Classic Non-Parametric Statistics

> [I]t is easy to . . . throw out an interesting baby with the nonsignificant bath water. Lack of statistical significance at a conventional level does not mean that no real effect is present; it means only that no real effect is clearly seen from the data. That is why it is of the highest importance to look at power and to compute confidence intervals.
>
> William Kruskal (1978, p. 946)

Throughout the main book I was constantly using non-parametric statistics. Non-parametric statistics include bootstrap analyses and other types of robust statistical tests. However, in this chapter I will give information about non-parametric statistics that mostly use ranking to estimate location instead of mean scores. These are the types of non-parametric statistics that have been in place for many years, are considered classics, and are one way to avoid the influence of outliers. I personally would rather turn to bootstrap analyses or means trimming as ways of looking at data that does not fulfill the ideals of parametric statistics, but I realize some of my readers may be interested in these classic rank-based statistics and so I present them briefly in this chapter.

Why Use Non-Parametric Statistics?

The first question I want to answer at the outset of this paper is why you would use non-parametric statistics. Non-parametric statistics are also called distribution-free statistics (Howell, 2002) because they do not require that the data be normally distributed. Maxwell and Delaney (2004) note that it is not accurate to say that nonparametric tests do not assume homogeneity of variances, however. They point out that classic non-parametric tests like the Kruskal-Wallis assume that population distributions are equal, which would clearly imply that variances are

equal as well. Other requirements of non-parametric tests are that sampling is random and observations are independent.

But wait a minute, you might say. Did we see that most of the datasets used in this book didn't satisfy the assumptions of parametric statistics? They weren't perfectly normal, because they weren't perfectly symmetrically distributed and/or they contained outliers. They often violated the assumption that the variances of the groups would be equal. And yet the authors of these studies continued to use parametric tests. So is it or isn't it OK to just use parametric statistics even when your data do not satisfy the assumptions of parametric statistics? It's hard to get a straight answer to this question when you consult the statistical experts. Some authors claim that parametric statistics are robust to violations of the assumptions, while others claim that even small violations can spell certain doom (OK, not certain doom, but cause you to conclude that there are no differences between groups or no relationship between variables when they do in fact exist). Statistical simulation studies have shown that problems with skewness, unequal variances, and outliers can have large effects on the conclusions you draw from statistical tests (Wilcox, 1998).

You already know that I prefer the robust statistical methods that were presented as alternatives in every chapter of the book, which were thought up early in the twentieth century but have only become possible since the advent of strong computing power (Larson-Hall & Herrington, 2009). Howell predicted in 2002 that robust methods would soon "overtake what are now the most common nonparametric tests, and may eventually overtake the traditional parametric tests" (p. 692), and to my way of thinking that day has come. The reason parametric and so-called non-

parametric statistics (the ones I will show you in this paper) were the ones that became well known was because their computing requirements were small enough that people could compute them by hand. The usually unreasonable assumptions of the parametric statistics were put in place so that the statistical test would be much easier to compute by hand. Thus, there may not be strong reasons to use these classic types of non-parametric tests anymore, but I also do not know of any sources that assert that robust methods are preferable to these classic non-parametric statistics in all cases, so I am presenting that information here.

Some authors assert that non-parametric statistics are less powerful than parametric statistics, but that is not always true. It really depends upon the problems that are found in the distribution of the data. If there are outliers, then a non-parametric test, which uses the median which is insensitive to outliers, might result in more power to find a statistical result than a parametric test.

Non-Parametric Statistics Tests

Table 1 lists the parametric counterpart to a number of non-parametric tests. The Spearman rank order correlation is also a non-parametric alternative to the parametric Pearson correlation, but this test has already been mentioned in Chapter 6 on correlation so I won't discuss it further in this paper. The last 4 tests in Table 1 are the ones that I will consider in this paper.

Non-parametric test	Parametric counterpart	Statistic used
Chi-square	–	χ^2
Binomial	–	p-value returned
Runs	–	p-value returned
1-sample K-S	–	p-value returned
2 independent samples	Independent-samples t-test (Chapter 8)	Mann-Whitney U or Wilcoxon rank-sum test W
K independent samples	One-way ANOVA (Chapter 9)	Kruskal-Wallis χ^2
2 related samples	Paired-samples t-test (Chapter 8)	Wilcoxon signed ranks test Z
K related samples	RM ANOVA with only one within-subject independent variable (Chapter 11)	Friendman χ^2

Table 1 Non-parametric tests and their parametric counterparts.

There are four non-parametric tests listed in Table 1 that I will not cover in this paper. The **chi-square test** is a non-parametric test, and information about that test can be found in the online chapter "Chi-square." There is no parametric alternative to the test.

The **binomial test** examines the proposition that the proportion of counts that you have fits a binomial distribution. It starts with the assumption that either of two choices is equally likely, although one can change that proportion to fit the circumstances. In the online chapter x "Chi-square" I explained how to use this test.

The **runs test** is designed to test whether a categorical level of your variable (with only two levels) is randomly distributed in your data. For example, you could use the runs test to see whether males and females were randomly distributed in your sample. This test is not frequently

used in the second language research field so I will not demonstrate how to use it in this paper.

The one-sample **Kolmogorov–Smirnov** test is sometimes used to test whether a variable has a normal distribution. This test can also be used to compare the distribution of a variable with other distributions besides the normal distribution, such as the Poisson distribution. As I have discouraged the use of such tests throughout this book because they are usually not sensitive enough to detect deviances from the normal distribution with small sample sizes and too sensitive to deviances for large sample sizes, I also will not demonstrate how to use this test in this paper.

Non-Parametric Statistics Tests in SPSS

The place to go to find ways to analyze statistics with non-parametric methods in SPSS is the ANALYZE > NONPARAMETRIC TESTS menu, shown in Figure 1. The first menu shows four choices. The first three choices (ONE SAMPLE, INDEPENDENT SAMPLES, and RELATED SAMPLES) are more general choices with a kind of statistical wizard that tries to guide you to the test you need. The fourth choice, LEGACY DIALOGS, allows you to directly choose the test you want, so I will be using this menu. You see eight different non-parametric tests in the LEGACY DIALOGS area, but I will only discuss the last four in this paper.

