Factoring out Differences with Analysis of Covariance: The Effect of Instruction on Derivational Morphology

> A covariate is, after all, nothing but an independent variable which, because of the logic dictated by the . . . issues of the research, assumes priority among the set of independent variables as a basis for accounting for . . . variance.
>
> Jacob Cohen (1968, p. 439)

Analysis of covariance (ANCOVA) is a statistical technique you can use when you want to focus on the effects of a main response variable with the effects of other interval-level variables factored out. Such a technique may be useful when:

- you assume that there is some external factor, such as pretest or TESOL score, which will affect how your students will perform on the response variable
- previous studies have shown that another variable, such as aptitude or writing scores, affects how your participants will perform on the variable of interest
- you find after the fact that an unplanned variable, such as age, affected the performance of participants on the response variable

In essence, ANCOVA works by simply including the additional variable (the **covariate**) in the regression, but, by doing so, it allows the effects of that variable (such as age, or aptitude scores) to be separated out from the response variable. ANCOVA is like partial correlation (information

for which can be found in the online document in Chapter 6 called “Other Kinds of Correlation”), because it includes the variable whose effects we want to “partial out” in the analysis in order to separate them from the other effects. ANCOVA works like the repeated-measures designs seen in Chapter 11 as well to reduce the amount of variability in the model that is unexplained. If we think that scores on an aptitude test help account for the variability on the response variable, then by including the aptitude test in the design we help reduce the amount of variability that is unexplained.

The ANCOVA design, then, is quite similar to the ANOVA design but includes one or more variables as explanatory variables. The example we will look at in this section involves a study by Lyster, Quiroga and Ballinger (2013), which looked at the effect of direct instruction on derivational morphology among second grade students in bilingual classrooms. The students differed in that they came from different schools and were taught by different teachers, had different language profiles (English-dominant, French-dominant or bilingual, identified by using the Peabody Picture Vocabulary Test), and either received direct biliteracy instruction on morphology or did not (the comparison group). Lyster, Quiroga and Ballinger (2013) had a number of measures that they could (and did) use as covariates in their analysis, including a test of phonological awareness and a pretest measure of morphological awareness. The response variable was a morphological awareness test with a maximum score of 146 (both English and French were tested, although each ANCOVA only examined one at a time).

Recent Examples of ANCOVA in the SLA Literature

Lee & Macaro (2013)

The authors wanted to investigate the question of whether using a shared L1 in the foreign language classroom would help or hinder vocabulary acquisition. They tested large numbers of Korean sixth graders (N=443) and Korean college freshmen (N=286) after each of 4 sessions where certain words in a reading were targeted and explained verbally, either in Korean or English, in the context of trying to understand the reading. The teachers used the same reading materials but were divided into whether they were native speakers of Korean, and thus used the shared L1 in defining vocabulary words, or native speakers of English who had low Korean skills and used only English in the classroom. A one-way ANCOVA on the posttest was conducted with Instructional type ("English only" or "Codeswitching") as the between-group variable and a vocabulary pretest score as the covariate, with the data split for the younger learners and the older learners. On a test of vocabulary recall, where learners had to actually produce a word, the younger learners in the Codeswitching group scored statistically higher than the English only group even when pretest vocabulary scores were taken into account, with an effect size of partial eta-squared = .23, while for older learners the ANCOVA also showed a statistical difference between the groups but at a lower measure of effect size, partial eta-squared = .14.

Van Beuningen, De Jong & Kuiken (2012)

This study assessed the value of comprehensive error correction. A large number (n=268) of secondary school children (mean age 14) learning Dutch as an L2 and with a variety of L1s (the largest being Moroccan Arabic at 31%) were tested. One group received direct corrective feedback, another received indirect corrective feedback with an error coding system, a third group were invited to revise their writing but did not receive any feedback (self-correction group), and a fourth group were not asked to revise but instead to write another essay (additional writing practice). In a pretest, all groups filled out a questionnaire, did a receptive vocabulary test as a measure of overall language proficiency, and had to write for 20 minutes on the topic of butterflies as their first writing task. In the second week, participants got their treatment on the initial writing task. The first three groups were invited to revise their papers after their treatment, while the fourth group was asked to write about honeybees. A posttest was given one week later and the students had to write another essay about ladybugs. Texts were coded for linguistic errors and clause types. The results were measured in an error to number of words ratio, and errors were divided into linguistic (word order, article error, additions of nonnecessary elements) and non-linguistic (lexical errors, orthographic errors, pragmatic errors). Structural complexity was measured as the number of subordinate clauses as a

percentage of the total number of clauses. Because the data came from different intact classes and schools, the statistical analysis started with a multilevel analysis (what is called a mixed-effects analysis in this chapter), but this analysis found that class and school did not make a difference in explaining variation, so the authors chose to use an ANCOVA design using language proficiency (score on the vocabulary test) and pretest performance as covariates to factor out the influence of individual differences of proficiency. The authors used the statistical approach explained in the online document in Chapter 7, "Finding the Best Fit in Multiple Regression", of starting with a maximal model and reducing the model to the minimal model. Many statistical tests were conducted, but I will report only on the ANCOVA that looked at the number of errors made at the initial posttest. This model found a statistical effect for group ($F_{3,231}$=11.34, p<.001, partial eta-squared=.13), and both the covariate of language proficiency ($F_{1,231}$=12.93, p<.001, partial eta-squared=.05) and pretest accuracy ($F_{1,231}$=87.9, p<.001, partial eta-squared=.28) were statistical. The effect of pretest accuracy explained the largest amount of the variance in the data, as can be seen from the very high effect size (partial eta-squared). Post-hoc comparisons on group found that those who had received direct or indirect correction used fewer errors in their subsequent writing than those who self-corrected or just practiced more writing, although there was not a statistical difference between either of the two correction conditions.

Miranda Casas, Soriano Ferrer & Baixauli Fortea (2013)

The authors begin by noting that in children with a clinical diagnosis of attention-deficit/hyperactivity disorder (ADHD), learning disabilities in writing are twice as common as problems in other academic areas such as math, reading or spelling. In order to investigate the effect that ADHD may have on writing ability, the authors compared the writing performance of children with ADHD (n=50, age range 9–14 years) with those without ADHD (n=50, same age range) on a written narrative task about a trip they had recently taken. The groups were unbalanced in the sense that of the 50 ADHD children, 49 were boys and 1 was a girl; the makeup of the non-ADHD group was also more weighted toward boys (36) than girls (14). For this reason, in the analysis, gender was used as a covariate in order to remove any possible effects of gender. In order to evaluate the writing, the authors used quite a host of measures, including measures of structure (setting information, conclusion), time sequence, content digressions, cohesion, number of words, mean length of utterance, syntactic complexity, type-token ratio, etc. ANCOVA analyses were performed on each of these measures. One of the measures that showed a large effect size was on text structure ($F_{1,97}$=41.5, p<.000, Cohen's d=1.3), with the ADHD children having a much harder time "articulating an organizational plan directed toward a purpose" (p. 453). Another measure that showed a large effect size was the number of words ($F_{1,97}$=56.2, p<.000, Cohen's d=1.5), with the ADHD children

producing on average 73 words while the comparison group produced 115 words.

Peters, Hulstijn, Sercu & Lutjeharms (2009)

Incidental vocabulary acquisition through reading has been found to be a slow and arduous process. This study asks whether readers will learn more vocabulary if one or more of three techniques is used in a classroom setting: 1) announcing that there is a vocabulary test after an in-class reading; 2) giving reading comprehension questions that crucially rely on unfamiliar words in the text to correctly complete, and 3) giving a vocabulary test after the reading. Students' vocabulary level was measured with a 50-item multiple choice test and used as a covariate in the analysis, since previous research has shown that students with higher levels of vocabulary are able to retain new words better. The researchers measured students' behavior while reading (whether they looked up the words online) and their word retention in (N=137) Dutch L1 German L2 college students. One of the three word retention tests was a recall test of meaning in context. There were a large number of research hypotheses tested, but one was what effect the three techniques mentioned above would have on how many words students retained on the test. There were 4 different groups tested, and here are their mean scores on the recall test at an immediate posttest and a delayed posttest 2 weeks later:

Experimental Group	Immediate	Delayed
INCID ONLY (- test announcement, -comprehension Qs)	8.12 (1.77)	6.04 (1.60)
INCID PLUS (- test announcement, +comprehension Qs)	11.77 (2.18)	8.27 (1.95)
INTENT ONLY (+ test announcement, -comprehension Qs)	8.64 (2.36)	6.33 (2.63)
INTENT PLUS (+ test announcement, +comprehension Qs)	12.17 (2.48)	8.72 (2.69)

A 4-way RM ANCOVA with 1 covariate was used to test the research question. It was a 2 ($\pm$Test announcement) × 2 ($\pm$Comprehension Qs) × 2 (Types of words: Targeted or not) × 2 (Testing time: Immediate and Delayed) ANCOVA with vocabulary size as a covariate. Results of this covariate were not reported; it was simply used as a way to look at participants' results while factoring out the effects of their differential vocabulary levels. The first three variables were between-subject variables while Testing time was a within-subjects variable (the repeated measures). The analysis found that the four-way interaction was statistical with a medium effect size ($F_{1,103}$=8.22, p=.005, eta-squared=.06), as was the three-way interaction between the between-subject variables (Test announcement, Comprehension Qs and Types of words) with a small-to-medium effect size ($F_{1,103}$=5.24, p=.02, eta-squared=.03). Obviously, with a four-way analysis there are quite a large number of results and I won't list all of them here, but in prose, the authors found that participants who did comprehension questions performed better than those who did not do it, and that word retention was affected by whether words were

targeted or not. Scores were lower on the delayed posttest than immediate posttest, but not for all students—those who had done the comprehension questions lost more vocabulary from immediate to delayed posttest than those who had not, but they remembered more of the targeted words than the non-targeted words.

Visually and Numerically Examining the Data

Numerically Examining the Data

Let's start by looking at Lyster, Quiroga and Ballinger's data numerically. In their 2013 paper, Table 3 listed descriptive statistics for the English and French versions of the morphological awareness test, both pretest and posttest, also divided into scores for students in the experimental class and the comparison group and then within each of those groups, into language dominance groups. Let's recreate that table. Use the Excel data file called LysterQuirogaBallinger2013.xlsx (this will give us practice getting an Excel file into both SPSS and R).

In SPSS, go to FILE > OPEN > DATA and change the type of file to Excel. Press the OPEN button and a box labeled "Opening Excel Data Source" appears. The Excel file has variable names along the top, so make sure the box that says "Read variable names from the first row of data" is checked and the first page of the worksheet is listed (for this data file, it's labeled in Spanish and says "Hoja 1[A1:AP80]"). Press OK and the file is loaded without any problems.

R Commander used to have a command for importing Excel files, but that is now gone, so one way to do this is to save the Excel file as a .csv file and then import through R Commander. Open up the Excel file and save the file as a .csv file. Now in R Commander, go to DATA > IMPORT DATA > FROM TEXT FILE, CLIPBOARD OR URL . . . Enter the name LQB and choose "Commas" as the "Field Separator." Press OK and navigate to where you have saved your file. You might want to pull down the menu that says "All Files" and change to just text files and .csv

files in order to narrow down the files you are searching. Once you find the file and press "Open" you should find that your file opened up correctly in R.

To get a table like the one we see in Table 1, in SPSS you might first want to label the CONDITION and LANGUAGE variables since it would be nice to have those defined when we call for data. Go to the VARIABLE VIEW tab and click on the cell for Language under the Values column. A blue button will appear. Press it, and define 1=English dominant, 2=French dominant, 3=Bilingual. For Condition, 1=Comparison, 2=Experimental. Now go to DATA > SPLIT file, and put Condition and Language into the box with the button “Compare groups” pushed. Then go to ANALYZE > DESCRIPTIVE STATISTICS > DESCRIPTIVES and move over all of the MAT files (4 of them) to the “Variable(s)” box. Open the Options button and tick off everything except “Mean” and “Standard deviation.” This will give a fairly compact table, although not arranged as nicely as Table 1.

Table 1 Descriptive statistics for Lyster, Quiroga & Ballinger 2013.

	English Version of Morphological Test		French Version of Morphological Test	
	Pretest	Posttest	Pretest	Posttest
	M (sd)	M (sd)	M (sd)	M (sd)
Experimental				
English dominant (n=9)	74.78 (15.77)	95.44 (15.43)	69.33 (18.16)	80.78 (10.64)
French dominant (n=16)	59.13 (15.46)	72.94 (25.48)	84.25 (14.41)	93.62 (13.60)
Bilingual (n=20)	71.10 (19.03)	80.45 (18.53)	81.95 (20.05)	95.25 (17.76)
Comparison				
English dominant (n=8)	69.13 (15.97)	76.38 (15.20)	53.63 (21.33)	60.25 (16.87)
French dominant (n=5)	42.40 (13.05)	58.60 (17.39)	82.20 (14.72)	80.80 (18.21)
Bilingual (n=7)	62.00 (14.35)	85.14 (19.08)	84.14 (16.55)	88.14 (16.65)

For R, first look at the structure of the imported data (some variables at the end of dataframe cut):

```
> str(LQB)
'data.frame':   65 obs. of  13 variables:
 $ School     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ ID         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Program    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Language   : int  1 2 3 1 1 1 3 1 3 1 ...
 $ Conditon   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ PA.pre.FR  : int  29 30 35 17 36 37 24 23 23 36 ..
 $ PA.post.FR : int  26 22 24 15 35 28 26 21 30 31 ..
```

The thing to notice is that our categorical variables are not factors yet. We would like Language and Condition to be factors, with the same category level labels attached that were listed above

in the SPSS paragraph. Turning a number into a factor is as easy at telling R to do so, and then we can list the level labels for that variable like this:

```
LQB$Language<-as.factor(LQB$Language)
levels(LQB$Language)=c("English dominant", "French dominant", "Bilingual")
LQB$Condition<-as.factor(LQB$Condition)
```

```
Error in `$<-.data.frame`(`*tmp*`, "Condition", value = integer(0)) :
  replacement has 0 rows, data has 65
```

Why do I get an error message here? I check the names of the dataset again:

```
> names(LQB)
 [1] "School"      "ID"          "Program"     "Language"    "Conditon"
 [6] "PA.pre.FR"   "PA.post.FR"  "PA.pre.EN"   "PA.post.Eng" "MAT.pre.FR"
[11] "MAT.post.FR" "MAT.pre.EN"  "MAT.post.EN"
```

Oh! Looks like Condition was spelled incorrectly in the original file. I'll just change the name while I'm making it into a factor, and remove the old "Conditon" column:

```
LQB$Condition<-as.factor(LQB$Conditon) #notice misspelling on right side but not left
LQB$Conditon=NULL
levels(LQB$Condition)=c("Comparison", "Experimental")
```

That's all I need for now, but you might want to change Program into a factor too, in case we need it later. For Program, 1=80% English, 2=80% French and 3=50/50.

```
LQB$Program<-as.factor(LQB$Program)
levels(LQB$Program)=c("80% English", "80% French", "50/50")
```

Looking at Table 1, we notice that test takers' scores improved from pretest to posttest, although some more dramatically than others (this trend holds for everyone except for the French-dominant children in the French morphological test). Also notice that in almost all cases, in the pretest the English-dominant children do better on the English version of the morphological test and the French-dominant children do better on the French version of the morphological test no matter whether they are in the experimental or comparison group (for the Comparison group for French the bilinguals are actually a bit higher). For the posttest, in the experimental group the English-dominant children do better than everyone else but the French-dominant children are bested a bit by the Bilinguals. In the Comparison group, by contrast, the bilinguals in both languages score the highest.

As for checking on data assumptions, we notice that the standard deviations of each group are not too different from one another and probably we are safe to assume homogeneity of variance for the data, at least divided up this way.

