: Routledge.

Brown, J. D. (2016). *Statistics corner: Questions and answers about language testing statistics.* Tokyo: JALT Testing and Evaluation SIG.

Champely, S. (2017). *Pwr: Basic functions for power analysis* (R package Version 1.2-1) [Software package]. Retrieved from https://cran.r-project.org/package=pwr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology, 1,* 1-9.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Del Re, A. C. (2013). *Compute.es: Compute effect sizes* (R package Version 0.2-2) [Software package]. Retrieved from http://cran.r-project.org/web/packages/compute.es

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63,* 591-601.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40,* 532-538.

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace *p* values. Some conceptual arguments and empirical demonstrations. *Zeitschrift für Psychologie* (Journal of Psychology)*, 217,* 27-37.

Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. London, England: Sage.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391-408). Thousand Oaks, CA: Sage.

Grosjean, P., & Ibanez, F. (2014). *Pastecs: Package for analysis of space-time ecological series* (R package Version 1.3-18) [Software package]. Retrieved from https://CRAN.R-project.org/package=pastecs

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* New York, NY: Psychology Press.

Hudson, T. (2015). Presenting quantitative data visually. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 78-105). New York, NY: Routledge.

Kirk, R. E. (2007). Effect magnitude: A different focus. *Journal of Statistical Planning and Inference, 137,* 1634-1646.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

LaFlair, G. T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, *t* tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46-77). New York, NY: Routledge.

Larson-Hall, J. (2015) *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). Oxford, England: Routledge.

Larson-Hall, J. (n.d.). *A guide to doing statistics in second language research using R.* Retrieved 3 January, 2017, from http://cw.routledge.com/textbooks/9780805861853/guide-to-R.asp

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31,* 368-390.

Lebowitz, A. (2015). Require confidence intervals for effect size estimates in JALT Journal. *The Language Teacher*, *39*(2), 42.

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly, 48,* 360-388.

Mair, P., Schoenbrodt, F., & Wilcox, R. (2016). *WRS2: Wilcox robust estimation and testing* (R package Version 0.9.1) [Software package]. Retrieved from https://cran.r-project.org/web/packages/WRS2/

Mair, P., & Wilcox, R. [ca, 2016] *Robust statistical methods in R using the WRS2 package*. Retrieved from https://cran.r-project.org/web/packages/WRS2/vignettes/WRS2.pdf

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195-244.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103-115.

Norris, J. M. (2015) Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning, 65*(Suppl. 1), 97-126.

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning, 65,* 470-476.

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* (pp. 23-45). New York, NY: Routledge.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning, 65*(Suppl. 1), 9-36.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416-428.

Turner, J. L. (2014). *Using statistics in small-scale language education research.* New York, NY: Routledge.

Uprichard, E., Burrows, R., & Byrne, D. (2008). SPSS as an "inscription device": From causality to description? *The Sociological Review, 56*, 606-622.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Appendix

R Code Snippets

Importing Data Into R

To import the data from Turner (2014) into R, it is probably easiest to save the data as a comma-separated or tab-separated file. Enter the data in a spreadsheet in Microsoft Excel or a similar application, with three columns:

Student	coached	score
1	1	76
2	1	76
3	1	76
4	1	77
.		
.		
41	2	70
42	2	70
43	2	70
44	2	70

Export this to the appropriate format (in this case, a text-delineated text file named *data.txt*) and use the following code in the R console to load the data into the dataframe *graphtxt* that will then be used for the subsequent analysis. Ensure the path to the file is set correctly:

```
> graphtxt <- read.delim("~/path to file/data.txt", header=TRUE, sep="\t")
```

you can check that the data loaded correctly by viewing the first few lines:

```
> head(graphtxt)
```

If all looks correct, you can then work with the dataframe for the analysis.

Code for Descriptive Statistics

The pastecs package provides a useful function to calculate descriptive statistics. As this package is not part of the core R installation, it needs to be enabled; this can be done simply with the *library* command.

```
> coached <- subset(graphtxt, graphtxt$coached == 1)

> noncoached <- subset(graphtxt, graphtxt$coached == 2)

> library(pastecs)

> Coached <- round(stat.desc(coached$score, norm=TRUE), digit=3)

> NonCoached <- round(stat.desc(noncoached$score, norm=TRUE), digit=2)
```

Checking for Normality of Data

The R code for the Q-Q plot, which shows the data in relation to a line representing a normal distribution, is:

```
> qqnorm(noncoached$score)

> qqline(noncoached$score)
```

and for the histogram (using the ggplot2 package)

```
> library(ggplot2)

> hist2 <- ggplot(noncoached, aes(score))

> hist2 + geom_histogram(aes(y= ..density..), binwidth= 1, colour = "black", fill = "white") + labs(title="Histogram of noncoached data") + stat_function(fun = dnorm, args = list(mean=mean(noncoached$score), sd= sd(coached2$score)), colour = "black")
```

Graphing Data

All the following graphs are produced using ggplot2. The code for Figure 2 is as follows:

```
> linegraph1 <- ggplot(graphtxt, aes(coached, score)) + stat_summary(fun.y = mean, geom = "point") + stat_summary(fun.y = mean, geom = "line", aes(group=1)) + stat_summary(fun.data = mean_cl_normal, geom = "errorbar", width = 0.1) + theme(legend.position = "none") + ylab("Mean Score") + ggtitle("Scores for Coached vs Noncoached Students") + coord_cartesian(ylim = c(75, 90)) + scale_x_discrete(name="Condition", limits=c("Coached", "Uncoached")) + scale_y_continuous(breaks=seq(75, 90, 2))

> linegraph1
```

The code to generate the boxplot shown in Figure 3 is as follows:

```
> library(ggplot2)

> p <- ggplot(graphtxt, aes(x=coached, y=score, group=coached))

> p <- p + geom_boxplot(fill="gray90") + scale_x_discrete(name = "", limits=c("-Coached", "Uncoached")) + scale_y_continuous(name = "Score", breaks = seq(70, 100, 2), limits=c(70, 100)) + theme_minimal() + scale_colour_manual(values = c("#E69F00", "#56B4E9"))

> p
```

The code for the violin plots shown in Figure 4 is as follows:

```
> p1 <- p + geom_violin(fill="gray90") + scale_x_discrete(name = "", limits=c("-Coached", "Uncoached")) + scale_y_continuous(name = "Score", breaks = seq(70, 100, 2), limits=c(70, 100)) + theme_minimal() + geom_dotplot(binaxis='y', stackdir='center', dotsize=.5, binwidth = 1)

> p1


> p2 <- p + geom_violin(fill="gray90") + scale_x_discrete(name = "", limits=c("-Coached", "Uncoached")) + scale_y_continuous(name = "", breaks = seq(70, 100, 2), limits=c(70, 100)) + theme_minimal() + geom_boxplot(width=0.1)

> p2
```