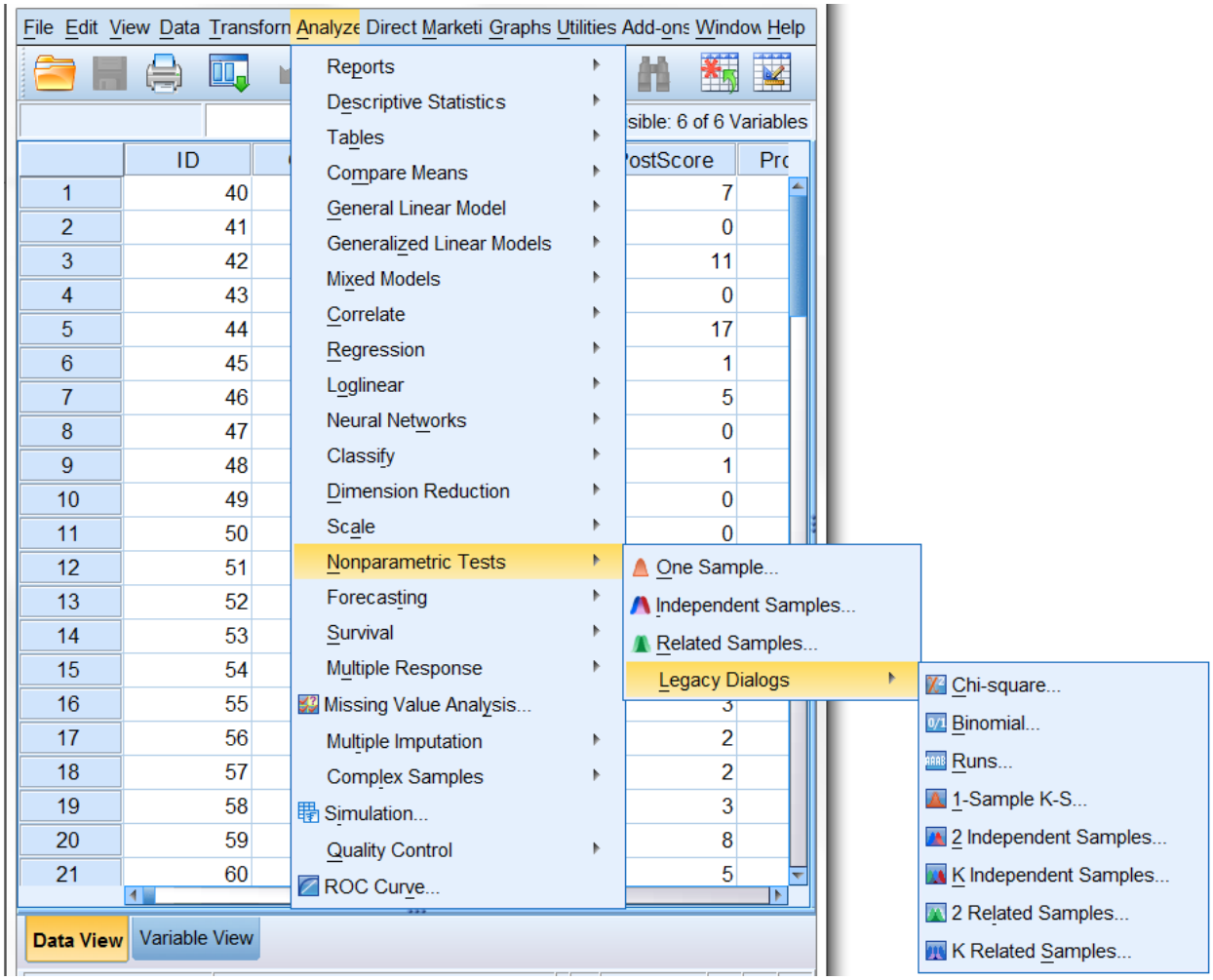

Figure 1 Non-parametric tests available in SPSS.

Non-Parametric Statistics Tests in R

R Commander is able to conduct non-parametric statistics on the last 4 tests listed in Table 1, so I will use R Commander to do this. Figure 2 shows the choices in the STATISTICS > NON-PARAMETRIC TESTS menu in R Commander. There are more choices of non-parametric tests available in R, but this chapter will focus on these 4 basic ones.

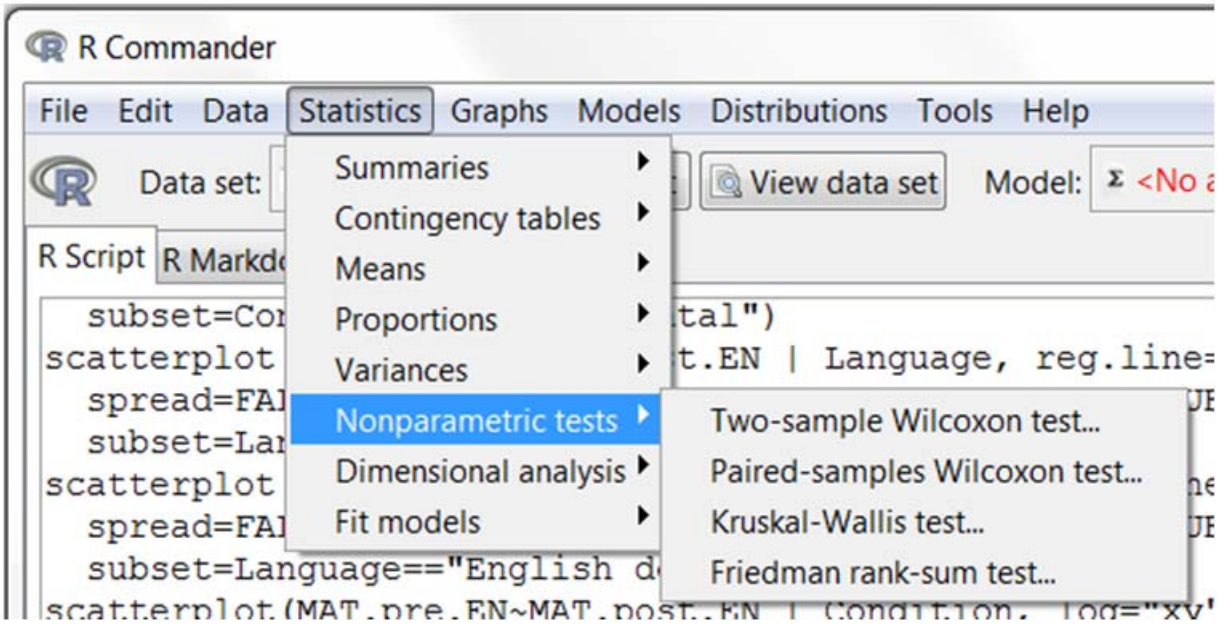

Figure 2 Non-parametric tests available in R Commander.

Non-Parametric Alternative to the Independent-Samples T-Test

Use an independent-samples *t*-test when you have two mean scores from two different groups or, in other words, two levels in your independent variable. In Chapter 8 I illustrated the use of the independent-samples *t*-test with Leow and Morgan-Short's (2004) study of comprehension ability when participants had to engage in a think-aloud task. I used a parametric test for the recognition post-score test and found a non-statistical difference between the think-aloud and non-think-aloud groups with a $p = .105$. Actually, Leow and Morgan-Short analyzed this variable with the non-parametric **Mann–Whitney U test** because their data did not fit the assumptions of a parametric test. Let's take the same variable and see if we get the same results as obtained in Chapter 8 with the parametric test (use the LeowMorganShort.sav file). The Mann-Whitney test (also known as the Wilcoxon's rank-sum test) tests the null hypothesis that "the two samples were drawn at random from identical populations (not just populations with the same mean)" (Howell, 2010, p. 673), so that rejection of the null hypothesis is "generally interpreted to mean

that the two distributions had different central tendencies, but it is possible that rejection actually resulted from some other difference between the populations" (ibid). Howell (2010) notes that we are not as certain in the result when we reject the null hypothesis of the non-parametric test, and this is the trade-off that we get for not having to fulfill all the assumptions of a parametric test.

The Mann–Whitney Test in SPSS (Two Independent Samples)

Go to ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > 2 INDEPENDENT SAMPLES.

Move RECPOSTSCORE to the "Test Variable List" and GROUP to the "Grouping Variable" box. Just as with the independent-samples *t*-test, you'll need to define the groups with numbers before moving on. I just called them 1 and 2, as shown in Figure 3. Notice that you have several choices for the type of test that you will use, but just keep the check on the default box, the "Mann–Whitney U." This test is exactly the same as the Wilcoxon Rank Sum Test that returns a statistic of W (Howell, 2002), in case you ever see that result and wonder what test it is. Open the EXACT button if you want to calculate the exact *p*-value of the test. You can choose how many minutes you want to let the exact calculation run, and the number of iterations will show at the bottom of the Output document. By default SPSS will use the asymptotic method, which is quicker if your dataset is large, but sometimes not as exact. If you have a small sample (less than 20 cases all together) SPSS will automatically calculate both the Asymptotic method and the Exact method. If you want a more exact *p*-value but you have a large sample, you can try the Monte-Carlo test method, as it will be quicker than the Exact method. The descriptive statistics returned by the choice of "Descriptive Statistics" in the OPTIONS button are not divided by group, so there is no need to open that button. The test will automatically return mean ranks (not the mean score) divided by groups without calling for descriptive statistics.