Visually Exploring the Data

This dataset is quite complex and we may wonder how to begin looking at it. We have at least four different factors we'd like to compare—scores on the French morphological awareness test (MAT) versus scores on the English MAT, how scores changed from pretest to posttest (testing time), membership in the experimental group or not (Condition), and which language is dominant (Language). We may also want to explore at some point how Phonological Awareness

in both French and English, at a pretest and posttest stage, may have affected or explained scores on the MAT. So there's a lot going on! In Chapter 11 we looked at a parallel coordinate plot for complex data, and with this plot we can examine the change on the MAT from pretest to posttest scores with plots split for Language and Condition. We'll just need to look at data from English and French separately. Here is the code I used to create Figure 1 in R:

```
library(lattice)
parallelplot(~LQB[9:10]|LQB$Condition*LQB$Language, data=LQB,
main="Lyster, Quiroga & Ballinger (2013)\nFrench Morphological Awareness Test",
varnames=c("French pretest", "French posttest"))
parallelplot(~LQB[11:12]|LQB$Condition*LQB$Language, data=LQB,
main="Lyster, Quiroga & Ballinger (2013)\nEnglish Morphological Awareness Test",
varnames=c("English pretest", "English posttest"))
```

We saw in Chapter 11 that SPSS can create parallel coordinate plots too. First, split your data according to both Language and Condition (DATA > SPLIT FILE), then go the GRAPHS > GRAPHBOARD TEMPLATE CHOOSER, and choose the variables MAT.pre.FR and MAT.post.FR. The choice of a parallel coordinate plot (PARALLEL) will appear and with the data already divided you should be able to call back all of the graphics you can see in R (although you will have to manually put them together to create the type of graphic you see in Figure 1).

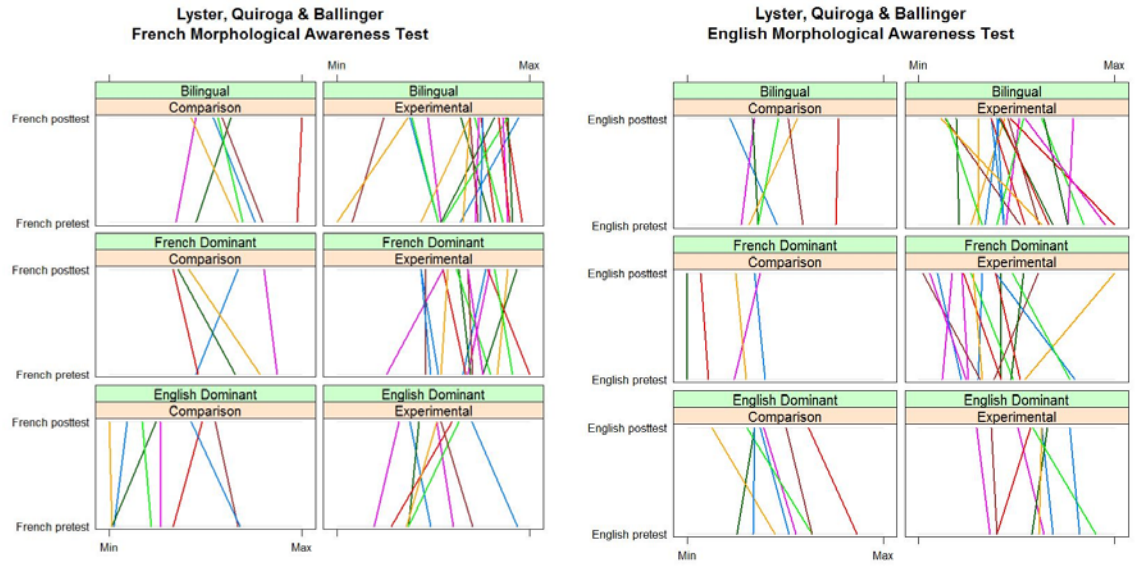

Figure 1 Parallel coordinate plots for Lyster, Quiroga & Ballinger (2013) morphological awareness test.

To see improvement from pretest to posttest we want to see the lines lean to the right, toward higher scores. Visually, does it look like more of the lines in the Experimental groups are leaning as compared to the Comparison groups? Well we see there are more participants in the Experimental group than the Comparison group, so that is something to keep in mind as less typical scores may have more influence, but to me there does appear to be more leaning to the right in the Experimental column than the Comparison column. Notice also how scores across the two language tests (French on the left, English on the right) differ for the Language groups. For example, the lines of the French dominant group are clustered more toward the right end of the window (more toward the max score) for the French MAT as compared to the English MAT.

Another graphic that we used with lots of variables was the coplot in Section 7.2.1 of the book. The coplot shows scatterplots conditioned by other variables, so let's look at scatterplots of the

relationship between pretest and posttest scores, conditioned by Language and Condition (see Figure 2 for the English test). Here is the R code for the coplot:

```
coplot(MAT.post.EN~MAT.pre.EN|Condition*Language, panel= function(x,y,...)
panel.smooth(x,y,span=1.0,...),data=LQB)
```

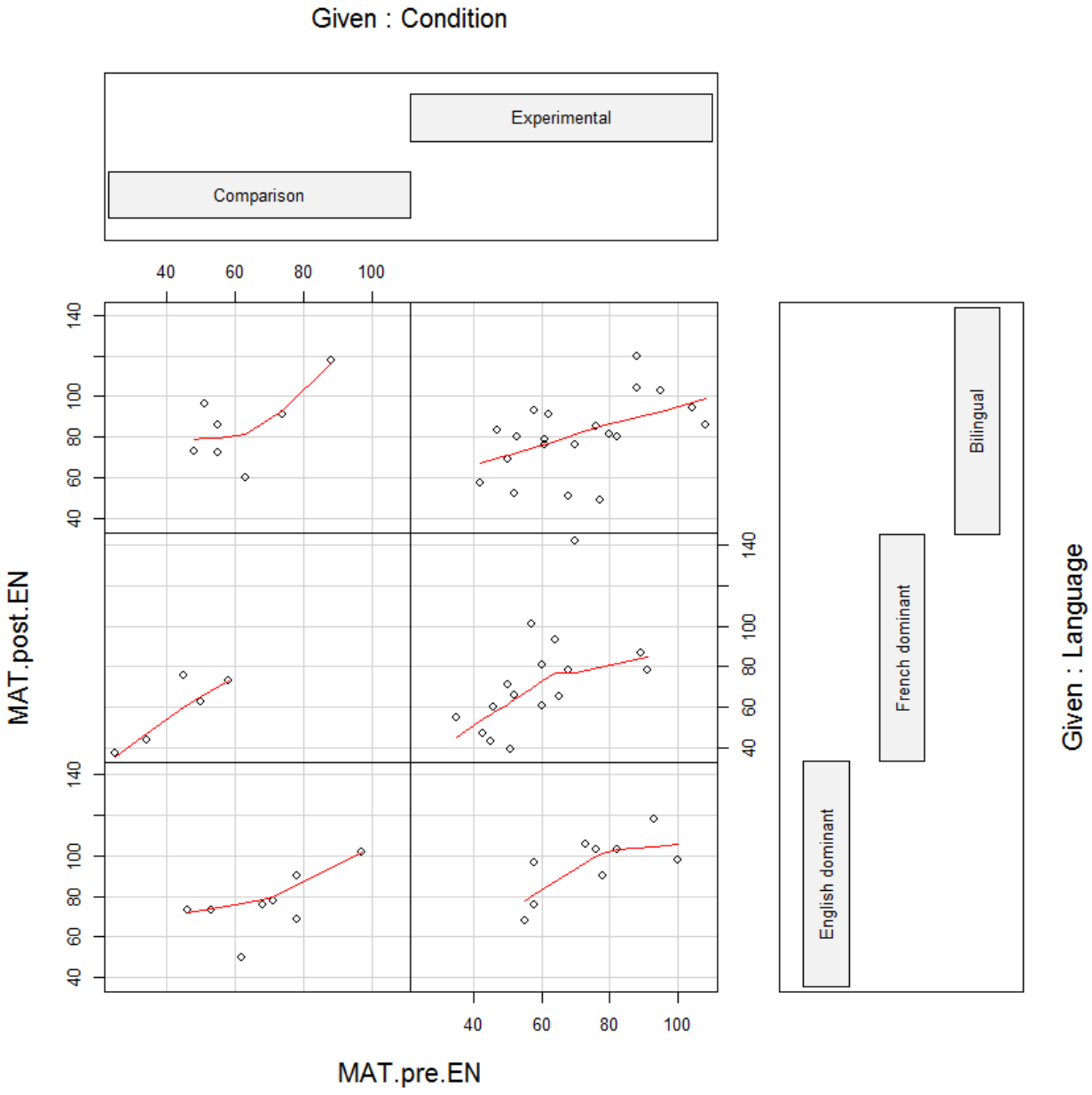

Figure 2 Coplot for Lyster, Quiroga & Ballinger (2013) English morphological awareness test.

Because the conditioning variables are all categorical, this coplot is basically just a collection of the six scatterplots that can be generated by the intersection of the 3 levels of Language and the 2 levels of Condition. Although I don't know of any way to make coplots in SPSS, one could certainly make scatterplots with the same division of data. The coplots have Loess lines on the data, showing not a straight regression line but rather an approximation to the trends in the data. Since the pretest is plotted on the x-axis and the posttest on the y-axis, the steeper the line slopes to the right, the more the gains from pretest to posttest. Again, we should note that the number of data points in some categories is small and we cannot rely on those trends as strongly, but it appears that the French-dominant students had quite steep slopes for both the control and experimental groups on the English MAT, steeper than the Bilinguals (and the English-dominant groups are both quite small so it is hard to draw conclusions from them).

The coplot excels when using continuous variables as conditioning variables, so let's try keeping the Language variable but using the pretest Phonological Awareness variable to condition scores. We'll see how Phonological Awareness affected the pretest to posttest learning on the English MAT (see Figure 3).

Here is the R code for this figure:

```
coplot(MAT.post.EN~MAT.pre.EN| Language*PA.pre.EN, panel= function(x,y,...)
panel.smooth(x,y,span=1.0,...),data=LQB)
```

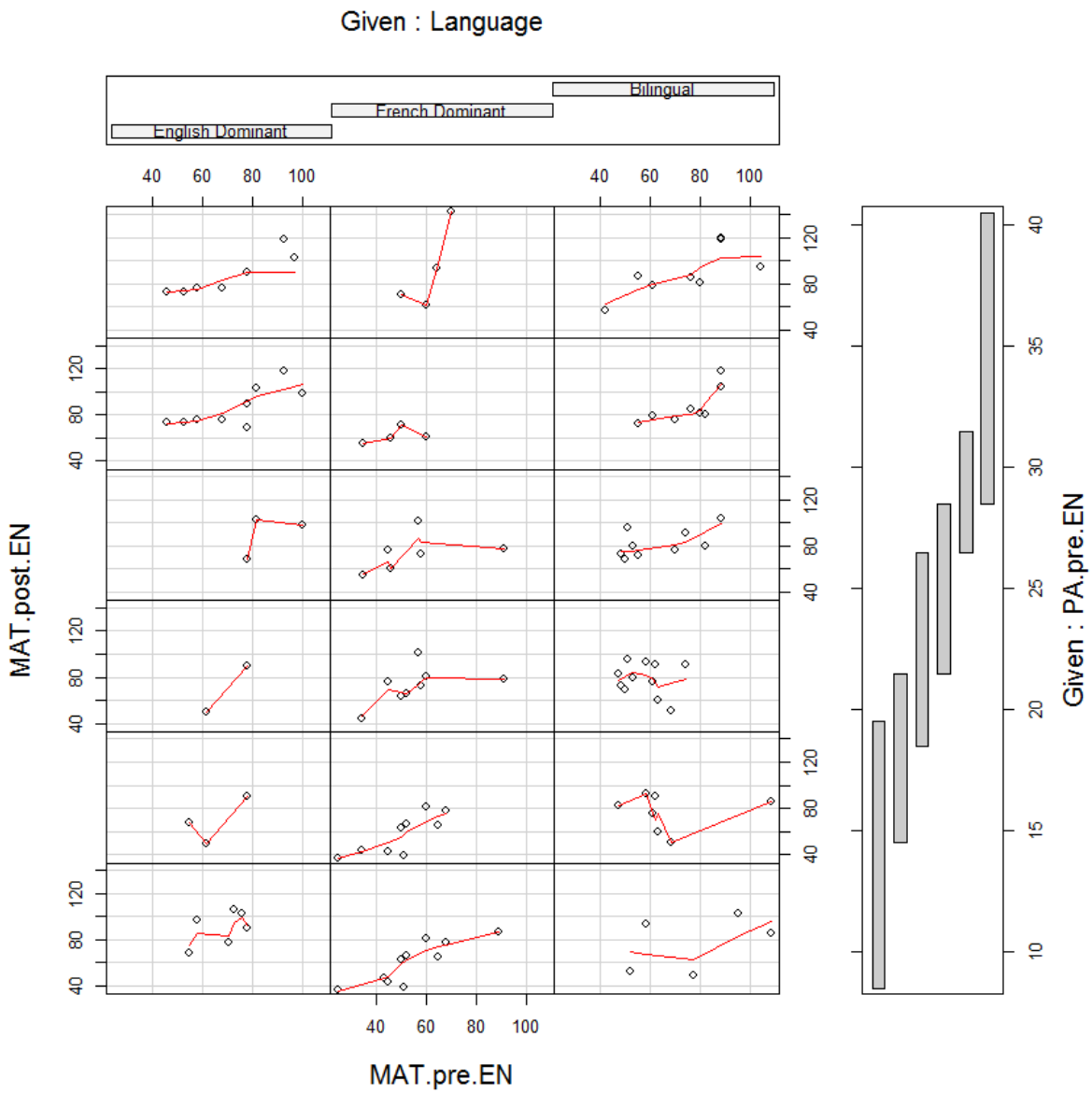

Figure 3 Coplot English morphological awareness test with conditioning variable of Phonological awareness.

Use Figure 3 to look at the differences in growth from low PA to high PA. Do you think slopes get steeper as participants have higher levels of PA to start with? It looks like perhaps having a high phonological awareness may be associated with students being able to make more gains

from pretest to posttest in morphological awareness. Of course, this plot mixes together the Experimental and Comparison students, which may not be the right thing to do. Anyway, you can see how you could use a coplot to explore your data and look for trends. The more you want the line to be a straight regression line and not follow the trends of the data, the higher you can set the span number in the argument for the coplot.

There's one other plot I'd like to make with this data, and it does not involve R or SPSS, but it might be helpful for understanding the data. This is a small multiple, a term coined by Tufte (2001). The small multiple takes the same information and repeats it multiple times. In this way, you can understand the graphic display of information for one individual and then your mind can easily create patterns by looking at that display for many individuals. My idea was to have a barplot with the MAT pretest and posttest information side by side, and put those barplots for the French and English MATs side by side. Then to see how that gain from pretest to posttest was affected by phonological awareness, I wanted to have pretest and posttest information for PA for each language in a barplot version underneath each MAT. These small graphs containing 8 pieces of information could then be arranged according to Language and Condition, so that the viewer could try to spot any trends across the combinations of Language and Condition (the small multiples could also of course be rearranged in other ways, by Program or School). The idea of such a graphic is to see all of the individual information in a visual way, which the brain is able to use to provide perceptual inferences at basically zero cost (Larkin & Simon, 1987). We humans are able to process data this way much more efficiently and to find patterns more readily than looking at the same information in numerical form. At least, that's my hope! Now someone who could program R better than me would probably be able to write a program to create these

barplot clusters (and I hope they will and will send it to me, and I'll post it online!), but since I can't, I've had them drawn by hand (thanks go to my son, Lachlan Hall), scanned them in, and rearranged them according to Condition and Language, and present them to you in Figure 4.