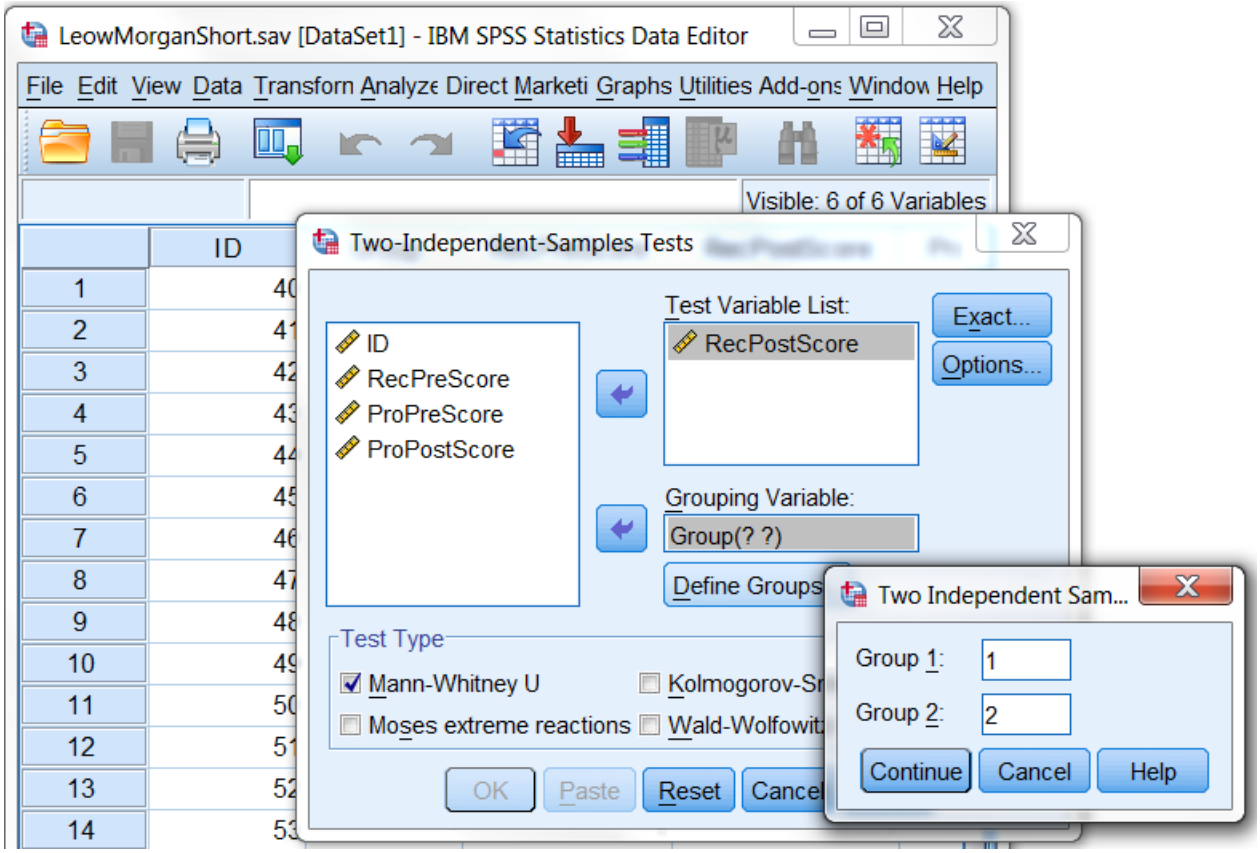

Figure 3 The Mann-Whitney test in SPSS (non-parametric alternative to the independent-samples *t*-test).

Your output will not show mean scores as it does not use mean scores, but instead will show mean ranks. The non-parametric test will rank the data for the whole dataset and then compare whether the ranks divided up by groups are different from the rank for the whole set. In this case, the mean rank of the non-think-aloud group is lower than the mean rank of the think-aloud group, but they are fairly similar (as shown in Table 2). The Test Statistics table in Table 2 shows the U-value (663.5) and the associated *p*-value ($p = .424$). Because the *p*-value is above $p = .05$, we cannot reject the null hypothesis that there is no difference between groups. The result is the same as with the parametric test we saw in Chapter 8—we conclude that there is no

difference between the groups.

Mann-Whitney Test

Ranks

	Group	N	Mean Rank	Sum of Ranks
RecPostScore	Non Think Aloud	39	37.01	1443.50
	Think Aloud	38	41.04	1559.50
	Total	77		

Test Statistics[a]

	RecPost Score
Mann-Whitney U	663.500
Wilcoxon W	1443.500
Z	-.799
Asymp. Sig. (2-tailed)	.424

a. Grouping Variable: Group

Table 2 Results from the Mann-Whitney U Test in SPSS (alternative to independent-samples *t*-test).

The Wilcoxon U Test in R (Two Independent Samples)

In Chapter 8 you imported the LeowMorganShort.sav file as leow into R. Use that data (or reimport it) and go to STATISTICS > NON-PARAMETRIC TESTS > TWO-SAMPLE WILCOXON TEST. This test is exactly the same as the Mann-Whitney Test that returns a statistic of U (Howell, 2002). Split your data by the group variable, and choose recpostscore under "Response Variable." In the OPTIONS tab, you can choose to have a two-sided or a one-sided hypothesis. You can also choose to have an exact test, which calculates an exact *p*-value but may take longer to compute than the default method, which is the normal approximation if your dataset is not quite small (in which case the exact method is used). There is an option to use a normal

approximation with a continuity correction. By default, if the program uses a normal approximation, it will include the **continuity correction**.

Here is the R code for this test:

```
wilcox.test(recpostscore ~ group, alternative="two.sided", data=leow)
```

The syntax looks like a regression equation, with the response variable listed first, then modeled by the group division. The output looks like this:

```
Warning in wilcox.test.default(x = c(7, 0, 11, 0, 17, 1, 5, 0, 1, 0, 0,  :
  cannot compute exact p-value with ties

        Wilcoxon rank sum test with continuity correction

data:  recpostscore by group
W = 663.5, p-value = 0.427
alternative hypothesis: true location shift is not equal to 0
```

The default method is to try to calculate the exact *p*-value, but in this case because there are ties, the exact *p*-value cannot be computed, so it is approximated based on the normal approximation. The *p*-value shows that this test is not statistical and we cannot reject the null hypothesis that there is no difference between groups. The result is the same as with the parametric test we saw in Chapter 8—there is no difference between the groups.

By the way, if you want to be able to choose the Monte Carlo method of calculating *p*-values (this was available in SPSS), or if you want the Z-score for calculating the effect size (see the next section in this document) you can try the wilcox_test( ) in the coin package (remember, to

install this package type install.packages("coin") in the R console and then type library(coin) to open the package). Use the argument distribution=c("approximate") to call for the Monte Carlo method of calculation.

```
wilcox_test(recpostscore ~ group, alternative="two.sided", distribution=c("approximate"), data=leow)
```

```
        Approximative Wilcoxon Mann-Whitney Rank Sum Test

data:  recpostscore by group (Non Think Aloud, Think Aloud)
Z = -0.7995, p-value = 0.449
alternative hypothesis: true mu is not equal to 0
```

Here is the analysis of this command:

wilcox_test(recpostscore ~ group, alternative="two.sided", distribution=c("approximate"), data=leow)

wilcox.test(x ~ y)	The command is set up to evaluate the dependent variable (x) modeled as a function of an independent variable (y)
alternative="two.sided"	This specifies a two-sided hypothesis and is the default (so you don't actually need to type it); for a one-sided hypothesis, use either greater or less. For these tests greater means "true location shift is greater than zero," and for less it means "true location shift is less than zero."
distribution=c("approximate")	The default method of calculating the *p*-value

	depends on the size of the sample; for small samples the exact method will be used, but one can always force the command to calculate the exact method (use exact), although it may take quite some time to calculate depending on the size of the dataset, and cannot be done if there are ties in the data. Other choices include approximate(B=9999), which calculates based on a Monte Carlo approximation (number of replicates set by the B); or (asymptotic), based on the normal distribution. If you don't type anything, the default will be used.
data= leow	Specifies what dataset should be used

Effect Size for the Mann–Whitney or Wilcoxon Test

You can calculate an effect size for any non-parametric test which returns a z-score (the capital "Z" in the output in Table 2 or the wilcox_test( ) output in the preceding section of this document) by using the following equation to turn it into a percentage variance measure of r:

$$r = \frac{Z}{\sqrt{N}}$$

where N = the total number of observations (Rosenthal, 1991, p. 19). The sign does not give any helpful information so just use the absolute value. For the Mann– Whitney U test we did for the Leow and Morgan-Short variable of recognition posttest, $r = \frac{.8}{\sqrt{77}} = .09$, a small effect size. Remember, this is an *r*-family effect size, which is a percentage variance effect size, not the *d*-family effect size, so we can estimate the percentage of variance it explains by squaring it.