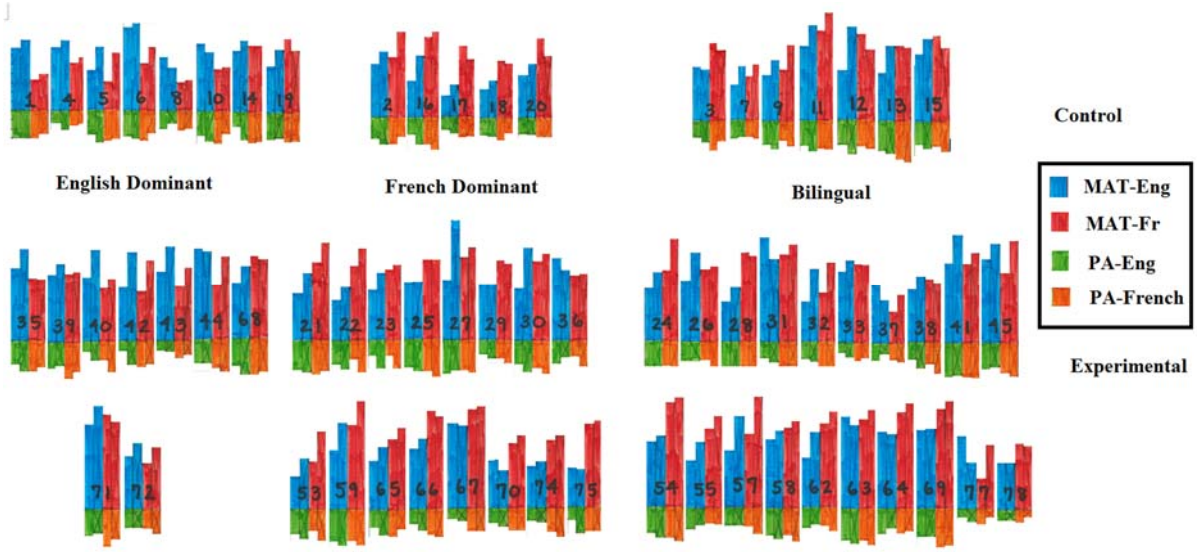

Figure 4 Small multiple boxplot clusters for the Lyster, Quiroga & Ballinger (2013) data.

With those views of the data, then we shall continue in our quest to analyze the data statistically.

Visually Examining the Data for Statistical Assumptions

We would like to examine the data for normality and homogeneity of variances, with the data divided up the way we will be looking at it. The one-way ANCOVA we look at will examine the pretest and posttest morphological awareness tasks (MAT) divided up by Condition (Experimental vs. Comparison), so let's call for boxplots of the data split this way.

In SPSS, here's a different way to make boxplots than I have before. Go to GRAPHS > REGRESSION VARIABLE PLOTS. Move all 4 MAT variables to the box labeled "Vertical-Axis

Variables." Move CONDITION to the box labeled "Horizontal-Axi Variables." Press the OPTIONS box and add a title if you like. Click on "Boxplots under "Categorical Variable Plots." Press CONTINUE. You will get 4 boxplots, one for each MAT variable split into data from the Comparison and Experimental groups.

In R Commander, you can choose GRAPHS > BOXPLOTS and then split one variable at a time by Condition into Experimental and Comparison groups (open the PLOT BY GROUPS button to pick Condition).[1] This can get you the syntax you need to paste into R and get all 4 boxplots you want. Alternatively, use the code I showed in Section 8.2.4 and with a little more work you can clean the boxplots up to all fit into one figure and look much nicer, as in the code here, which results in Figure 5:

```
par(mfrow = c(2, 2)) #Change to 2 rows and 2 columns so it will all fit in one window
boxplot(MAT.pre.EN~Condition, data=LQB,
names=c("Comparison", "Experimental"),main="MAT pretest English",
las=1,notch=FALSE,col="grey", boxwex=.5,medcol="white")
```

. . . continue with the same syntax for 3 more boxplots, just changing the underlined part to different files and fix the name.

[1] Remember that if you don't seem to be able to choose any groups, you may need to pick a different active dataset, then come back to the LQB data file, since we changed the structure of the dataset to turn some variables into factors, but R Commander doesn't know that until we detach and reattach the file.

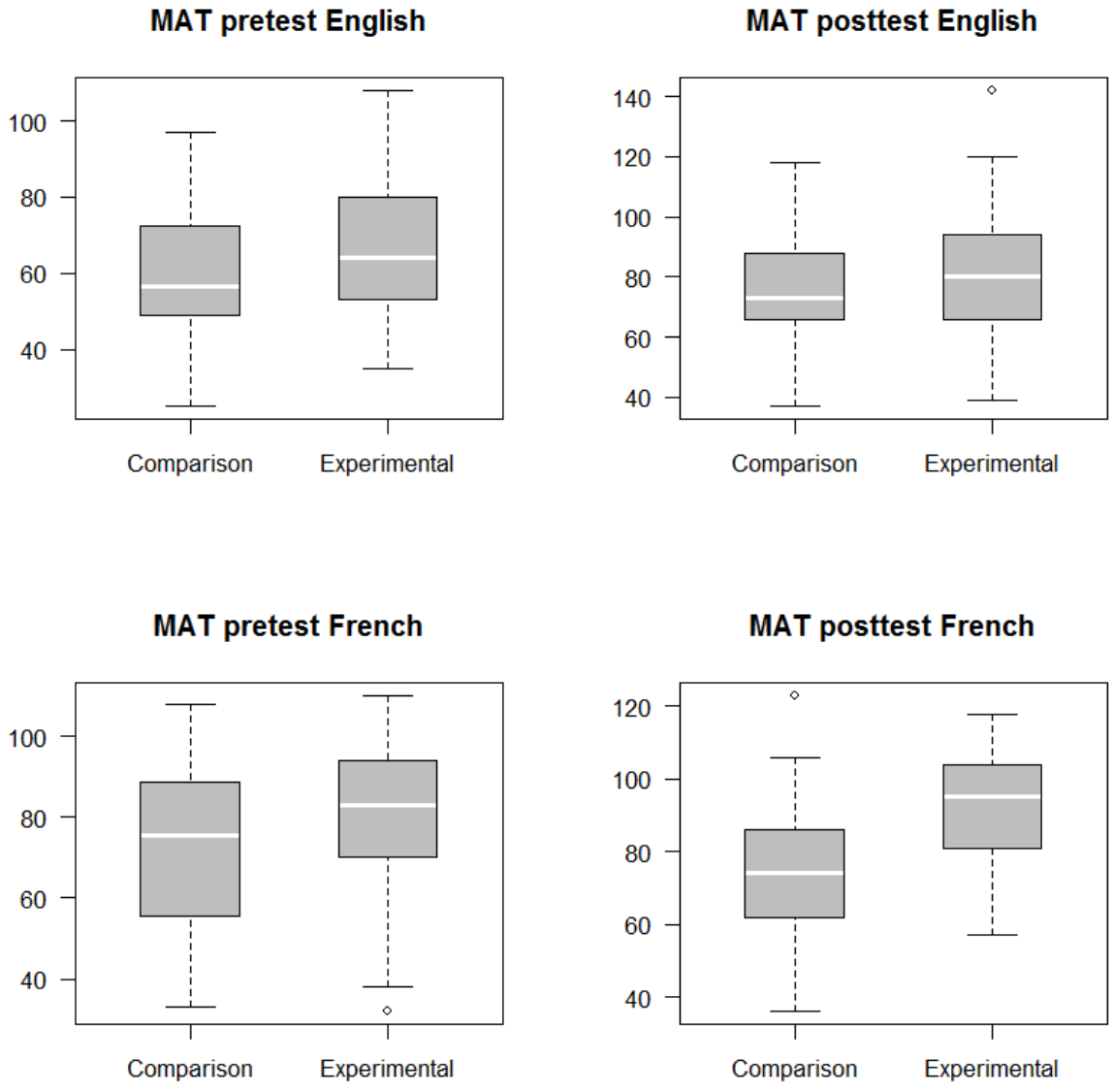

Figure 5 Boxplots of the MAT data split by Condition (Experimental vs. Comparison).

The boxplots in Figure 5 show us several things. First, they show that in general, collapsing all of the language groups, the Experimental group consistently has a higher median score than the Comparison group, but usually not by too much. There are departures from normality in that there is one outlier in all but the pretest English MAT, and some slight skewness in some distributions.

One more graph that can be helpful in assessing what is happening in a situation with covariates is to look at a scatterplot of the continuous variables with data divided by groups. The scatterplot should have regression lines drawn for the groups because the hypothesis that is being tested in a standard ANCOVA is that the regression lines for the groups are parallel. The one-way ANCOVA tested in this chapter looks at the relationship between the posttest MAT and the posttest PA. This ANCOVA will divide the data into the two Conditions, Experimental and Comparison, so this is the type of scatterplot we will look at.

In SPSS, choose GRAPHS > LEGACY DIALOGS > SCATTER/DOT, choose SIMPLE SCATTER and DEFINE, and then put "MAT-post-FR" in the "Y-Axis" box and "PA-post-FR"in the "X-Axis" box (or vice versa), and "Condition" in "Set Markers by." Press OK. To draw regression lines on the scatterplot, double-click on the plot and the CHART EDITOR will open. In the CHART EDITOR, choose ELEMENTS > FIT LINE AT SUBGROUPS to get separate regression lines on the two different groups. A PROPERTIES box will open and if you want to choose a different type of line besides "Linear" (which you don't right now) you can choose that and press APPLY. Otherwise, just press CLOSE. Since SPSS separates lines only by color, you'll probably want to make different groups have different symbols. To get a different symbols, click twice, slowly, on one particular point. A different PROPERTIES box opens to the MARKER tab. You can choose a different symbol under the "Marker": "Type" box. You can change the color too. Press APPLY and CLOSE and then close the CHART EDITOR.

In R Commander, you can choose GRAPHS > SCATTERPLOT and pick "MAT.post.FR" for the x-variable and "PA.post.FR" for the y-variable (or vice versa). We want to split by Condition, so

open the PLOT BY GROUPS button and pick Condition (by the way, in this command you can only pick one variable to split by). Go to the OPTIONS tab and click off “Marginal boxplots,” “Smooth line” and “Show spread” (these are useful things but not what we want right now!), leaving only “Least-squares line” (the straight regression line). Go ahead and change “Identify Points” to “Do not identify.” Press OK. To make Figure 6 I then used the code from this command but added a main title to the end of the code: main="French MAT", and then copied both scatterplots to be side by side. Here’s the code:

```
scatterplot(MAT.post.FR~PA.pre.FR | Condition, reg.line=lm, smooth=FALSE,
spread=FALSE, boxplots=FALSE, span=0.5, by.groups=TRUE, data=LQB,
main=“French MAT”)
scatterplot(MAT.post.EN~PA.pre.EN | Condition, reg.line=lm, smooth=FALSE,
spread=FALSE, boxplots=FALSE, span=0.5, by.groups=TRUE, data=LQB,
main=“English MAT”)
```

Figure 6 shows scatterplots between the posttest MAT score and the posttest PA score, with separate regression lines for Condition.

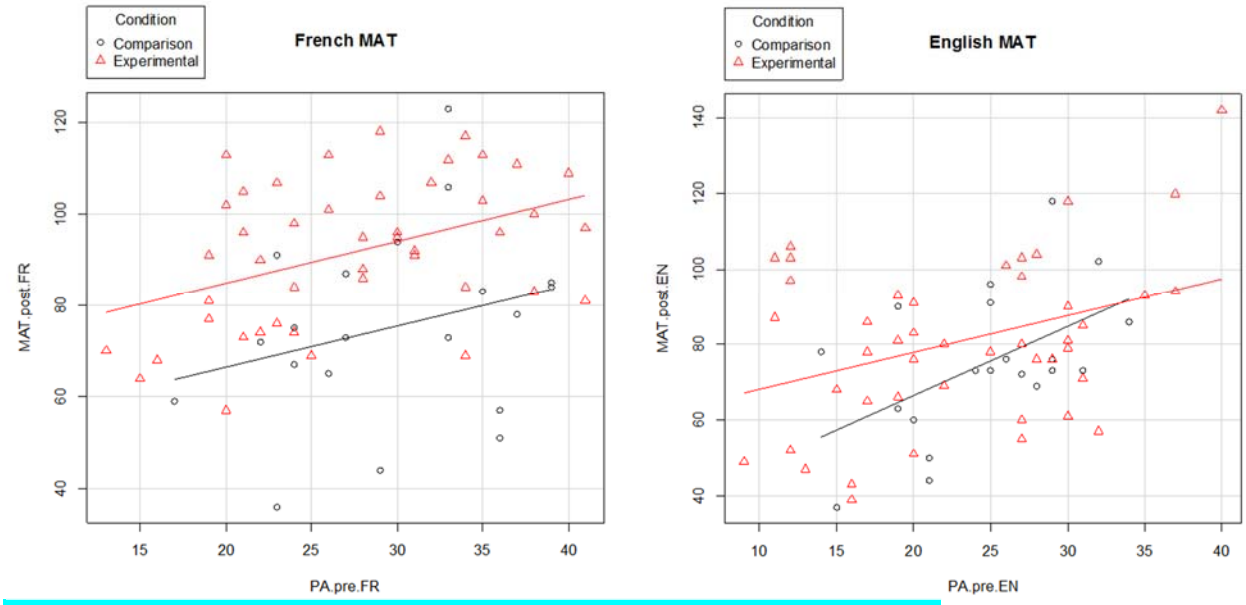

Figure 6 Scatterplots of MAT posttest and PA posttest split by Condition.

The regression lines for Condition look pretty parallel for the French MAT, but they intersect for the English MAT.

Another type of scatterplot in R uses the lattice package and can separate the scatterplots based on Condition and put them side by side. See if you like this graph better than Figure 6:

```
library(lattice)
xyplot(MAT.post.EN~PA.post.Eng|Condition,layout=c(2,1),col="black",
type=c("p","r"), data=LQB) #the argument "r" adds a regression line; to get a smooth line, use
#"smooth"
```

#you can add it in addition to the "r" or instead of it; "p" plots points for the scatterplot

Application Activity (No Answers Given)

1 In the first edition of the book the data featured in the chapter on ANCOVA was the object identification task from Lyster (2004). Use the Lyster.Oral.sav file to get the data for the object identification task. Do a numerical summary of the mean scores and standard deviations for the task by dividing the data into Condition (4 levels) and testing Times (3 times).

2 Using the same data as in Exercise #1, make a parallel coordinate plot for the object identification task with three time periods, dividing the graphs into the 4 different conditions. Can you make any observations about which groups had more gains from pretask to immediate posttask? Did gains seem to hold from immediate to delayed posttask?

3 Figure 2 shows a coplot for the English MAT. Make a similar coplot for the French MAT. Talk about any patterns you see with regards to Language or Condition.

4 Figure 3 shows a coplot for the English MAT with the English Phonological Awareness (PA) pretest and Language. Make a similar coplot for the French MAT. Do you notice any trends that change as participants have higher levels of PA?

5 Figure 6 shows scatterplots for the MAT vs. PA posttest scores. The two-way ANCOVA we'll look at later uses pretest scores and posttest scores as the continuous covariates, and divides data by Condition and Language. Make scatterplots for the French and English MAT pretest and posttest scores, and divide the data by Condition. Are the regression lines parallel?

6 Do the same analysis as in #5 but divide the data by Language. Are the regression lines parallel?

ANCOVA Design

While an ANOVA design includes one continuous dependent variable and one or more categorical independent variables, an ANCOVA design differs by being able to have continuous independent variables. Covariates can be either categorical or continuous (Howell, 2002), although in the field of second language research they are by and large continuous. Covariates are considered independent variables.