Non-Parametric Alternative to the One-Way ANOVA

Use a one-way ANOVA when you have three or more levels of your independent variable or, in other words, you want to compare three or more mean scores on one dependent variable. In Chapter 9 I illustrated the use of the one-way ANOVA with the Ellis and Yuan (2004) dataset, which looked for differences in groups that received differing amounts of planning and writing time (EllisYuan.sav, imported into R as EllisYuan). We examined group differences with the dependent variable of how much syntactical variety was found in each participant's writing sample. With the one-way ANOVA we found a statistical result ($F_{2,39} = 9.05$, $p = .0006$), and further post-hoc tests showed that the pre-task planning (PTP) group was better than both the online planning (OLP) and no-planning (NP) groups. Let's see what happens when we use the non-parametric alternative, the **Kruskal–Wallis test**. The null hypothesis tested by the Kruskal–Wallis test is that "all samples were drawn from identical populations" (Howell, 2010, p. 683).

Kruskal-Wallis H Test in SPSS (Alternative to the One-Way ANOVA)

To call for the test, go to ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > K INDEPENDENT SAMPLES. Move the SYNTAXVARIETY variable to the "Test Variable List" and GROUP to the "Grouping Variable" box. Figure 4 shows that you can't continue until you define your groups. Click the box under "Grouping Variable" that says DEFINE RANGE. For

"Minimum," enter the number of the lowest level of your group (I put 1), and in "Maximum" put the number of the highest level of your group (I put 3 because I have 3 groups). Press CONTINUE. Leave the "Kruskal-Wallis H" test box ticked. Use the Jonckheere–Terpstra test to get more power if there is some a priori ordering to your groups (this is only available if you have the Exact Tests add-on module for SPSS). If you open the EXACT button you will be able to choose to calculate an exact *p*-value (see the explanation in the section of this chapter called "Why Use Non-Parametric Statistics?" about the different options here). The descriptive statistics do not split the data up between groups so I do not see any use calling for them in the OPTIONS button, as the regular command will call up mean rankings divided by groups.

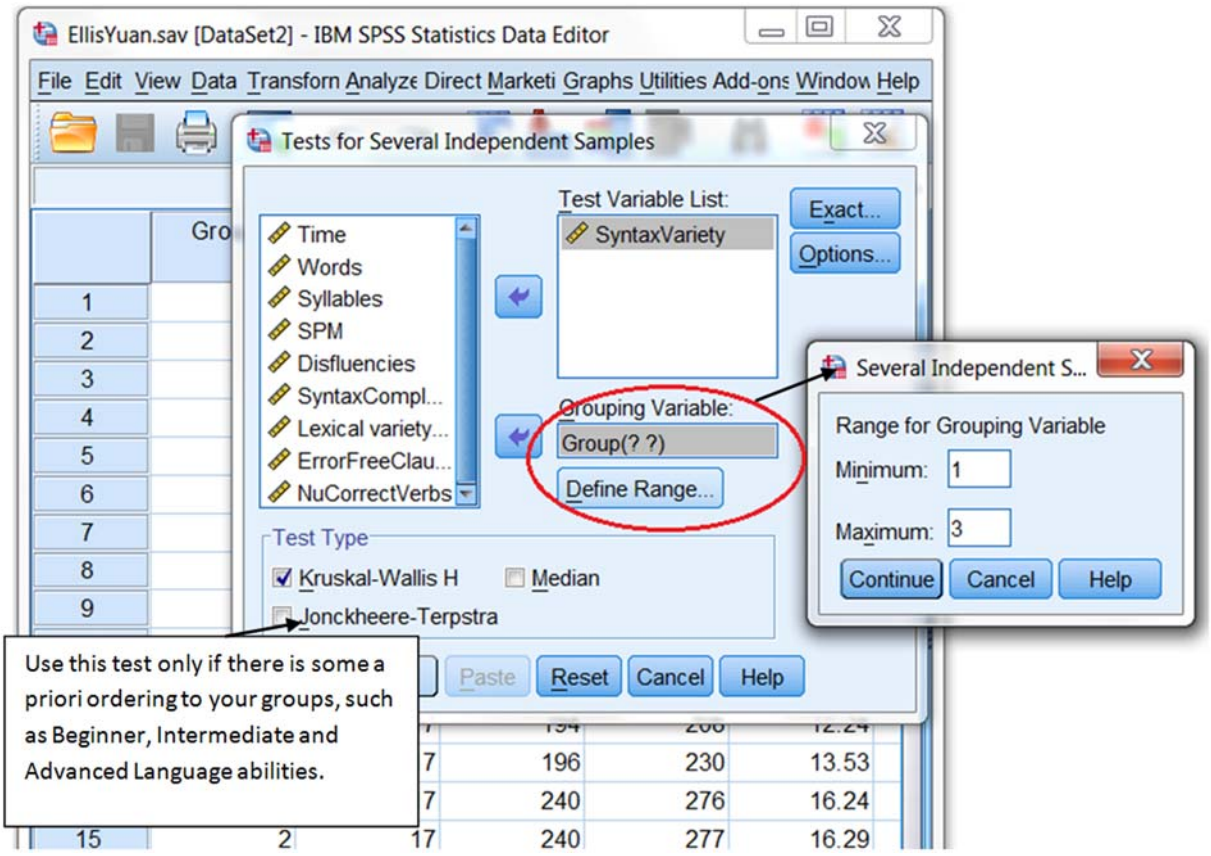

Figure 4 The Kruskal–Wallis test in SPSS (alternative to the one-way ANOVA test).

The Kruskal–Wallis test is an extension of the Mann–Whitney to the case of more than two levels, so the same type of ranking is taking place in this test and the first table of the output will show the rankings of the groups. The highest ranking is for the PTP group at 30.3; then there is 20.9 for the OLP group, and 13.3 for the NP group. The output, shown in Table 3, returns a chi-square statistic that has a probability of $p = .001$ at 2 degrees of freedom (for the Jonckheere–Terpstra test you will get a J-H statistic and an associated *p*-value, which you can use in the same way). We conclude that there are statistical differences between the three groups. This is the same conclusion we drew from the parametric tests.

Kruskal-Wallis Test

Ranks

	Group	N	Mean Rank
SyntaxVariety	NP_No planning	14	13.32
	PTP_PretaskPlanning	14	30.32
	OLP_OnlinePlanning	14	20.86
	Total	42	

Test Statistics[a,b]

	SyntaxVariety
Chi-Square	13.635
df	2
Asymp. Sig.	.001

a. Kruskal Wallis Test

b. Grouping Variable: Group

Table 3 Results from the Kruskal–Wallis Test (alternative to one-way ANOVA).

There's just one problem, which is that the Kruskal–Wallis test does not provide post-hoc tests in the same way as the one-way ANOVA did, so we can't be sure which groups are statistically different from one another, and this is probably something we want to figure out. If you want to

keep using a rank-based non-parametric test, your option would be to run Mann-Whitney tests and pit only two groups at a time against each other. Since there are three groups, you would need to do three tests. For three tests I won't worry about the familywise error rate, but if you run a large number of tests you might consider doing something to minimize the error rate, such as using the FDR adjustment. Alternatively, although the Bonferroni is too conservative, it is easy to calculate—simply divide your alpha (.05) by the number of tests that you are using and that is your critical value.