Covariates can be entered into any of the ANOVA designs—one-way ANOVAs, factorial ANOVAs, and even repeated-measures ANOVAs. Therefore, an analysis of covariance does not so much tell you about the design of the study as much as the fact that covariates will be included in it.

Any number of covariates may be entered into the research design, although Howell (2002) cautions that interpreting an analysis of covariance may be difficult enough with just one covariate, let alone more. Can covariates enter into any interactions with the other independent variables? The research design used in SPSS will not allow it. The research design in R is more flexible and it is possible, but you should not enter the covariates into any interactions with the other independent variables except in the special case where you are checking on one of the assumptions of ANCOVA (see the sections of this paper called “Assumptions of ANCOVA” and “Checking the Assumptions for the Lyster, Quiroga & Ballinger (2013) Data: Assumption 1, correlations” below). What you are looking for by including a covariate is what’s happening with the other variables when the effect of the covariate is taken away, and also whether that variable (the covariate) is statistical or not. If a covariate is found to be statistical then it has an independent effect on the variance of the dependent variable. Basically, a statistical covariate

means that the covariate does affect scores on the dependent variable. In fact, this would be the same interpretation you would make if any simple main effect of an independent variable were statistical. However, be careful not to interpret a statistical covariate as implying causation. In other words, if we find that the covariate of Phonological Awareness is a statistical predictor in a model where the dependent variable is the MAT score, it does not mean that Phonological Awareness causes differences in scores on the MAT.

Howell (2002) also warns against a use of the ANCOVA when it would result in a situation that would go against logic or common sense. If controlling for your covariate results in a design that does not exist in reality, then it doesn't make much sense to test for it statistically. For example, you probably wouldn't want to factor age of acquisition out of a research design involving early and late bilinguals. Would you really want to examine, say, context of acquisition (naturalistic, instructed, or both) among early and late bilinguals while ignoring the effects of age? Age is an important factor and it would be silly to ignore it while examining the effects of a different variable.

One controversy surrounding ANCOVA is using it to make groups "equal" when the subject cannot be randomly assigned to experiments. Although using ANCOVA for experimental designs where there is some "noise" in the data, things like individual variation that we want to remove to compare the effects of a treatment, is seen as appropriate, there are a number of arguments against using covariates to try to make subjects "equal" if they are not. Tabachnick and Fidell (2001) say that in research when subjects cannot be randomly assigned to groups it is legitimate to use ANCOVA to try to reduce differences of group means on the dependent

variable “as long as the [covariate] differences are not caused by the IV” (p. 280). In other words, in terms of our research design here, if the differences in MAT pretests (the covariate) were caused by being in the Experimental group versus the Comparison group, or if the differences in MAT pretests were caused by being a French-dominant or English-dominant or Bilingual speaker, this would be problematic.

Other authors argue that problems are increased if assignment to groups is not random, however. Clark (2014) lists some problems that arise if you use intact groups in ANCOVA: correlation between the covariate and the IV (the fact that this is a problem is detailed in the preceding paragraph), if groups differ because of the IV then partialling out the covariate may mean you are actually partialling out the effects of treatment, and the adjusted means may not represent any situation in the real world and so interpretation is problematic. Clark quotes Anderson (1963), who states that covariance may be a useful tool for reducing error variability, but “if the between groups differences on the covariate are systematic rather than chance, one may well wonder what exactly it means to ask what the data would be like if they were not what they are” (p. 170). Tabachnick and Fidell (2001) likewise caution that “adjusted means must be interpreted with great caution because the adjusted mean DV score may not correspond to any situation in the real world” (p. 280).

Figure 7 shows the research design for the Lyster, Quiroga and Ballinger (2013) analysis I will be showing in this chapter. Figure 7 shows the design for the English MAT, although the French MAT was similar and the total number of points was the same as the English MAT. The research design will depend upon the number and type of variables involved in the test if the covariate is

ignored. Figure 5 shows that, without the covariate, the design includes one categorical independent variable and one continuous dependent variable. This is a one-way ANOVA design, so with the covariate we would call it a one-way ANCOVA design.

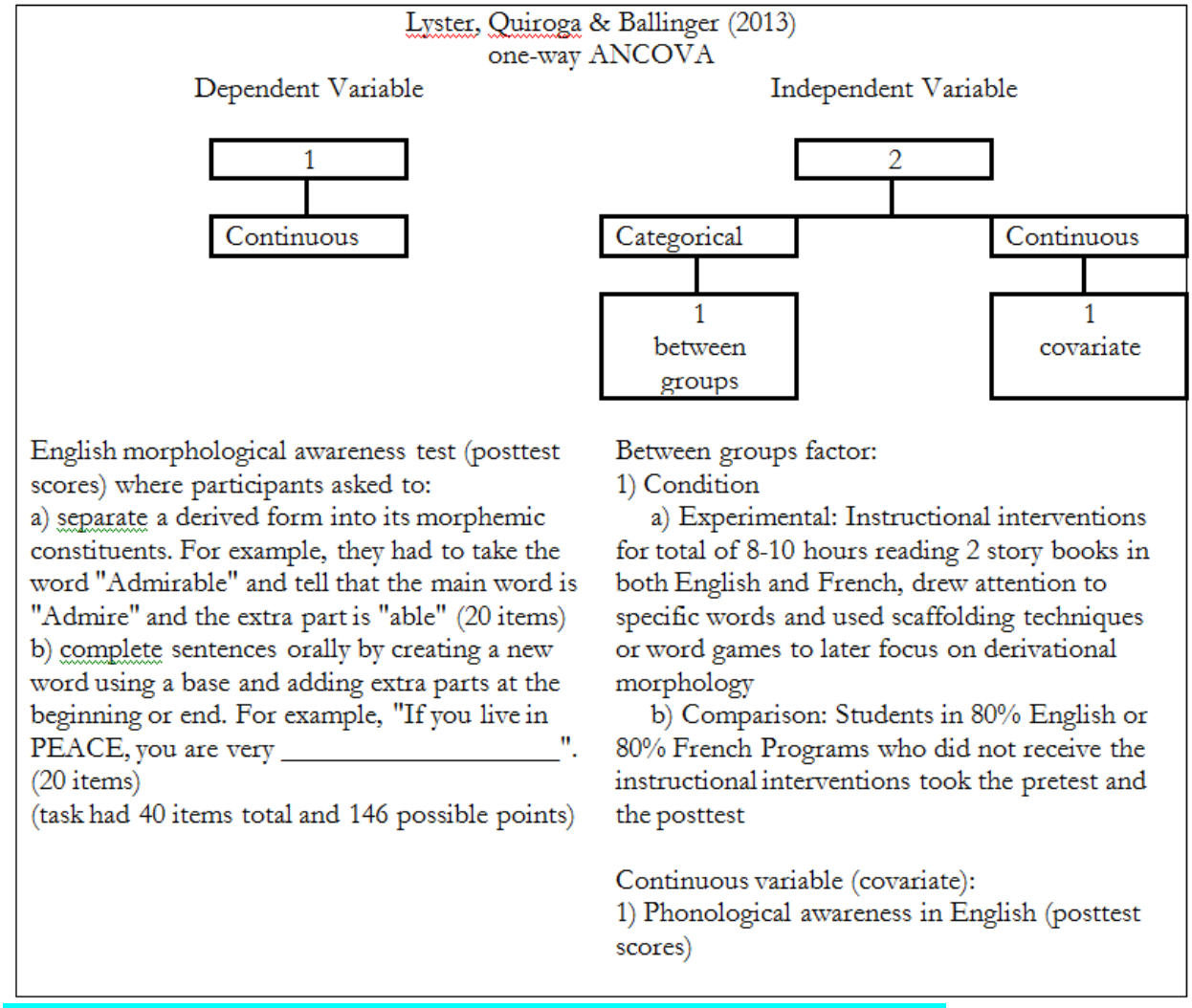

Figure 7 Lyster, Quiroga & Ballinger (2013) one-way ANCOVA design box.

Lyster, Quiroga and Ballinger (2013) also performed a two-way ANCOVA analysis on the MAT posttest scores by including both Language dominance (English-dominant, French-dominant, and Bilingual) and Group (Experimental vs. Comparison) and then used the MAT pretest score

as a covariate. Figure 8 shows the research design of this study is a factorial 3 (Language dominance) × 2 (Group) ANOVA, so we can call it a two-way ANCOVA (or 3 × 2 ANCOVA) design. If the information for the design is the same as I already gave in Figure 7, I won't repeat it in detail for Figure 8.

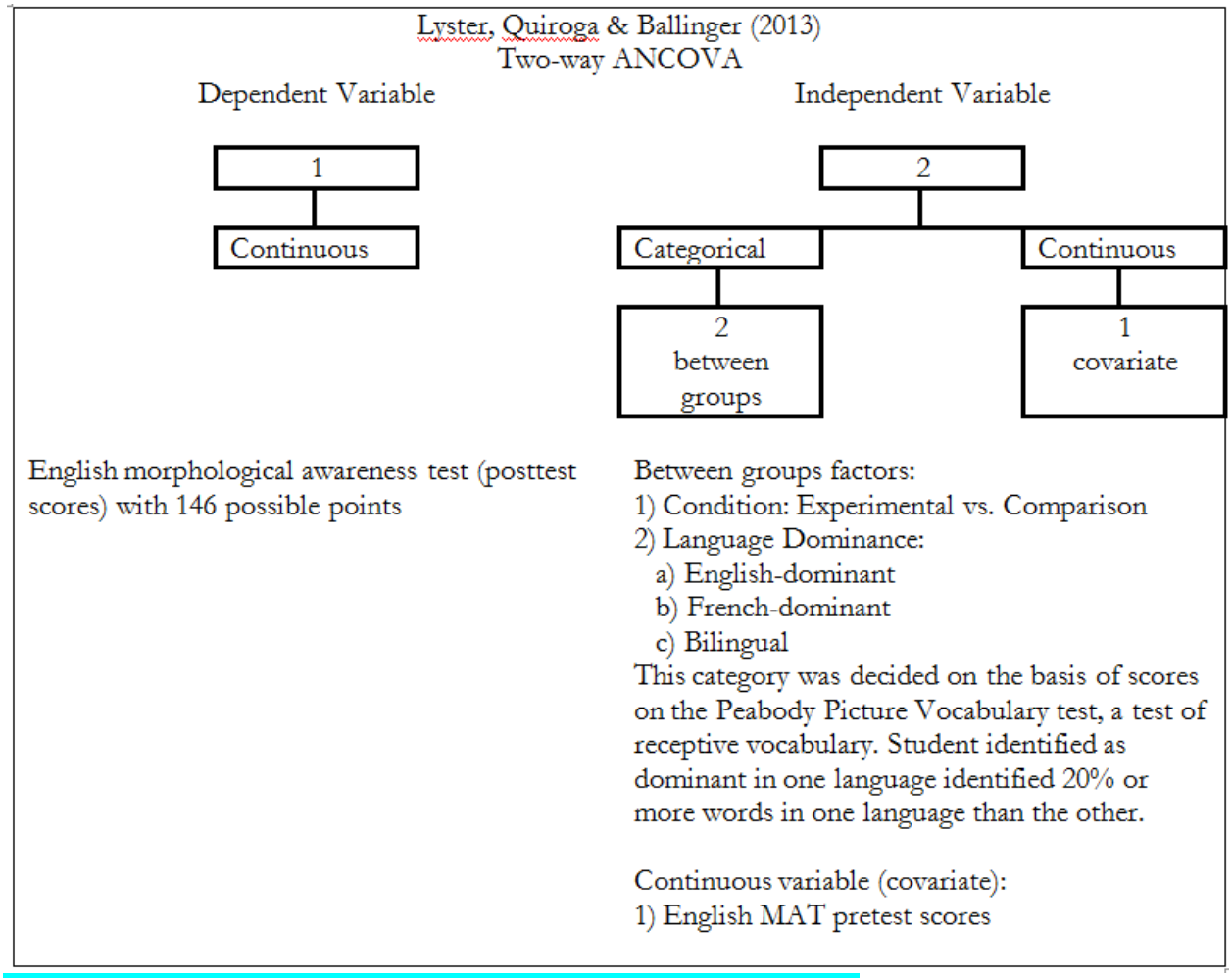

Figure 8 Lyster, Quiroga & Ballinger (2013) two-way ANCOVA design box.

Application Activity: Identifying Covariate Designs

Look at the following descriptions of experimental studies in the second language research field. Decide whether the design is one-way ANCOVA, factorial ANCOVA, or RM ANCOVA. Remember that the design depends on what test you would use if the covariates were not

included in the design. The requirements for each of these research designs is listed in Table 2.

	Dependent	*Independent*
one-way ANOVA	one, continuous	one, categorical
factorial ANOVA	one, continuous	two or more, all categorical, all between-groups
RM ANOVA	one, continuous	one or more, all categorical, at least one within-groups (repeated)

Table 2 ANOVA design requirements

1. Fraser (2007). The author wanted to compare the performance of two groups of Mandarin Chinese users of English (one living abroad, one not) on five reading tasks. The same tasks were given to the participants in both their L1 (Mandarin) and their L2 (English). There was also a covariate, which was scores on the listening portion of a measure of English language proficiency called the CELT. Using this study as a covariate would factor out differences between participants due to their English listening proficiency. Fraser specified her research design (2007, p. 380): "Thus, there was one between-subject factor (group with two levels: Canada group and China group), and two within-subject factors (language condition with two levels: L1 and L2; and Task with five levels: rauding [normal, ordinary reading], scanning, skimming, learning, memorizing).... In addition, to examine the impact of L2 proficiency on L2 reading rate and task performance, the CELT scores were used as a covariate in the analyses of the L2 data."

Research Design:	One-way ANCOVA	Factorial ANCOVA	RM ANCOVA

2. Lim and Hui Zhong (2006). The authors wanted to see how computer-assisted learning (CALL) compared to traditional reading classes in promoting reading comprehension in Korean college students learning English. There were two groups of students whose reading comprehension was measured at the beginning and end of the semester. The authors found that scores on the comprehension task were higher for the traditional learners (X = 54) than for the CALL class (X = 49) on the 100-point test, and thus decided to use the pretest comprehension task as a covariate. In this way the authors could compare the scores of the two groups, adjusted by subtracting out variation due to the pretest scores.

Research Design:	One-way ANCOVA	Factorial ANCOVA	RM ANCOVA

3 Beech and Beauvois (2005). These researchers begin their study with the assumption that in-utero influence of sex hormones can affect auditory development, which in turn can affect phonology. Problems with phonology have in turn been linked to reading disorders. The authors assert that the influence of sex hormones can be measured by a ratio between the length of the index and ring fingers. Thus, one of the variables in their study is digit ratio, and participants were split into three groups: top, middle, or bottom. The authors wanted to control (or even out) the effects of intelligence on their participants, so they used the Baddeley reasoning task. One of the statistical tests they performed looked at the effects on a silent reading task (the dependent variable) of the covariate reasoning task and the digit ratio as a categorical independent variable.

Research Design:	One-way ANCOVA	Factorial ANCOVA	RM ANCOVA

4 Larson-Hall (2008). I examined Japanese college users of English in order to see whether an earlier start in learning English would result in any advantages on an English grammaticality

judgment test (GJT). Thus my dependent variable was scores on the GJT, while my independent variables were a categorical division into earlier and later starters (those who began learning English before age 12 or 13, when it is a required subject in public schools), a continuous variable of language aptitude, and a continuous variable of the amount of total input the participants reported in English before they reached college (this was, of course, estimated!). Both language aptitude and amount of input were covariates.

Research Design:	One-way ANCOVA	Factorial ANCOVA	RM ANCOVA

5 Culatta, Reese, and Setzer (2006) (slightly adjusted from the original). The authors examined the effects on several different reading tasks of presenting skills in the first six weeks or second six weeks of instruction in a dual-language immersion kindergarten. A pretest and posttest were also given, so that time was a categorical independent variable. Whether the skill of alliteration or rhyme was presented first was categorized as the class independent variable. The dependent variable was word recognition. In order to control for differences in reading ability, scores from a standardized test of reading were used as a covariate.

Research Design:	One-way ANCOVA	Factorial ANCOVA	RM ANCOVA

Answers to Application Activity: Identifying Covariate Designs

1 RM ANCOVA because there are two within-subject variables, meaning all of the participants completed every level of these.

2 One-way ANCOVA because there is only one independent variable beside the covariate—that is group (it doesn't matter that there are only two levels—if there is a covariate we cannot use a *t*-test, we must use a univariate ANOVA design).

3 One-way ANCOVA because there is only one independent variable beside the covariate—that is the digit ratio.

4 One-way ANCOVA because there is only one categorical IV. The two covariates are continuous and do not enter into any interactions with the IV.

5 RM ANCOVA because one of the two categorical variables is within-subjects (repeated), that of Time (meaning, the same people were tested at two different times). If a researcher wanted to ignore the repeated measures and classify Time as a between-subjects variable (which would cause a loss of power and wouldn't be recommended), then this could be a factorial ANCOVA because it would have two between-subject variables, that of Time and Class.

Assumptions of ANCOVA

ANCOVA carries with it the normal assumptions of any ANOVA test, including normal distribution of data and homogeneity of variances. However, ANCOVA also carries a couple more requirements that are special to the covariate situation. I won't specifically list the assumptions of ANOVA here (they can be found in the book in Section 9.3 for one-way ANOVA, Section 10.3 for factorial ANOVA, and Section 11.4 for RM ANOVA), just the additional requirements in Table 3.

Table 13.3 Additional Assumptions for Covariates

Meeting assumptions		Factorial ANOVA
1 No strong correlations among the covariates themselves	Required?	Yes
	How to test assumption?	If you have more than one covariate, perform a correlation test on your covariates; Tabachnick and Fidell (2001) say that any covariate which correlates with another covariate at R^2=.5 or higher should be eliminated, as it is not adding much additional information independent of the other variable
	What if assumption not met?	Eliminate one of the covariates
2 The relationship between the covariate and response variable should be linear	Required?	Yes
	How to test assumption?	Look at correlation statistics or scatterplots between the covariate and the response variable; impose a regression line and a Loess line to see if the relationship is linear "enough"; this should be done with data divided into the separate groups used in your analysis, as shown in Figure 9 in this chapter.
	What if assumption not met?	1) Use Robust ANCOVA analysis which does not require linearity; 2) Try transformation of one or both variables; 3) Do not use ANCOVA
3 The slopes for each group of the regression should be the same (homogeneity of regression slopes)	Required?	Yes
	How to test assumption?	1) Check scatterplot to see if all groups are similar in their slopes; 2) include an interaction term between the covariate and the treatment—if it is statistical ($p<.05$) then you have a problem (I will demonstrate how to check this assumption in the following section)
	What if assumption not met?	1) Do not use ANCOVA; 2) use robust ANCOVA analysis which does not require homogeneity of regression

Table 3 Additional Assumptions for Covariates.

Checking the Assumptions for the Lyster, Quiroga & Ballinger (2013) Data: Assumption 1, Correlations

For the first assumption that there may be a strong correlation between covariates, neither one of the designs used by Lyster, Quiroga and Ballinger (2013) had more than one covariate, so this requirement is moot for those designs. However, there *should* be some correlation between the response (dependent) variable and the covariate. One of the reasons for conducting an ANCOVA is to look at what the relationship between the DV and IV would be if everyone were to score equally on the covariate. If the covariate has no relationship to the DV, there is nothing to adjust for!

So let's look at the correlation for the one-way ANCOVA between the English MAT posttest and the English Phonological Awareness posttest. If there is no correlation between these two variables, there is not much reason to keep the covariate in. Remember that to call for correlations in SPSS, go to ANALYZE > CORRELATE > BIVARIATE and choose the variables you want to test. For R Commander, go to STATISTICS > SUMMARIES > CORRELATION MATRIX, where you can choose as many variables as you like. Table 4 shows correlations among all of the continuous posttest variables, and the correlation between the English MAT and the English PA is r=.42, a moderate correlation, and one that argues for keeping the covariate in.

Table 4 Correlations among continuous posttest variables in Lyster, Quiroga & Ballinger 2013.

	MAT.post.EN	MAT.post.FR	PA.post.EN	PA.post.FR
MAT.post.EN	1	.44	.34	.42
MAT.post.FR		1	.31	.43
PA.post.EN			1	.78
PA.post.FR				1

For the two-way ANCOVA, again we do not have to worry about unwanted correlations between covariates as there is only one, but we do want there to be a correlation between the English MAT posttest and the English MAT pretest. Table 5 shows correlations among the pretest-posttest pairs of continuous variables. The correlation between the English MAT pretest and posttest is r=.62, a large correlation.

Table 5 Correlations between pretest and posttest variables in Lyster, Quiroga & Ballinger 2013.

	MAT. pre. EN	MAT. post. EN	MAT. pre. FR	MAT. post. FR	PA. pre. EN	PA. post. EN	PA. pre. FR	PA. post. FR
pretest		.62		.75		.77		.70

Let's also imagine a situation where we have two covariates with this data. Say we conducted a two-way ANCOVA with two covariates, that of Phonological Awareness (at the posttest) and pretest scores on the MAT for the English data. In that case we would need to test the

assumption that there is no correlation between these covariates. The correlation is r=.03, 95% CI [-0.21, 0.27]. Thus there is no problem with strong correlation between the two variables. If there is a strong correlation between the covariates, there is less room for finding any reduction of "noise" in the data due to these factors, but keeping both covariates means we will lose statistical power.

For this hypothetical two-way ANCOVA we have already tested the strength of the correlation between the covariates and the dependent variable in the paragraphs above, so we would also have reasons for keeping the two covariates. We are happy, since Tabachnick and Fidell (2001) say that ideally, "we want a very small number of [covariates], all correlated with the DV and none correlated with each other" (p. 279), and that is what we have!

Checking the Assumptions for the Lyster, Quiroga & Ballinger (2013) Data: Assumption 2, Linearity

For the second assumption of linearity between the covariate and the dependent variable (divided by groups), for the one-way ANCOVA design in Figure 7 the dependent variable is the English MAT posttest score and the covariate is Phonological Awareness in English, posttest score, divided into experimental groups. Figure 6 showed scatterplots of the correlation between the tests with regression lines drawn, but this is not exactly what we want, as we are trying to assess whether a straight line (the regression line) is the correct assumption for the data rather simply imposing the straight line as was done here. So trying out the Lattice package side-by-side scatterplots as detailed for Figure 6 but adding a Loess line along with the regression line results in Figure 9 for the English data. A straight line is definitely the best choice for the Comparison group but the Experimental group shows some curvature in the Loess line, indicating that a

straight line may not be the best choice. We can use a robust regression (and later, we will) to examine the data without the assumption of linearity. Here is the code I used in R to generate Figure 9:

```
xyplot(MAT.post.EN~PA.post.Eng|Condition,layout=c(2,1),col="black",
type=c("p","r", "smooth"), data=LQB)
```

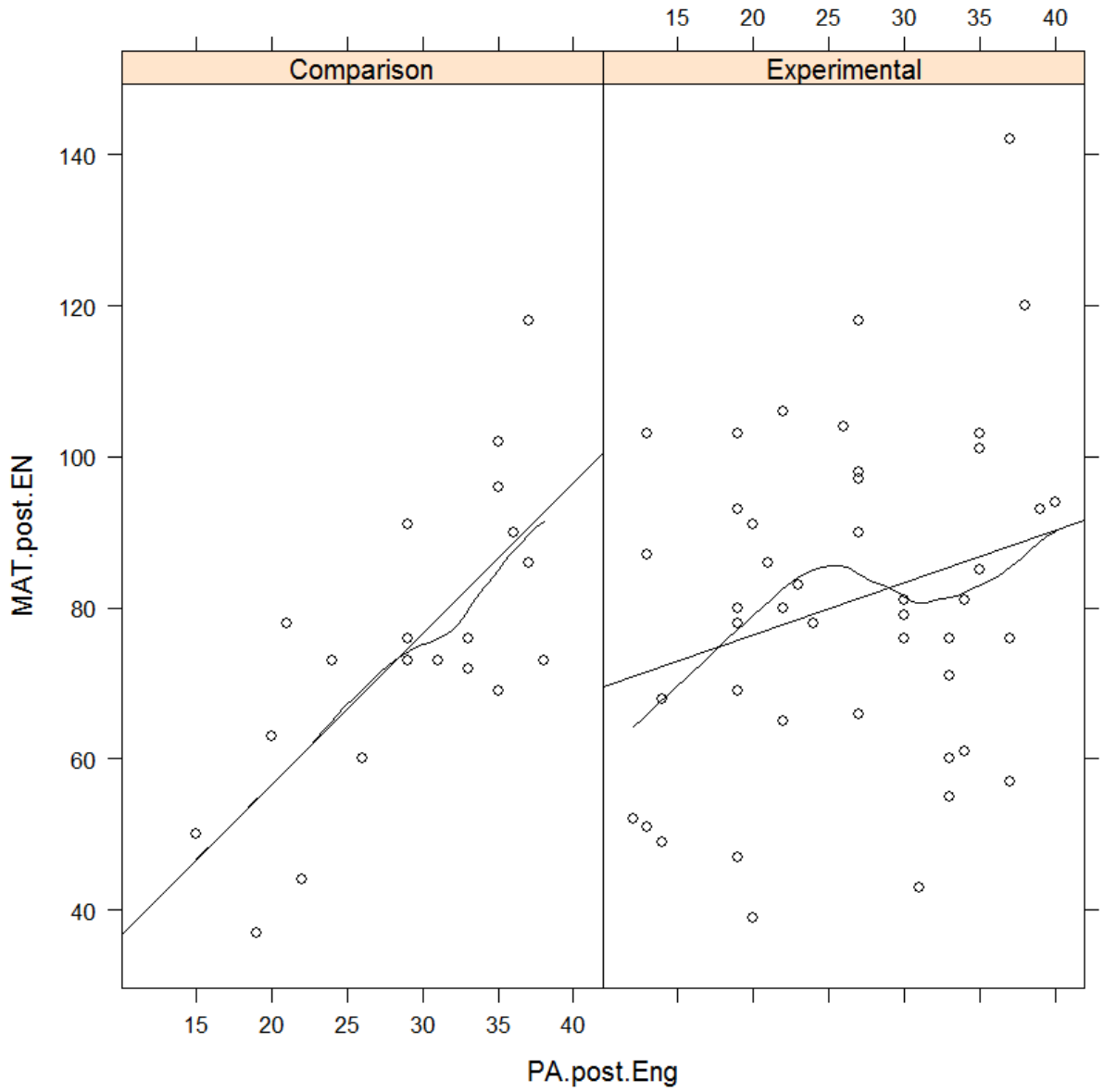

Figure 9 Examining the assumption of linearity between the covariate and the dependent variable divided by groups for the dependent variable of MAT.post.EN and the covariate of PA.post.EN.

For the two-way ANCOVA design in Figure 8 the dependent variable is still the English MAT posttest score, but now the covariate is the English MAT pretest score, and we want to divide the

data by both Language dominance and Condition, which makes things a little trickier if we want to have an understandable graph. I ended up doing this in R by subsetting the data into just one condition at a time, then calling for the data to be split by Language so I would get 3 side-by-side scatterplots with Loess and regression lines drawn on them. The code for the graph for the Experimental group is this:

```
xyplot(LQB$MAT.post.EN[subset=LQB$Condition=="Experimental"]
~PA.post.Eng[subset=LQB$Condition=="Experimental"]|Language,layout=c(3,1),col="bl
ack",
type=c("p","r", "smooth"), xlab="MAT.pre.EN", ylab="MAT.post.EN",
main="Experimental group", data=LQB)
```

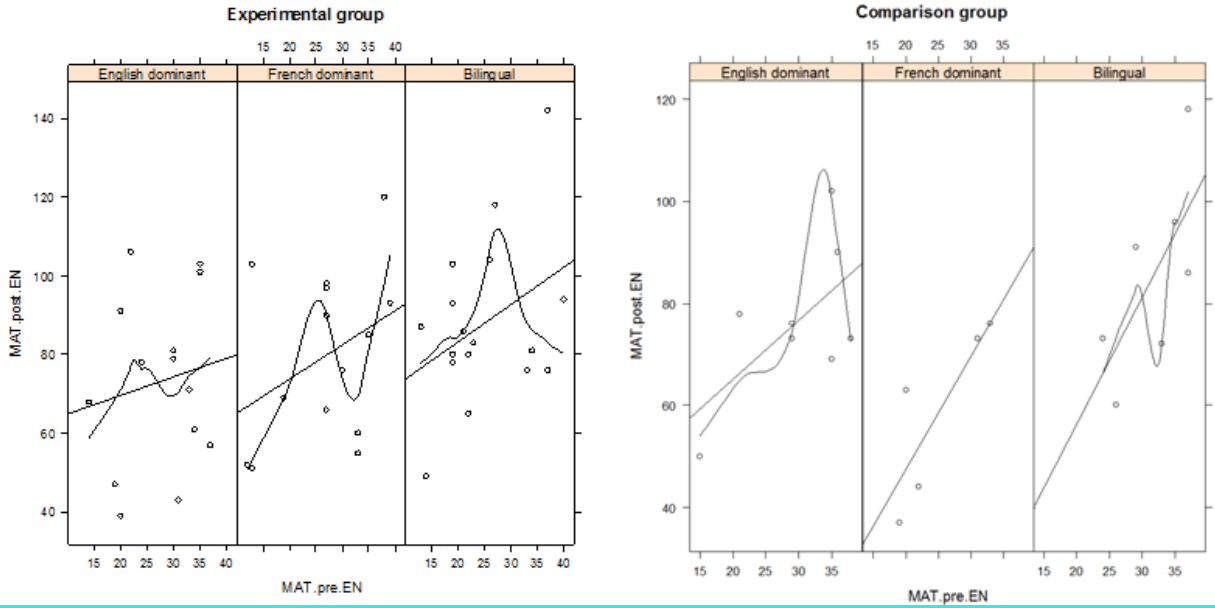

Figure 10 Examining the assumption of linearity between the covariate and the dependent variable divided by groups for the dependent variable of MAT.post.EN and the covariate of MAT.pre.EN.

Figure 10 shows the result, although for the Comparison group: French dominant there is no Loess line as the xyplot( ) command had a problem with plotting a Loess line over so few data points. In fact, we have to be careful with Loess lines if they fit the data too carefully, as then they will never look linear. Although none of the Loess lines looks like it fits a line well, we also do not see any kind of curvilinear pattern in the data, so for now we will hope that a line best describes the data. The worst that can happen is that we will lose statistical power if we violate this assumption.

If there is more than one covariate, there is also an assumption that the relationship between the covariates is linear (Tabachnick and Fidell, 2001). However, if there is no correlation between the covariates according to assumption #1 in the section of this chapter called "Assumptions of ANCOVA," I'm not sure how the scatterplot between the two variables will be linear. I guess the best we can hope for is that there are not any clear patterns and that the data are mostly randomly scattered. Examining a scatterplot between posttest phonological awareness in English and the pretest MAT in English, the data do just seem to be randomly scattered.