If you go back to the Mann–Whitney dialogue box (to ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > 2 INDEPENDENT SAMPLES), move the SYNTAX VARIETY variable to the right as your dependent variable, and then the GROUP variable to the "Grouping Variable" box. This box gives you a place to define groups (see Figure 2), so we'll first put in Groups 1 and 2, then 1 and 3, and then 2 and 3 using the same independent and dependent variables that we used for the Kruskal–Wallis test. A faster way to do this would be to put the correct variable in the boxes for the Kruskal–Wallis test, define Groups 1 and 2, and then push the PASTE button. The syntax for the first comparison will be shown. Copy the line starting at NPAR TESTS twice more, changing the group numbers to cover all of the permutations needed, as shown in Figure 5. Then choose RUN > ALL from the menu.

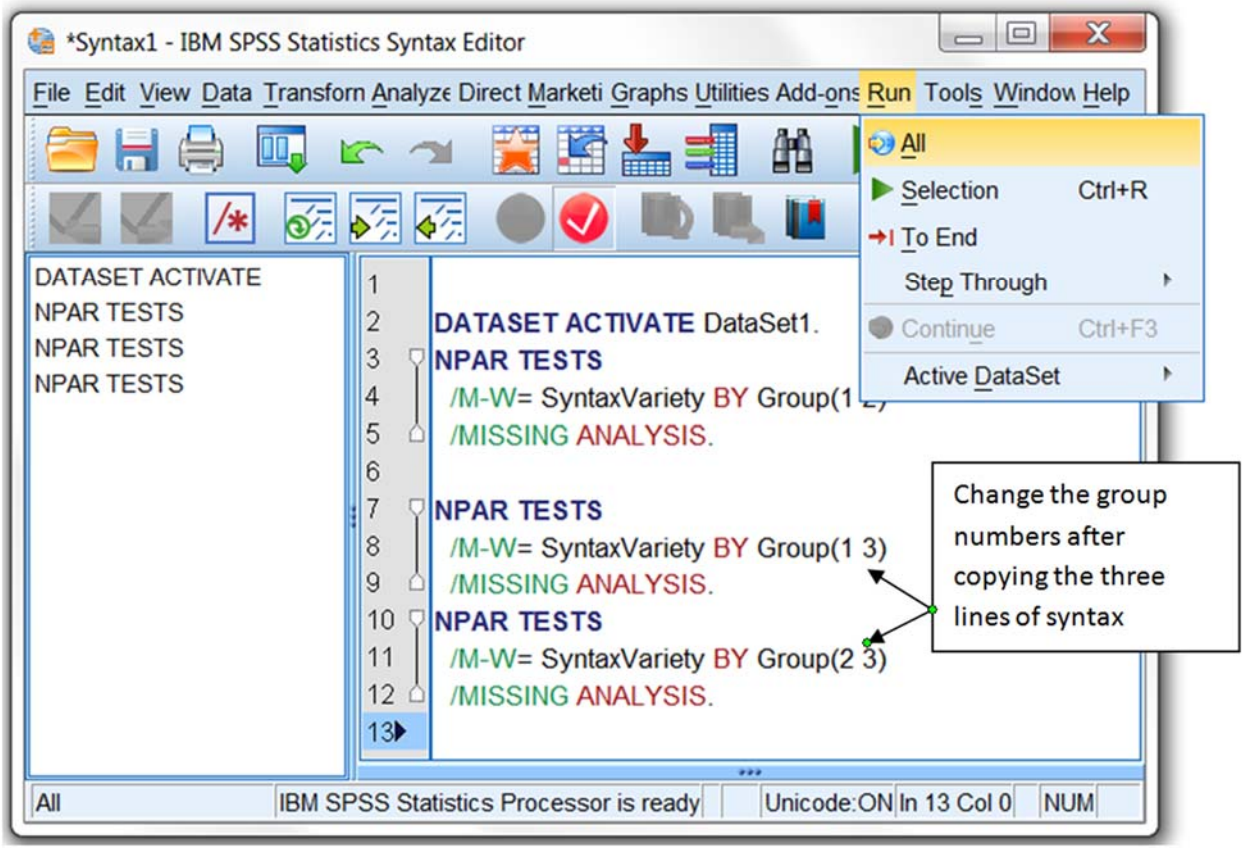

Figure 5 Pasting in syntax to run three Mann-Whitney test as post-hocs for the Kruskal–Wallis test.

The results show that there is a difference between the PTP and NP groups (U = 25.0, p = .001, r = 0.64) and the PTP and OLP groups (U = 47.5, p = .02, r = 0.36), but not between the OLP and NP groups (U = 56.5, p = .054, r = 0.44), although this p-value is quite close to the cut-off point and may be argued to show that all of the groups showed differences from each other, especially as the effect size is even larger than the comparison between the PTP and OLP groups and in general is a fairly large effect size (r effect sizes are calculated as shown in the section called "Effect size for the Mann–Whitney or Wilcoxon Test"). This result is different from the one we received with the parametric test, and might be said to have more power to find differences than the parametric test. We might remember from Sections 9.4.5 and 9.4.6 of the book, however, that

confidence intervals, which we cannot get from SPSS for these tests, showed that the true difference may pass through or be quite close to zero for both the NP vs. OLP and PTP vs. OLP comparisons, so that a *p*-value analysis may be misleading.

Kruskal-Wallis H Test in R (Alternative to the One-Way ANOVA)

In Chapter 9 you imported the EllisYuan.sav file as EllisYuan into R. Use that data (or reimport it) and in R Commander go to STATISTICS > NON-PARAMETRIC TESTS > KRUSKAL-WALLIS TEST. Choose the group variable under "Groups" and the syntaxvariety variable under "Response Variable." Press OK. Here is the R code for this test:

```
kruskal.test(syntaxvariety ~ group, data=EllisYuan
```

```
        Kruskal-Wallis rank sum test

data:  syntaxvariety by group
Kruskal-Wallis chi-squared = 13.6354, df = 2, p-value = 0.001094
```

The output shows there is a statistical effect of group on the Syntax Variety variable, χ^2 =13.6, df = 2, p = .001. The Kruskal–Wallis test is an extension of the Mann–Whitney to the case of more than two levels, so the same type of ranking is taking place in this test, although R does not return any numbers for ranking, the way that SPSS does.

There's just one problem, which is that the Kruskal–Wallis test does not provide post-hoc tests in the same way as the one-way ANOVA did, so we can't be sure which groups are statistically different from one another, and this is probably something we want to figure out. The "Pairwise Multiple Comparison of Mean Ranks Package" (PMCMR) will run multiple comparisons and control the error rate for rank-order statistics. The test we want is called

posthoc.kruskal.nemenyi.test( ) (the name Nemenyi is the researcher who proposed this test). The syntax is a little different from the Kruskal–Wallis test but puts the arguments in the same order, calling them "x" for the DV and "g" for the IV.

```
install.packages("PMCMR")
library(PMCMR)
posthoc.kruskal.nemenyi.test(x=EllisYuan$syntaxvariety, g=EllisYuan$group,
method="Tukey")
```

```
Warning in posthoc.kruskal.nemenyi.test(x = EllisYuan$syntaxvariety, g = EllisYuan$group,
  Ties are present, p-values are not corrected.

        Pairwise comparisons using Tukey and Kramer (Nemenyi) test
                   with Tukey-Dist approximation for independent samples

data:  EllisYuan$syntaxvariety and EllisYuan$group

    NP      PTP
PTP 0.00072 -
OLP 0.23496 0.10251

P value adjustment method: none
```

The warning in the output shows that there were ties in the data, so *p*-values are not adjected to take into account that we have conducted more than one test. We can see from the *p*-values that there is a difference between the NP and PTP groups (p = .0007), but not between the NP and OLP groups (p = .23) or the PTP and OLP groups (p = .10). I do not know of any built-in function to calculate the mean rank of each group (SPSS gives it automatically in the output) to say which group did better than the other.

Effect Sizes for the Kruskal-Wallis H Test (Alternative to the One-Way ANOVA)

The only way I know of to calculate effect sizes is for the two-way comparisons using the Z statistic. This means that to get effect sizes for the Kruskal–Wallis test we need to go back to the calculations for the r effect size given in the section called "Effect Size for the Mann–Whitney or Wilcoxon Test." In the SPSS output I calculated individual Mann-Whitney tests for each of the three comparisons in the Ellis and Yuan (2004) data, and these are the Z-scores and associated N (total observations for the comparison) for those:

NP-PTP, Z=-3.37, N=28

NP-OLP, Z=-1.93, N=28

PTP-OLP, Z= -2.33, N=28

That means the for the NP-PTP comparison, r = .64, for the NP-OLP comparison, r = .36, and for the PTP-OLP comparison r = .44.

For R, you would need to run the wilcox_test( ) command from the coin package to get the Z-score, but this can only be done with two groups at a time. Here's how you could modify your data from the EllisYuan dataset so you can run it for two groups at a time (I'm sure there is a more elegant way to do this, but this is how I figured it out!):

1 Subset the EllisYuan dataframe by selecting one level of the group to exclude. Here I exclude the third group, "OLP," by using the "not equal" syntax, "!=":

```
NP_PTP <- subset(EllisYuan, subset=group!="OLP")
```

2 You might think you're done here but you need to drop one of the group levels, as str( ) will still report 3 levels for the IV:

```
NP_PTP$group <- factor(NP_PTP$group)
levels(NP_PTP$group) #check the levels if you want, but it should work!
```

3 Run the test to obtain the Z-score (remember, this test comes from the coin package):

```
wilcox_test(syntaxvariety~group, data=NP_PTP)
```

4 Use tapply( ) to get counts for your data:

```
tapply(NP_PTP$group, list(group= NP_PTP$group), function(x) sum(!is.na(x)))
```

Repeat the process for the other pairs to get the correct information.

Non-Parametric Alternative to the Paired-Samples T-Test

Use a paired-samples *t*-test when you have two mean scores you want to compare and these scores come from the same group of people. In other words, use a paired-samples *t*-test when your independent variable has only two levels and those levels are repeated measures. I illustrated the use of the paired-samples *t*-test with data from Kim (2013) (Kim2013.sav in SPSS, imported as kim2013 in R), which asked whether participants improved on their scores in understanding sarcasm from a pretest to a posttest when they were taught for 9 weeks about how to identify sarcasm. The parametric paired-samples *t*-test showed that there was a statistical difference between the pretest and immediate posttest, with a CI of [-14.4, -3.8]. Let's see what happens when we use the Wilcoxon matched-pairs signed-ranks test to compare the two times the tests were taken. The null hypothesis for this test is that the two variables being compared came from identical populations (so there is no difference between the groups), and more specifically, "it tests the null hypothesis that the distribution of difference scores (in the

population) is symmetric about zero" (Howell, 2010, p. 678).

Wilcoxon Matched-Pairs Signed Ranks Test in SPSS (Alternative to the Paired-samples T-test)

To call for the test, go to ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > 2 RELATED SAMPLES. Move the variables PRETEST and POSTTEST to the right to form Pair 1, as shown in Figure 6. Leave the box ticked for test type as "Wilcoxon," which will return the **Wilcoxon signed ranks test**. Use the **McNemar test** when you have nominal data and want to see how many people changed their categories over time and you only have two categories. If you want to calculate the exact *p*-value or use a Monte-Carlo approximation to the exact *p*-value, open the EXACT button (see the explanation in the section called "Why Use Non-Parametric Statistics?" about the different options here). Unlike the case with the non-parametric tests we've seen up until now, the descriptive statistics will be useful, so also open the OPTIONS button and tick "Descriptive statistics."

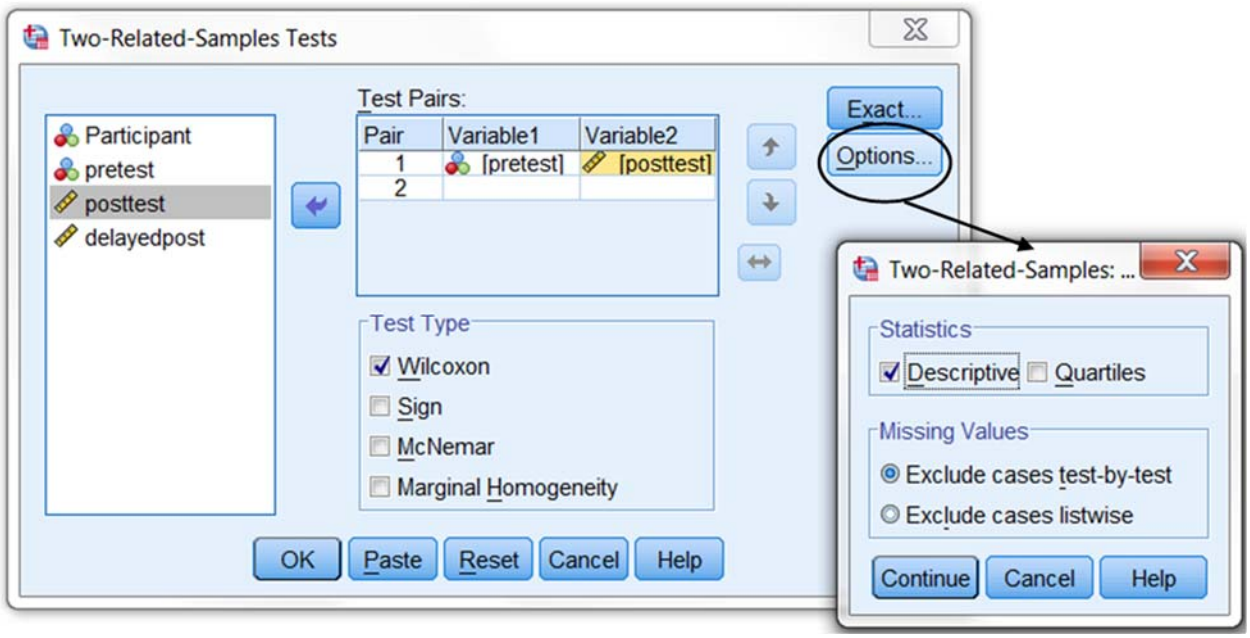

Figure 6 The Wilcoxon Signed Ranks test in SPSS (alternative to the paired-samples *t*-test).

Just like the Mann–Whitney and Kruskal–Wallis tests, the Wilcoxon signed ranks test ranks data. This is why we see mean ranks for the two groups in the output shown in Table 4, after the descriptive statistics give the mean score, standard deviation and counts. Positive ranks mean that an individual scored more highly at Time 2; negative ranks mean they scored lower at Time 2. For those instances where an individual's score did not change, these ties are dropped out of the analysis. From Table 4 you see that there's quite a lot of difference between ranks for the pretest and posttest, with almost all of the participants gaining in scores rather than losing.

Table 4 Results from the Wilcoxon Signed Ranks Test (alternative to the paired-samples *t*-test).

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
pretest	9	7.44	5.247	1	18
posttest	9	16.56	2.963	13	20

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
posttest - pretest	Negative Ranks	1[a]	2.00	2.00
	Positive Ranks	8[b]	5.38	43.00
	Ties	0[c]		
	Total	9		

a. posttest < pretest

b. posttest > pretest

c. posttest = pretest

Test Statistics[a]

	posttest - pretest
Z	-2.433[b]
Asymp. Sig. (2-tailed)	.015

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

The Z-score calculated as a statistic for this test has a probability of $p = .02$, so we conclude that the groups are different.