Checking the Assumptions for the Lyster, Quiroga & Ballinger (2013) Data: Assumption 3, Homogeneity of Regression Slopes

For the third assumption for the one-way ANCOVA we can go to Figure 6 to see if slopes are parallel for the 2 conditions (Experimental vs. Comparison). They are not parallel and cross at one point for the English test, indicating an interaction. Tabachnick and Fidell (2001) say that if there is an interaction for the lines of different groups, this indicates that the relationship between the DV and the covariate is different at different levels of the groups. This would make finding a statistical result for the ANCOVA difficult, and Tabachnick and Fidell recommend not using an

ANCOVA in that situation. If you have read the online section for Chapter 11 entitled "Performing an RM ANOVA the Mixed Effects Way," you know that a mixed-effects model is an excellent way to handle situations where groups have different slopes because mixed-effects models can let the slopes of different groups vary. So if your data violates these assumptions, you might want to think about a mixed-effects model. Other authors suggest a plain multiple regression if you violate this assumption for ANCOVA.

Another way to test whether there is **homogeneity of regression slopes** is to test for the presence of an interaction between the covariate and the treatment or grouping variable. If the interaction is not statistical, I can proceed with the normal model, according to Tabachnick and Fidell (2001, p. 292).

Checking the Homogeneity of Regression Slopes Assumption in SPSS

In order to test the homogeneity of regression slopes assumption in SPSS, we will request the interaction between the covariate and the grouping variable in our initial model, whatever it is. For the one-way ANCOVA case we are considering, we want to use the one-way ANOVA procedure. To do this, open ANALYZE > GENERAL LINEAR MODEL > UNIVARIATE. In the Univariate dialogue box, move the variables you want to analyze to the right. For the one-way ANCOVA data, move MAT.POST.EN to the "Dependent variable" box, CONDITION to the "Fixed Factor(s)" box, and PA.POST.EN to the "Covariate(s)" box. Now open the MODEL button.

Figure 11 shows a box where you can build a custom ANOVA model. Click on the button that says "Custom." Click on CONDITION, and in the "Build Term(s)" area use the arrow to move it to the right under "Model." Do the same for PA.POST.EN. Then use the Ctrl button to click on both

variables at the same time under the left-hand box “Factors & Covariates.” In the “Build Term(s)” area the “Type” button should say INTERACTION. This will let you set up the third term in the “Model” area as I have. Notice also that I have changed the sum of squares to Type II. Press CONTINUE and then OK to run the analysis.

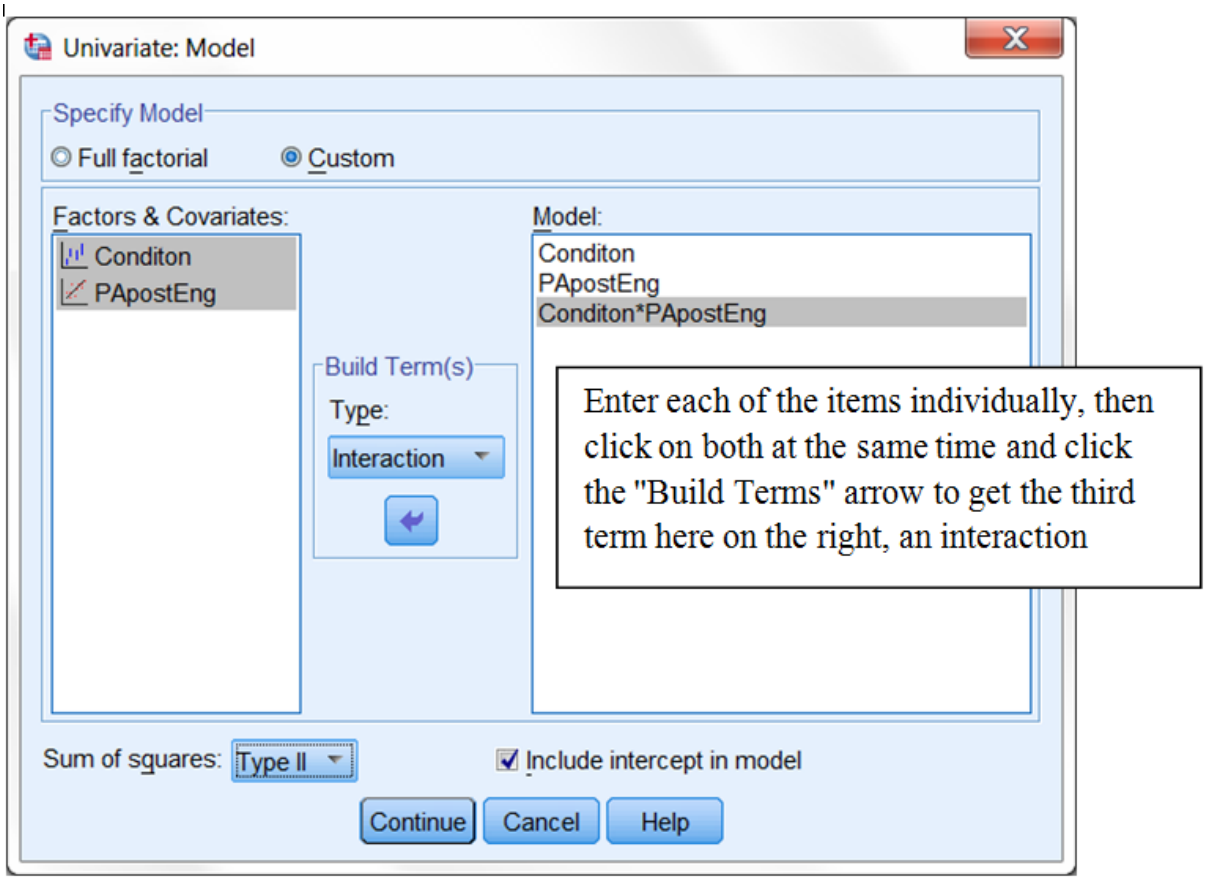

Figure 11 Creating an interaction term in a custom ANCOVA model.

The main output table, the “Tests of Between-Subjects Effects,” shows that the interaction (Condition*PApostEng) is not statistical (p = .08). This is one of those times when we are hoping the p-value will be *larger* than p = 0.05. If it is, we can conclude that the slopes of the groups on the covariate are parallel enough and that there is homogeneity of regression. If there were a statistical interaction, then that would mean that the groups performed differently on the

covariate. In the output, shown in Table 6, the interaction (Cond*PreObjectID) is the only thing that you need to look at; you can ignore the other parts of the output.

Tests of Between-Subjects Effects

Dependent Variable: MAT-post-EN

Source	Type II Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5556.751[a]	3	1852.250	4.893	.004
Intercept	13913.942	1	13913.942	36.756	.000
Conditon	1711.972	1	1711.972	4.523	.038
PApostEng	3899.353	1	3899.353	10.301	.002
Conditon * PApostEng	1195.176	1	1195.176	3.157	.081
Error	23091.249	61	378.545		
Total	434313.000	65			
Corrected Total	28648.000	64			

a. R Squared = .194 (Adjusted R Squared = .154)

Table 6 Testing assumptions in a one-way ANCOVA in SPSS.

The fact that the *p*-value of the interaction is quite low as well as the fact that the scatterplot showed the lines crossing may make us cautious about using ANCOVA in this situation, although if we interpret Tabachnickand Fidell (2001) strictly, we could do it. It seems it may be more prudent to use it in the case of the French MAT (which did not show any interaction of groups in Figure 6).

In the case of an ANCOVA where you have two independent variables or more than one covariate, I do not know whether it is necessary to simply check the relationship of each of the independent variables with each covariate separately, or whether all of the variables in the model need to be checked together. In other words, for Lyster, Quiroga and Ballinger's (2013) two-way ANOVA, do we need to look at two interactions—one between the covariate of the pretest MAT and Condition (Experimental vs. Comparison), and one between pretest MAT and Language

dominance (French-dominant, English-dominant or Bilingual), or should we look at the three-way interaction between the pretest MAT, Condition and Language dominance? None of the books that I have examined have directly address this issue, and I don't know the answer. My hunch is that the interactions should be considered separately. Tabachnick and Fidell (2001) give an example of an ANCOVA where the question is whether political attitude is affected by geographic region and religious affiliation (so 2 IVs), with two covariates of socioeconomic status and age, and say that "[w]ith more than one IV, separate statistical tests are available for each one" (p. 278). It seems to me that this means that assumptions should also be checked separately.

Checking the Homogeneity of Regression Slopes Assumption in R

In order to test the homogeneity of regression slopes assumption in R, we will request the interaction between the covariate and the grouping variable in our initial model, whatever it is. For the one-way ANCOVA case we are considering, we want to use the one-way ANOVA model but add in the covariate *plus* an interaction between the dependent variable and the covariate. Basically this will mean setting up a full-factorial ANOVA with the two variables in the case of the one-way ANOVA, like this:

```
Model1=aov(MAT.post.EN~Condition*PA.post.Eng, data=LQB)
summary(Model1)
```

```
                    Df Sum Sq Mean Sq F value  Pr(>F)
Condition            1    462     462   1.221 0.27349
PA.post.Eng          1   3899    3899  10.301 0.00212 **
Condition:PA.post.Eng  1   1195    1195   3.157 0.08057 .
Residuals           61  23091     379
```

The ANOVA output shows that the interaction is not statistical (p = .08). This is one of those times when we are hoping the p-value will be *larger* than p = 0.05. If it is, we can conclude that the slopes of the groups on the covariate are parallel enough and that there is homogeneity of regression. If there were a statistical interaction, then you can see that that would mean that the groups performed differently on the covariate. In the case of checking for the assumption of homogeneity of slopes, the interaction is the only row of the ANOVA that you need to look at; you can ignore the other parts of the output.

The fact that the p-value of the interaction is quite low as well as the fact that the scatterplot showed the lines crossing may make us cautious about using ANCOVA in this situation, although if we interpret Tabachnick and Fidell (2001) strictly, we could do it. It seems it may be more prudent to use it in the case of the French MAT (which did not show any interaction of groups in Figure 6).

In the case of an ANCOVA where you have two independent variables or more than one covariate, I do not know whether it is necessary to simply check the relationship of each of the independent variables with each covariate separately, or whether all of the variables in the model need to be checked together. In other words, for Lyster, Quiroga and Ballinger's (2013) two-way ANOVA, do we need to look at two interactions—one between the covariate of the pretest MAT and Condition (Experimental vs. Comparison), and one between pretest MAT and Language dominance (French-dominant, English-dominant or Bilingual), or should we look at the three-way interaction between the pretest MAT, Condition and Language dominance? None of the books that I have examined have directly address this issue, and I don't know the answer. My

hunch is that the interactions should be considered separately. Tabachnick and Fidell (2001) give an example of an ANCOVA where the question is whether political attitude is affected by geographic region and religious affiliation (so 2 IVs), with two covariates of socioeconomic status and age, and say that "[w]ith more than one IV, separate statistical tests are available for each one" (p. 278). It seems to me that this means that assumptions should also be checked separately.

Application Activity for Checking ANCOVA Assumptions (No Answers Given)

1 The tests in the section of this paper called "Assumptions of ANCOVA" looked mainly at the ANCOVA that used the English MAT, but Lyster, Quiroga and Ballinger (2013) also used a one-way ANCOVA to look at the French MAT as well. Go through the the assumptions and see whether the French MAT satisfies the extra ANCOVA assumptions.

2 Lyster, Quiroga and Ballinger's two-way ANCOVA has one dependent variable (MAT posttest), one covariate (MAT pretest) and two independent variables (Condition and Language dominance). Examine the data to see if it satisfies the three extra assumptions of ANCOVA (use the English test):

   a) Check to make sure there *are* correlations between the MAT pretest (the covariate) and MAT posttest (the dependent variable

   b) Check whether the relationship between the MAT pretest and the MAT posttest is linear

   c) Check whether the slopes for Condition in a scatterplot of MAT pretest and MAT posttest are parallel and use a one-way ANOVA model to check for interaction between MAT posttest and Condition; check whether the slopes for Language

dominance in a scatterplot of MAT pretest and MAT posttest are parallel and use a one-way ANOVA model to check for interaction between MAT posttest and Language.

3 Do the same as #2 for the French MAT.

4 Look at the data file Lyster.Oral.sav (if using R, import as LysterO). Assume that you want to run a one-way ANCOVA on the posttest oral object identification task (PostObjectID) with Condition (treatment group) as the IV and pretest scores on the task as covariates. Check to see if the data satisfy the three extra assumptions of ANCOVA.

5 Later in the chapter we will look at the Larson-Hall 2008 data (if using R, import as larsonhall2008). This study was described in an earlier section of this paper entitled "Application Activity: Identifying Covariate Designs." The response variable is scores on a GJT and the independent variable is classification as an earlier or later starter (Early). This study used 2 covariates, one language aptitude and the other amount of input. Examine the data to see if it satisfies the three extra assumptions of ANCOVA:

a) Check to make sure there *are* correlations between the covariates and the dependent variable; make sure the covariates themselves are not correlated

b) Check whether the relationship between aptitude and the IV of GJT is linear, and also whether the relationship between amount of input and GJT is linear.

c) Check whether the slopes for Early in a scatterplot of GJT and aptitude are parallel and use a one-way ANOVA model to check for interaction between GJT and Early; check whether the slopes for Early in a scatterplot of GJT and amount of input are parallel and use a one-way ANOVA model to check for interaction between GJT and Early.

Performing an ANCOVA

Performing a One-way ANCOVA with One Covariate in SPSS

Performing an ANCOVA involves using SPSS's GENERAL LINEAR MODEL choices in the ANALYZE menu. The three pertinent choices in that menu are: UNIVARIATE, MULTIVARIATE, AND REPEATED MEASURES (the VARIANCE COMPONENTS choice is for mixed-effects models, which are discussed in the online document called "RM ANOVA the mixed effects way", but SPSS is not illustrated). Use the UNIVARIATE command for one-way or factorial ANOVA. Use the MULTIVARIATE command for MANOVA. I don't discuss MANOVA in this book, but it would involve analyzing more than one dependent variable in the same test. Use the REPEATED MEASURES command when you have any independent variable that measures the same people more than once.

The first ANCOVA I will analyze from the Lyster, Quiroga and Ballinger (2013) research design is the one-way ANCOVA, so I'll use the UNIVARIATE choice. So I choose ANALYZE > GENERAL LINEAR MODEL > UNIVARIATE and move MATPOSTEN to the "Dependent variable" box, CONDITION to the "Fixed Factor(s)" box, and PAPOSTENG to the "Covariate(s)" box. This corresponds to the ANCOVA for the English MAT reported in Lyster, Quiroga and Ballinger (2013) on p. 184.

You should now open some of the buttons that are found on the right side of the UNIVARIATE dialogue box. First, open the MODEL button and choose a Type II sum of squares analysis if you agree with me that this is the best choice (see Section 10.4.4 for arguments). If you previously explored a "Custom" model to check for interactions between the covariate and the

IV, change this back to the "Full Factorial" model. Because there's only one non-covariate independent variable for the Lyster, Quiroga and Ballinger (2013) data, a means plot would not be possible, so there's no reason to open the PLOT button. However, if you had two or more non-covariate independent variables you might like to call for some means plots using this button. To get diagnostic tests of your residuals, open the Save button and check "Unstandardized" under "Predicted Values" and "Cook's Distance" under Diagnostics.

If your independent variable has more than two levels, Tabachnick and Fidell (2001, p. 313) state that one should be able to run post-hoc tests after an ANCOVA run, but in SPSS the POST HOC button will become unavailable if a covariate is entered, as shown in Figure 12. To obtain post-hoc tests on the independent variable you can instead open the OPTIONS button and move the independent variables to the "Display Means for" box, as shown for CONDITION in Figure 12. If you tick the box that says "Compare main effects," pairwise comparisons will be done for all of the levels of the IV. The "Confidence interval adjustment" drop-down menu in the SPSS Options dialogue box gives only three choices for ways to adjust the *p*-values of the pairwise comparisons—LSD, which means no adjustments are made, Bonferroni, which means 0.05 is divided by the total number of comparisons that are made, and Sidak, which is a conservative familywise error rate adjustment, but slightly less conservative than the Bonferroni. As I am an advocate for higher power, I recommend using the LSD choice. If you are nervous about having too many comparisons, I would recommend still using LSD and then going to R to use the FDR adjustment (R code shown in Section 8.3.1) to adjust *p*-values.