Wilcoxon Signed Ranks Test in R (Alternative to the Paired-samples T-test)

In Chapter 9 you imported the Kim2013.sav file as kim2013 into R. Use that data (or reimport it) and in R Commander go to STATISTICS > NON-PARAMETRIC TESTS > PAIRED-SAMPLESWILCOXON TEST. Choose the pretest variable under "First variable" and posttest under "Second Variable." If you want to calculate the exact *p*-value or test a one-sided hypothesis, open the Options tab (see the explanation in this document called "The Wilcoxon U test in R (Two independent samples)" about the different options here). The R code for this test is:

```
wilcox.test(kim2013$pretest, kim2013$posttest, alternative='two.sided', paired=TRUE))
```

but I prefer the test from the coin package as it will return a Z-score:

```
wilcoxsign_test(kim2013$pretest~ kim2013$posttest, zero.method=c("Pratt"))
```

In this test, put the first observation first in the formula (the pretest) and the second observation after the tilde (the posttest). Here I have specified the default method, which is the Pratt method, which differs from the Wilcoxon method in how it evaluates ties. Please read the help files for the wilcoxsign_test( ) command if you would like to know more about these methods. The other arguments for this test are the same as they are for the Wilcoxon W test (the analog of the independent samples *t*-test), so see the earlier section entitled "The Wilcoxon U test in R (Two Independent Samples)" for more detail about different choices such as one-sided hypotheses or exact *p*-values. Here is the output from the wilcoxsign_test( ) command:

```
        Asymptotic Wilcoxon-Signed-Rank Test

data:  y by
         x (neg, pos)
         stratified by block
Z = -2.4329, p-value = 0.01498
alternative hypothesis: true mu is not equal to 0
```

The *p*-value is *p* = .01, so we reject the null hypothesis and assume there is a difference between groups. This is the same as was found for the parametric test.

Effect Size for the Wilcoxon Signed Ranks Test (Alternative to the Paired-samples T-Test)

The effect size for the tests can be calculated the same way as stated in the section "Effect size for the Mann–Whitney or Wilcoxon Test" where N will be the total number of negative and positive ranks, with the ties dropped out. This value is given explicitly in the SPSS output, but I don't know of any way to calculate it for R, and so I would just use the number of rows (which is probably the number of participants) used in the calculation. For the sarcasm pretest-posttest, *r* = 2.43/sqrt(9) = 0.81, which is quite a large effect size, accounting for 64% of the variance.

Non-Parametric Alternative to the One-Way RM ANOVA Test

Use a one-way RM ANOVA test when you have tested the same people more than once. You would need to use a one-way RM ANOVA when the one independent variable that you have has more than two levels (if you only had two levels that were repeated measures you could use a paired-samples *t*-test). The parametric RM ANOVA can be used with any number of independent variables, but for the non-parametric alternative you can only use the **Friedman** test when there is just one independent variable. The null hypothesis tested in a Friedman test is that scores for different levels of the independent variable were "drawn from identical populations"

(Howell, 2010, p. 685). We saw an RM ANOVA illustrated through the data of Murphy (2004) and Lyster (2004). However, both of these ANOVAs used more than one independent variable, so we would not be able to apply a non-parametric test to either design as a whole. However, to illustrate the non-parametric test we could ask if there were differences between verb similarity for the Murphy (2004) data for just regular verbs for the NNS adults. This would give us just one independent variable of verb similarity, with three levels (prototypical, intermediate, distant).

In the parametric RM ANOVA that was conducted in Chapter 11 of the book there was a statistical difference for the simple main effect of similarity, but we didn't care too much about that since there was a statistical three-way interaction between verb type, similarity, and group. In the analysis in Chapter 11 I didn't specifically test whether the levels of similarity were statistically different for the regular verbs for the NNS adults. We can look at a means plot with the data (see Figure 7) and see that, as far as the mean score goes, there is not that much difference in the scores on similarity for the NNS adults for the regular verbs (only about .3 points of difference on a 1–5 scale). We may suspect that the comparison would not have shown any effect, but let's see what the non-parametric tests do.

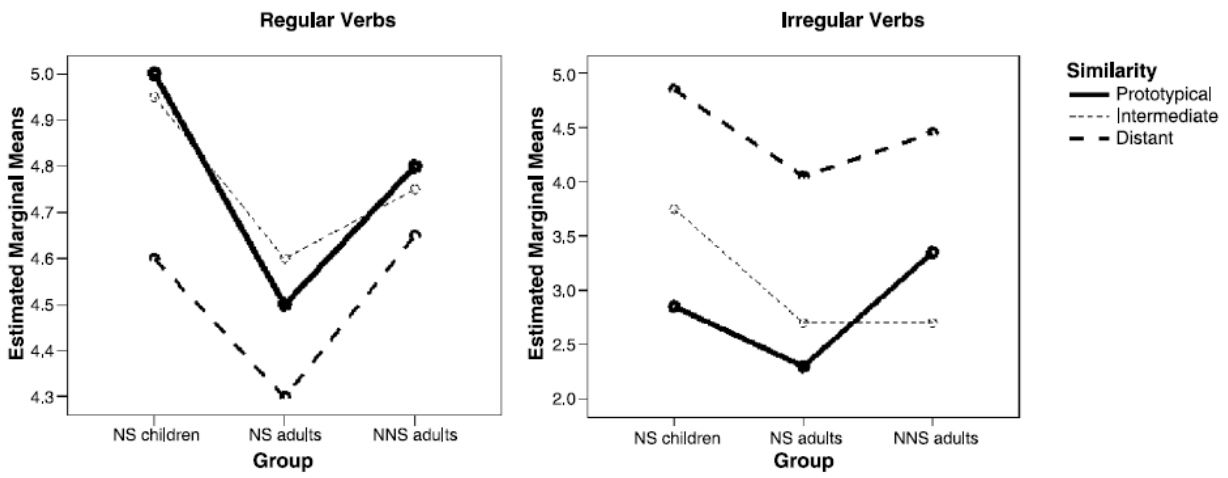

Figure 7 Means plots for the Murphy (2004) data.

The data for a Friedman's test will need to be arranged in the "wide" format, the same as was necessary to run the RM ANOVA. In other words, there needs to be one column with the data for the regular verbs that are prototypical, one column for regular verbs that are intermediate, and one column for regular verbs that are distant if I want to test those three levels of similarity.

Friedman Test in SPSS (Alternative to the One-way RM ANOVA)

Use the SPSS dataset Murphy.RepeatedMeasures.sav. In order to test just the NNS adults I need to select only specific cases of the Murphy (2004) dataset. I go to DATA > SELECT CASES, hit the "If condition is satisfied" radio button, and push the IF button. I want to select only cases where the group = NNS adults, so I move the group variable to the right. I can't actually remember what number the NNS adults are, but looking back I see they are Group 3. So inside the IF button I say "GROUP = 3" and press CONTINUE (remember, I want to choose who I want to *keep*, not throw away!). I'll leave the "Output" button alone to simply filter out the cases I don't want, and press OK. Checking, there are lines over all cases except those where the group is 3.

Now I can run the non-parametric test.

To call for the Friedman test, go to ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > K RELATED SAMPLES. Move the variables that represent your levels to the box labeled "Test Variables." In my case I am testing the three levels of similarity within the regular verbs, so I move REGPROTO, REGINT, and REGDISTANT to the right as shown in Figure 8. It's worth choosing the descriptive statistics for this box, so open the STATISTICS button and tick "Descriptive." Leave the test type box at "Friedman." Kendall's W is used for looking at the agreement between raters, and in that case each separate variable would be one judge's ratings for all of the people they rated. Cochran's Q is used when your data are dichotomous, and in that sense is like an extension to any number of levels of the McNemar test (for more information about these tests, open the HELP button when you are looking at the dialogue box that is shown in Figure 8).

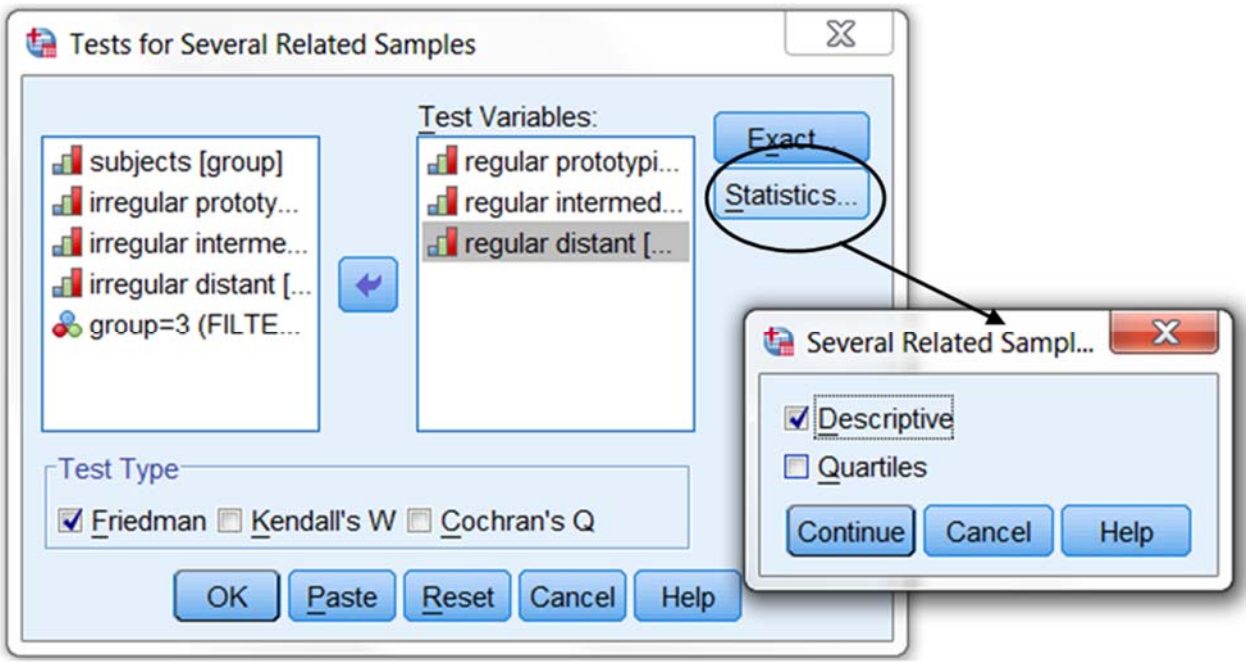

Figure 8 The Friedman test in SPSS.

The mean scores in the descriptive statistics (not shown) are not very different from one another. They are Regular Prototypical, X = 4.8 (sd = 0.5), Regular Intermediate, X = 4.75 (sd = 0.4), and Regular Distant, X = 4.65 (sd = 0.5). The mean ranks are not very different either, as shown in the output in Table 5. A chi-square statistic is returned, and the associated probability of this chi-square value given the degrees of freedom is p = .42. We cannot reject the hypothesis that there is no difference between verb similarities for regular verbs among the NNS adults.

Ranks

	Mean Rank
regular prototypical	2.10
regular intermediate	2.02
regular distant	1.88

Test Statistics[a]

N	20.000
Chi-Square	1.750
df	2.000
Asymp. Sig.	.417

a. Friedman Test

Table 5 Results from the Friedman test (alternative to the one-way RM ANOVA test).

The result is $\chi 2$ = 1.75, df = 2, p = .42, so there is no effect for differences between the different verb similarities.

If we had found a statistical difference, we would be left in the situation of not having post-hocs

to ascertain where the difference lay, and we would need to go back to the test with only two levels, in this case the Wilcoxon signed ranks test, and test only two levels at a time.

Friedman Test in R (Alternative to the One-way RM ANOVA)

In Chapter 11 you imported the Murphy.RepeatedMeasures.sav file as murphy.wide into R. Use that data (or reimport it) and we'll subset the data so that only NNS adults are in it:

```
NNS <-subset(murphy.wide, subset=group=="NNS adults")
NNS$group <- factor(NNS$group)
```

Now you are ready to work with the data. In R Commander with NNS as the active dataset, go to STATISTICS > NON-PARAMETRIC TESTS > FRIEDMAN RANK-SUM TEST. Choose the reg_distant, reg_int, and reg_proto variables and press OK. Here is the R code for this test:

```
.Responses <- na.omit(with(NNS, cbind(reg_distant, reg_int, reg_proto))) #removes NAs and puts the three variables into one object named .Responses
apply(.Responses, 2, median) #asks for medians for columns (the 2 represents columns)
friedman.test(.Responses)
remove(.Responses)
```

```
        Friedman rank sum test

data:  .Responses
Friedman chi-squared = 1.75, df = 2, p-value = 0.4169
```

The result is $\chi 2 = 1.75$, df = 2, $p = .42$, so there is no effect for differences between the different

verb similarities. If we had found a statistical difference, we would be left in the situation of not having post-hocs to ascertain where the difference lay. In this case there is a convenient function that needs the data in exactly the same arrangement as the friedman.test( ) command did, so we can go back and input the lines that R Commander created, all except the last one that removes the data in the correct form (remove(.Responses)), and then use that object in the command that will test multiple comparisons of the groups:

```
friedmanmc(as.matrix(.Responses))
```

```
Multiple comparisons between groups after Friedman test
p.value: 0.05
Comparisons
    obs.dif critical.dif difference
1-2     3.0     15.14086      FALSE
1-3     4.5     15.14086      FALSE
2-3     1.5     15.14086      FALSE
```

```
remove(.Responses)
```

The way to interpret this output is that the difference between Group 1 (Regular Distant) and Group 2 (Regular Intermediate) is 3.0 points, but it would need to be at least as big as 15.14 to be statistical, so this means we can't reject the null hypothesis that there is no difference between groups. The "difference" column says FALSE, meaning the *p*-value is above .05. Of course, we didn't find a main effect for differences between verb similarities, so it does not surprise us that we do not find any differences between paired groups.

Effect Size for the Friedman Test (Alternative to the One-sample RM ANOVA)

I wouldn't recommend giving an effect size for the Friedman test, as it is an omnibus test, and I also don't know how to do one! So I recommend you give an effect size for the individual

pairings of groups, and this can be done using the Wilcoxon Signed Ranks Test, obtaining Z-scores and N, and then using the equation given in the section called “Effect Size for the Mann–Whitney or Wilcoxon Test” that involves the z-score statistic and the total N that has been used throughout this paper.

Summary

In this document I have shown how to perform the non-parametric tests that are available in SPSS and R as counterparts to some of the parametric tests demonstrated in this book. These tests are the classic rank-based parametric tests and are well-known and accepted. The reason to use such tests is if your data do not fulfill the requirements of parametric tests, then you lose power to find differences, and since the non-parametric tests avoid the assumption that the data are normally distributed, you can gain power to find differences in some cases. However, through the entire book I have provided, in each chapter, other and I think better ways to handle data that does not satisfy parametric assumptions. However, because these non-parametric tests are well-known, I provided information about how to use them in SPSS and R.

Bibliography

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, *26*, 59–84.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury/Thomson Learning.

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Pacific Grove, CA: Duxbury/Thomson Learning.

Kim, J. (2013). *Developing conceptual understanding of sarcasm in a second language through concept-based instruction* (Unpublished doctoral dissertation). The Pennsylvania State University, State College, PA.

Kruskal, W. H. (1978). Significance, Tests of. In W. H. Kruskal & J. M. Tanur (Eds.), *International encyclopedia of statistics* (vol. 2, pp. 944–958). New York: Free Press.

Larson-Hall, J., & Herrington, R. (2009). Examining the difference that robust statistics can make to studies in language acquisition. *Applied Linguistics*.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud. *Studies in Second Language Acquisition*, *26*(1), 35–57.

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition*, *26*(4), 399–432.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: LEA.

Murphy, V. A. (2004). Dissociable systems in second language inflectional morphology. *Studies in Second Language Acquisition*, *26*(3), 433–459.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Wilcox, R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*(3), 300–314.