While the OPTIONS button is open, also check the "Descriptive statistics," "Estimates of effect

size," "Spread vs. level plot" and "Residual plot" boxes. If you want a Levene's test of homogeneity of variances, tick "Homogeneity Tests." When finished, press CONTINUE. Open the BOOTSTRAP button and click on the box that says "Perform bootstrapping" and set the number of samples to 2000. Change the type of "Confidence Intervals" to "BCa" and press CONTINUE, and then press OK in the main dialogue box.

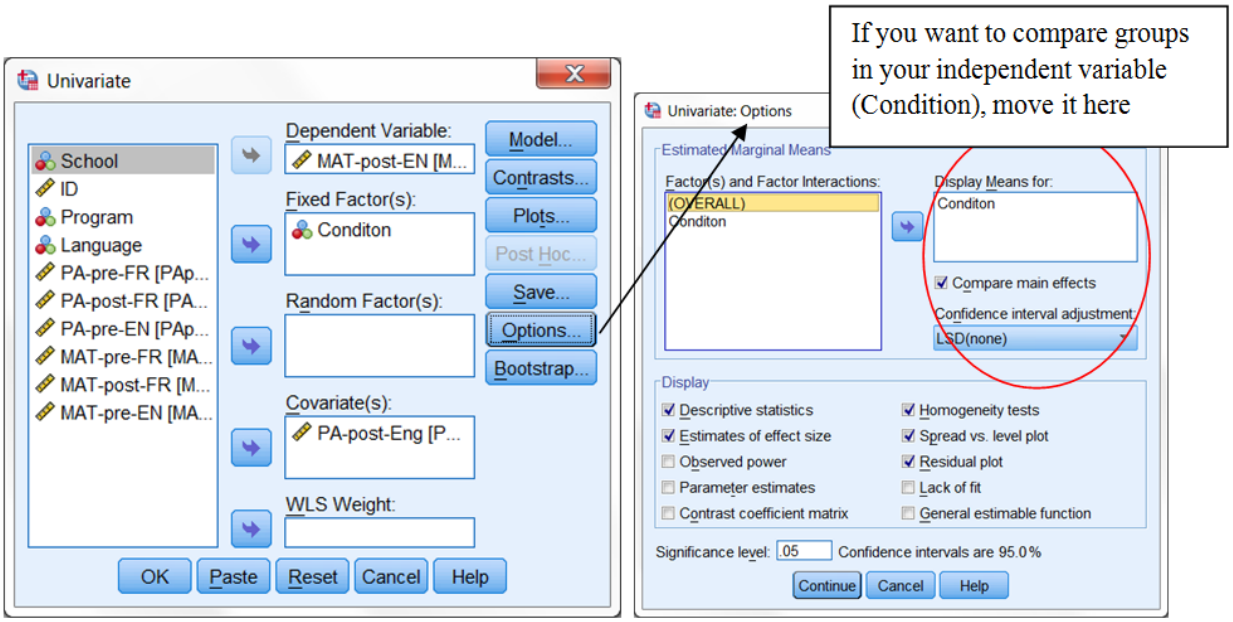

Figure 12 Calling for ANCOVA in SPSS.

ANCOVA Output in SPSS

The beginning of the output for a one-way ANCOVA in SPSS gives a box for "Between-Subjects Factors," just listing the groups, their labels and counts, and then "Descriptive Statistics." For the English posttest MAT, the data are divided into just scores for the Comparison group (M=75.0, SD=19.3, N=20) and the Experimental group (M=80.8, SD=21.9, N=45). Next comes "Levene's test of equality of error variances", which has a *p*-value of $p = .146$, which gives us an additional reason to believe that variances are equal (it is only when $p<.05$ that we worry about

variances being unequal across groups). The next box, “Tests of Between-Subjects Effects,” is the main output we want to examine, and is in Table 7.

The two main lines we care about are the ones labeled PApostEng and Condition. We see that the effect of Condition is $p = .11$, and would report this result by saying that “the participants did not differ in their performance by the experimental group they belonged to, when scores were adjusted for posttest Phonological Awareness, ($F_{1,62} = 2.61$, $p = .11$, partial eta-squared = .04).” We can see the effect size (partial eta-sqared) is also very small, and we could see from the boxplots in Figure 5 that median differences between the groups were small. This means that, when MAT posttest scores are adjusted for PA scores, condition is not a factor in explaining variance in the model.

We also are interested to know whether our covariate, Phonological Awareness (posttest, in English) was statistical, as this would mean that it provides adjustment of the scores on the dependent variable, and it can be interpreted as any independent variable in a regression model would be (Tabachnick and Fidell, 2001). Table 7 shows that the effect of the posttest PA is statistical ($F_{1,62}$=9.96, p=.002, partial eta-squared = .14), meaning that the scores are adjusted for the effect of this variable. This means that differences in Phonological Awareness accounted for 14% of the variance in the Morphological Awareness Test, if differences in Condition are held constant. By the way, if you compare the results for the English MAT given here and those given in Lyster, Quiroga and Ballinger (2013), the general outcome is the same but the actual numbers for the statistical test differ because the authors excluded one outlier from this analysis, a French-

dominant participant who scored 70 on the pretest but 142 on the posttest. You can see this participant's outlier score in the MAT posttest English boxplot in Figure 5.

Overall, this model accounted for R^2=15% of the variance in the model (13% for adjusted R^2). This is a rather small amount of the variance and means we have a lot of unexplained error in the model.

Table 7 Main results for testing a one-way ANCOVA in SPSS.

Tests of Between-Subjects Effects

Dependent Variable: MAT-post-EN

Source	Type II Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	4361.575[a]	2	2180.787	5.567	.006	.152
Intercept	14683.581	1	14683.581	37.485	.000	.377
PApostEng	3899.353	1	3899.353	9.955	.002	.138
Conditon	1020.942	1	1020.942	2.606	.112	.040
Error	24286.425	62	391.717			
Total	434313.000	65				
Corrected Total	28648.000	64				

a. R Squared = .152 (Adjusted R Squared = .125)

We can say more about the effect of PA scores by looking at the next piece of the SPSS output found under the title "Estimated Marginal Means." These results are shown in Table 8. The estimates for the means shown in Table 8 are the posttest means for the English MAT, but adjusted for PA, so these means are different from those seen in the descriptive statistics at the beginning of the output. The **adjusted mean** is the mean score with the influence of the covariate factored out, and in this dataset you can see there is more difference between scores of the Comparison and Experimental groups in these adjusted means than in the original means: now the Comparison group's mean is smaller (M=73.0, SE=4.47, N=20) and the Experimental group's mean is higher (M=81.7, SE=2.96, N=45). These point estimates look fairly different,

but if we look at the 95% confidence intervals for the estimated means for the groups, they are quite wide for the Comparison group (almost a 20-point difference, from 64.0 to 81.9). The confidence interval for the mean for the Experimental groups is smaller [75.8, 87.6] but the large and overlapping range means that we are not able to find a difference in effect between groups. As a note, it is not the case that just because confidence intervals overlap this automatically means that there is no difference between groups. Cumming and Finch (2005) give as a "rule of eye" for the interpretation of 2 CIs of two independent groups that they can overlap up to about half of the average length of both arms of the CI[2] and still be statistical at the .05 level. On the other hand, if the confidence intervals do not overlap, this is evidence that there is a strong difference between groups. Anyway, my point is that the estimated adjusted means gives you a point estimate, but the CIs show us these point estimates are not very precise and we cannot be very confident of where the true means lie.

Table 8 Scores on the dependent variable adjusted for the covariate in the one-way ANCOVA.

Estimated Marginal Means

Conditon

Estimates

Dependent Variable: MAT-post-EN

Conditon	Mean	Std. Error	95% Confidence Interval		Bootstrap for Mean[xq]			
							BCa 95% Confidence Interval	
			Lower Bound	Upper Bound	Bias	Std. Error	Lower	Upper
Comparison	72.965[a]	4.472	64.025	81.905	.203	3.823	65.381	80.921
Experimental	81.682[a]	2.964	75.757	87.608	.005	3.201	75.482	87.913

a. Covariates appearing in the model are evaluated at the following values: PA-post-Eng = 27.169.

xq. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Next in the output comes the "Pairwise Comparisons" table, which shows comparisons between the different groups. In this example there are only two groups, so if there had been a real

[2] Precisely, this average length of both of the arms of the CI is called the margin of error, and for this example would be [(81.9-64.0)/2 + (87.6-75.8)/2]/2= 7.4, meaning the CIs could overlap by about 3.5 points (half of the average length of both arms).

difference between the groups we would have known which group was better than the other just by looking at the mean scores. If you have more than two groups, the pairwise comparison box will help you know which groups are different from the others. These pairwise comparisons which have been adjusted for the covariate can be interpreted as for any other post-hoc tests, but remember that we asked for no adjustment on our *p*-values (the LSD choice for pairwise comparison adjustments). The SPSS output returns confidence intervals so we can report those instead of *p*-values. If you called for bootstrapping there will also be a different box with bootstrapped confidence intervals for the pairwise comparisons.

Table 9 Pairwise comparisons between groups, normal and bootstrapped, in SPSS.

Pairwise Comparisons

Dependent Variable: MAT-post-EN

(I) Conditon	(J) Conditon	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Comparison	Experimental	-8.718	5.400	.112	-19.512	2.077
Experimental	Comparison	8.718	5.400	.112	-2.077	19.512

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Bootstrap for Pairwise Comparisons

Dependent Variable: MAT-post-EN

(I) Conditon	(J) Conditon	Mean Difference (I-J)	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
Comparison	Experimental	-8.718	.198	4.911	.091	-18.411	1.005
Experimental	Comparison	8.718	-.198	4.911	.091	-.974	18.343

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

At the end of the output will be interaction plots if you called for them, and diagnostic plots for the residuals. The spread versus level plots test for the assumption of homogeneity of variances in the residuals. There is also a table with various combinations of standardized residuals,

observed and predicted residuals. Examine the Std. Residual vs. Predicted table; the data should look randomly distributed. To examine the assumption of normality of residuals, look to the columns appended to the end of your dataset called PRE_1 and COO_1. The PRE_1 stands for the unstandardized predicted values from the regression but we aren't going to do anything with that.The COO_1 stands for Cook's distance, which is a measure of influence (it measures the effect of deleting a given observation) and you can examine this data by either looking for values larger than 1 or plotting a scatterplot of the Cook's distance value crossed with the ID number of individuals and looking for values that stand out.

Performing a One-way ANCOVA with One Covariate in SPSS

1. Choose ANALYZE > GENERAL LINEAR MODEL > UNIVARIATE.
2. Put dependent variable in "Dependent Variable," independent variables in "Fixed Factor(s)," and covariate in "Covariate(s)."
3. Open the MODEL button and create a custom model that includes an interaction between the covariate and your fixed factor(s). If this is statistical, stop and do not continue with your ANCOVA. If this is not statistical, go back and click the "Full Factorial" button, which removes the interaction, and also change to a Type II Sum of Squares.
4. Open the OPTIONS button and tick "Descriptive statistics" and "Estimates of effect size. Move between-group variables over to "Display Means for" in order to get post-hoc comparisons.

Performing a One-way ANCOVA with One Covariate in R

We will need to use R for this, not R Commander. Performing an ANCOVA in R requires no more effort than simply including an extra term in the regression equation for an ANOVA. For example, this would be the regression equation if we simply wanted to perform a one-way ANOVA to examine the effect of Condition on MAT posttest English scores:

```
aov(MAT.post.EN~Condition, data=LQB)
```

Now we can just add the PA.post.Eng (the covariate) term to the right of the tilde to make an ANCOVA model:

```
aov(MAT.post.EN~Condition + PA.post.Eng, data=LQB)
```

We could use either aov( ) or lm( ) to model the regression; the only difference is in the format of the output, which in any case can be changed, so there is essentially no difference. I will show the aov( ) modeling here.

Crawley's (2007) recommendation is to start out the analysis with a maximal model, one which involves the covariate in an interaction with the independent variable (this is indicated by the "*" between the IV of Condition and the covariate of PA.post.Eng). Although the syntax in R is simple, Crawley says the maximal model will have different slopes and intercepts for each level of the factor. Remember that we tried a model with an interaction term in the section of this paper called "Checking the Assumptions for the Lyster, Quiroga & Ballinger (2013) Data: Assumption 3, homogeneity of regression slopes." We were hoping that there was no interaction between the covariate and the IV, so we can think of this maximal model as starting out assuming the worst, but hoping that the model with no interaction will be better.

```
Model1= aov(MAT.post.EN~Condition*PA.post.Eng, data=LQB)
```

Look then at the output by using the anova( ) command:

```
> anova(Model1)
Analysis of Variance Table

Response: MAT.post.EN
                     Df  Sum Sq Mean Sq F value   Pr(>F)
Condition             1   462.2   462.2  1.2210 0.273494
PA.post.Eng           1  3899.4  3899.4 10.3009 0.002121 **
Condition:PA.post.Eng 1  1195.2  1195.2  3.1573 0.080575 .
Residuals            61 23091.2   378.5
```

We see that the effect of Condition is not statistical, at p = .27, while the effect of the covariate, PA in English at the posttest, is statistical (p=.002), and the effect of the interaction between the covariate and the IV has a p-value of p = .08. We are not finished, however, as we want to find the best model that fits our data.

Note that order can matter in the regression equation (Crawley, 2007), so that we could actually get a different outcome if we model with the covariate first. We don't get an overall different result in this case, but the p-values are different as the order changes the sum of squares, which affects the calculation of p-values. Regression effect sizes (the "Estimate" in the lm( ) model), however, will not vary, nor will standard errors.

```
OrderMatters=aov(MAT.post.EN~ PA.post.Eng* Condition, data=LQB)
anova(OrderMatters)
```

```
                      Df  Sum Sq Mean Sq F value   Pr(>F)
PA.post.Eng            1  3340.6  3340.6  8.8249 0.004246 **
Condition              1  1020.9  1020.9  2.6970 0.105682
PA.post.Eng:Condition  1  1195.2  1195.2  3.1573 0.080575 .
Residuals             61 23091.2   378.5
```

Next we will try a model that does not include the interaction term. Use the update( ) command, which lets you change your model without typing everything again. Be careful with the syntax of the update( ) command; it is "comma tilde dot minus":

```
Model2=update(Model1,~.-Condition:PA.post.Eng)
```

Now we will use anova( ) to compare the two models and see if there is any statistical difference between them:

```
> anova(Model1, Model2)
Analysis of Variance Table

Model 1: MAT.post.EN ~ Condition * PA.post.Eng
Model 2: MAT.post.EN ~ Condition + PA.post.Eng
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     61 23091
2     62 24286 -1   -1195.2 3.1573 0.08057 .
```

There is no statistical difference between groups since $p > .05$, so we will choose the simpler model, Model2, and we are glad there is no need to assume an interaction between the covariate and the IV. The next step, according to Crawley (2007), is to remove the IV from the model and see if there is any difference in the models:

```
Model3=update(Model2,~.-Condition)
```

```
anova(Model2, Model3)
```

```
Analysis of Variance Table

Model 1: MAT.post.EN ~ Condition + PA.post.Eng
Model 2: MAT.post.EN ~ PA.post.Eng
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     62 24286
2     63 25307 -1   -1020.9 2.6063 0.1115
```

There is no difference between models, so logically the next step would be to remove the covariate, which would leave just the number 1, indicating that the best model is the intercept, which is a model with just the overall grand mean.

```
Model4=update(Model3,~.-PA.post.Eng)
anova(Model3, Model4)
```

```
Analysis of Variance Table

Model 1: MAT.post.EN ~ PA.post.Eng
Model 2: MAT.post.EN ~ 1
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     63 25307
2     64 28648 -1   -3340.6 8.3162 0.005371 **
```

This is too extreme a step, as there is a difference between models, so we conclude that the minimal adequate model is one that models variation in scores on the English MAT only by Phonological Awareness scores on the posttest (Model3). However, for traditional reports of ANCOVA, this may not be satisfactory, and one might want to go back to the model with both

terms to report on it (Model2). Because this is modeled with aov( ), either the anova(Model2) or the summary(Model2) command will return this table:

```
Analysis of Variance Table

Response: MAT.post.EN
             Df  Sum Sq Mean Sq F value   Pr(>F)
Condition     1   462.2   462.2  1.1800 0.281563
PA.post.Eng   1  3899.4  3899.4  9.9545 0.002476 **
Residuals    62 24286.4   391.7
```

We see that the effect of Condition is not statistical as the *p*-value is above $p = .05$, and would report this result by saying that "the participants did not differ in their performance by the experimental group they belonged to, when scores were adjusted for posttest Phonological Awareness, ($F_{1,62} = 1.18$, $p = .28$)." We could see from the boxplots in Figure 5 that median differences between the groups were small. The results mean that, when MAT posttest scores are adjusted for PA scores, condition is not a factor in explaining variance in the model.

We also are interested in knowing whether our covariate, Phonological Awareness (posttest, in English) was statistical, as this would mean that it provides adjustment of the scores on the dependent variable, and it can be interpreted as any independent variable in a regression model would be (Tabachnick and Fidell, 2001). The output shows that the effect of the posttest PA is statistical ($F_{1,62}=9.95$, $p =.002$), meaning that the scores on the MAT are in fact affected by this variable. By the way, if you compare the results for the English MAT given here and those given in Lyster, Quiroga and Ballinger (2013), the general outcome is the same but the actual numbers for the statistical test differ because the authors excluded one outlier from this analysis, a French-dominant participant who scored 70 on the pretest but 142 on the posttest. You can see this participant's outlier score in the MAT posttest English boxplot in Figure 5.

Alternatively, we could report that the minimal adequate model for the MAT posttest scores was one that included only the PA posttest scores and then describe that model.

For effect sizes, the SPSS output gives partial eta-squared effect sizes for each term of the regression equation. In R we have more say over what type of effect size we want to call for. We could get a partial eta-squared effect size by using the heplots package, etasq( ) command, like this:

```
install.packages("heplots")
library(heplots)
etasq(Model2, anova=T)
```

```
Anova Table (Type II tests)

Response: MAT.post.EN
             Partial eta^2  Sum Sq Df F value    Pr(>F)
Condition         0.040342  1020.9  1  2.6063 0.111516
PA.post.Eng       0.138345  3899.4  1  9.9545 0.002476 **
Residuals                  24286.4 62
```

So we can see that the effect size for the IV, experimental Condition, has a partial eta-squared = .04, and the covariate, the posttest PA test in English, has a partial eta-squared = .14, meaning that the Condition explains 4% of the variance accounted for when all other factors are fixed, and the PA test explains 14%.

But since this is a regression, we might also want to use the calc.relimp( ) command from the relaimpo package, which we saw previously in the book in Section 7.4.5. The lmg metric

assesses the contribution of each term in the regression to the total variance of the DV, and is basically the squared semipartial correlation (sr^2). To me this is an intuitively more understandable effect size than partial eta-squared.

```
library(relaimpo)

calc.relimp(Model2)
```

```
Response variable: MAT.post.EN
Total response variance: 447.625
Analysis based on 65 observations

2 Regressors:
Condition PA.post.Eng
Proportion of variance explained by model: 15.22%
Metrics are not normalized (rela=FALSE).

Relative importance metrics:

                  lmg
Condition   0.02588599
PA.post.Eng 0.12636111
```

The output giving us the overall multiple R-squared statistic says that the model overall accounts for 15% of the variance in scores on the MAT, and the PA.post.Eng regression entry for lmg shows that the effect of PA accounted for 12.5% of the variance while Condition accounted for only 2.5%. Note that calc.relim( ) cannot be used to calculate relative importance of a model with only one term (such as Model3) as it computes *relative* importance. In this case you could use the summary.lm( ) regression-type output to look at the overall R^2 of the model. For example:

```
summary.lm(Model2)
```

returns this part at the bottom:

```
Residual standard error: 19.79 on 62 degrees of freedom
Multiple R-squared:  0.1522,    Adjusted R-squared:  0.1249
F-statistic: 5.567 on 2 and 62 DF,  p-value: 0.005975
```

which shows that this model explains 15% of the variance (adjusted R^2=12%).

One note here is that the beginning of the output for a one-way ANCOVA in SPSS gives descriptive statistics on the English posttest MAT for scores for just the Comparison group and the Experimental group. We don't get this information automatically included in the R output, but we can call for it with the tapply( ) command as below, which contains the syntax for counts (substitute in mean and sd for those statistics):

```
tapply(LQB$MAT.post.EN, list(Condition= LQB$Condition), function(x) sum(!is.na(x)))
```

The result is that the Comparison group has a lower mean (M=75.0, SD=19.3, N=20) than the Experimental group (M=80.78, SD=21.9, N=45).

We can say more about the effect of PA scores by calling for the mean scores and standard errors of the dependent variable (MAT) adjusted for the effect of the covariate. Use the effects package and the effect( ) command to get this information (note that the package has a plural "s" but the command does not):

```
install.packages("effects")
library(effects)
```

lqb<-effect("Condition", Model2, se=TRUE) #put name of factor you want effects for in first #argument

```
 Condition effect
Condition
  Comparison Experimental
    72.96470     81.68236
```

```
 Lower 95 Percent Confidence Limits
Condition
  Comparison Experimental
    64.02459     75.75683
```

```
 Upper 95 Percent Confidence Limits
Condition
  Comparison Experimental
    81.90481     87.60789
```

summary(lqb)

```
[1] 4.472356 2.964289
```

The effect( ) command gives us the adjusted mean scores for both groups as well as the 95% confidence intervals for this mean. These means are different from those obtained with the tapply( ) command. The **adjusted mean** is the mean score with the influence of the covariate factored out, and in this dataset you can see there is more difference between scores of the Comparison and Experimental groups in these adjusted means than in the original means: now the Comparison group's mean is smaller (M=73.0, SE=4.47, N=20) and the Experimental group's mean is higher (M=81.7, SE=2.96, N=45). These point estimates look fairly different, but if we look at the 95% confidence intervals for the estimated means for the groups, they are quite wide for the Comparison group (almost a 20-point difference, from 64.0 to 81.9). The confidence interval for the mean for the Experimental groups is smaller [75.8, 87.6] but the large and overlapping range means that we are not able to find a difference in effect between groups. As a note, it is not the case that just because confidence intervals overlap this automatically means that there is no difference between groups. Cumming and Finch (2005) give as a "rule of eye" for the interpretation of 2 CIs of two independent groups that they can overlap up to about

half of the average length of both arms of the CI[3] and still be statistical at the .05 level. On the other hand, if the confidence intervals do not overlap, this is evidence that there is a strong difference between groups. Anyway, my point is that the estimated adjusted means gives you a point estimate, but the CIs show us these point estimates are not very precise and we cannot be very confident of where the true means lie.

After you have considered your ANCOVA, it might turn out that you will need to conduct pairwise comparisons on the variables. In this case, Condition is not statistical, and even if it were, since there are only two groups we would just look at the mean scores to see which group did better than the other. But in a case where we want to conduct pairwise comparison with the effect of the pretest factored out, use the glht( ) command from the multcomp package I used in Chapter 9, but use it on the ANCOVA model, like this (change the parts in red for your own data):

```
library(multcomp)
summary(glht(Model2, linfct=mcp(Condition="Tukey")))
```

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = MAT.post.EN ~ Condition + PA.post.Eng, data = LQB)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
Experimental - Comparison == 0    8.718      5.400   1.614    0.112
(Adjusted p values reported -- single-step method)
```

[3] Precisely, this average length of both of the arms of the CI is called the margin of error, and for this example would be [(81.9-64.0)/2 + (87.6-75.8)/2]/2= 7.4, meaning the CIs could overlap by about 3.5 points (half of the average length of both arms).

To get confidence intervals for your comparisons, just use confint( ):

```
confint(glht(Model2, linfct=mcp(Condition="Tukey")))
```

```
Linear Hypotheses:
                                Estimate lwr     upr
Experimental - Comparison == 0  8.7177  -2.0766 19.5119
```

Assumptions for ANOVA can be checked in the normal way:

```
plot(Model2)
```

Section 7.4.6 of the book covered these plots in more detail, but basically you will see a Residuals vs. Leverage plot (which should not show any points beyond the Cook's Distance dashed lines), Residuals vs. Fitted plot (which should show a random scattering of points), the Normal Q-Q plot (where points should cluster around the line), and the Scale-Location Diagnostic plot, which should also show a random scattering of data. Although the plot( ) command has called on the computer to label the 3 most extreme scores so that it may look like there are outliers, the results show that the residuals basically conform to the assumptions of normality and homogeneity.

Performing a One-way ANCOVA with One Covariate in R

1. Model your ANOVA with the aov( ) command (lm( ) can also be used but a summary will result in regression output, not ANOVA tables). Add your covariate to your equation; you can test out a maximal model by first having the covariate enter into an interaction with the IV (although you hope there is no statistical interaction or you will be violating the assumption of ANCOVA that the covariate has an independent effect on the dependent variable). Here is an example of a maximal model where scores on the English Morphology Awareness task (measure on a posttest) are modeled according to the experimental Condition that the subjects were in with a covariate of scores on Phonological Awareness in a posttest (N.B. items in red should be replaced with your own data name):
   Model1=aov(MAT.post.EN~Condition*PA.post.Eng, data=LQB)
2. Examine output using the anova( ) command. Find the best model by using the update( ) function and taking out model terms one by one. For example, here is the next model that takes out the interaction between Condition and the PA posttest:
   Model2=update(Model1,~.-Condition:PA.post.Eng)
3. Use the anova( ) command to compare models to get to a model that has only statistical terms. Or you may choose to report on the model which includes both the IV and the covariate, whether or not they are statistical.
4. Whether you pay attention to the statistical significance of the covariate depends on your question. You will, however, want to note whether your variable of interest (the IV) is statistical even when the effects of the covariate are factored out.
5. Use tapply( ) to get basic descriptive statistics for your data, and the effect( ) command from the effects package to get adjusted means, standard errors, and confidence intervals for your adjusted data.
6. If you would like to perform post-hoc comparisons use the regression model in the glht( ) procedure from the multcomp package to compare adjusted means.

Performing a Two-way ANCOVA with One Covariate in SPSS

This section assumes you have already read the sections of this document called "Performing a One-way ANCOVA with One Covariate in SPSS" and "ANCOVA Output in SPSS," describing how to call for a one-way ANCOVA with one covariate in SPSS. Here I will work to recreate the

analysis done by Lyster, Quiroga and Ballinger (2013) and shown in Figure 8 that has the English MAT posttest scores as the DV, Condition and Language as the IVs, and the English MAT pretest score as the covariate.

Follow the same directions for calling for the ANCOVA as were given in the section "Performing a One-way ANCOVA with One Covariate in SPSS," but additionally add Language to the "Fixed Factor(s)" box, and change the covariate from PA.post.Eng to MAT.pre.EN. Open the MODEL button and test out the CUSTOM MODEL with an interaction between Condition and MAT.pre.EN and another between Condition and MAT.pre.EN. Press CONTINUE and run the ANCOVA (perhaps check off Bootstrap to make the analysis run faster while you are checking the assumptions right now). Look for the box labeled "Tests of Between-Subjects Effects" and you will see that the *p*-values for the interactions are well above $p = .05$, so we may assume there are no interactions between the covariate and the IVs.

Open up the same analysis and change the Model back to full factorial. Open the Options button and since Language has 3 levels, move it over to the right (under "Display Means for") as well as the interaction between the IVs (take the main effect of Condition out if it is still there from the one-way ANCOVA analysis). Tick on bootstrapping again if you unticked it for the test of assumptions.

What is noteworthy from the output is that neither the effects of experimental Condition nor Language dominance are statistical influences on the posttest English MAT, nor does the interaction between these two variables fall under the $p = .05$ level (the relevant table of output is

found in Table 10). The covariate of the MAT pretest, however, is statistical, indicating that there is an influence of pretest scores on the variability of posttest scores. Overall, this model accounted for R^2 = 44% of the variance in the model (39% for adjusted R^2).

Table 10 Main results for testing a two-way ANCOVA with one covariate in SPSS.

Tests of Between-Subjects Effects

Dependent Variable: MAT-post-EN

Source	Type II Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	12727.062[a]	6	2121.177	7.727	.000	.444
Intercept	4763.245	1	4763.245	17.353	.000	.230
MATpreEN	7263.104	1	7263.104	26.459	.000	.313
Conditon	33.468	1	33.468	.122	.728	.002
Language	280.850	2	140.425	.512	.602	.017
Conditon * Language	1596.445	2	798.223	2.908	.063	.091
Error	15920.938	58	274.499			
Total	434313.000	65				
Corrected Total	28648.000	64				

a. R Squared = .444 (Adjusted R Squared = .387)

Notice that order matters in an ANOVA analysis! I have Condition first, then Language. If you put in the factor of Language first, you don't get an overall different result, but the *p*-values are different as the order changes the sum of squares, which affects the calculation of *p*-values.

Here is how you could report on the results of this analysis:

A two-way ANCOVA on the English MAT posttest revealed no interaction between group and language dominance ($F_{2,58}$ = 2.91, *p* = .063, partial eta-squared = .09), although the effect size suggests that there is some effect for the interaction in spite of the fact that the *p*-value is below .05. No main effects were found for Language ($F_{2,58}$ = .51, p=.60, partial eta-squared = .02) or

Condition ($F_{1,58}$ = .12, p = .73, partial eta-squared = .002). There was a statistical effect for the covariate of English MAT pretest, however ($F_{1,58}$ = 26.5, p<.0005, partial eta-squared = .31), indicating that pretest scores did have a strong effect on posttest scores. Overall, this model accounted for R^2=44% of the variance in the model (39% for adjusted R^2).

By the way, Lyster, Quiroga and Ballinger (2013) reported that by removing one outlier from their analysis, they found a statistical interaction between experimental Condition and Language dominance, with further tests showing that there was a difference between Experimental and Comparison students who were English-dominant. This shows that this dataset might be a good candidate for a robust ANCOVA where outliers would be removed systematically and without compromising independence (read on for that!).

Performing a Two-way ANCOVA with One Covariate in R

This section assumes you have already read the section "Performing a one-way ANCOVA with one covariate in R" describing how to call for a one-way ANCOVA with one covariate in R. Here I will analyze the two-way ANCOVA used by Lyster, Quiroga and Ballinger (2013) and shown in Figure 8 that has the English MAT posttest scores as the DV, Condition and Language as the IVs, and the English MAT pretest score as the covariate.

In the section "Performing a one-way ANCOVA with one covariate in R" the maximal model was:

```
Model1= aov(MAT.post.EN~Condition*PA.post.Eng, data=LQB)
```

To model two IVs, just add the second IV in (because order matters, I will model it after Condition):

```
ModelA= aov(MAT.post.EN~Condition*Language*MAT.pre.EN, data=LQB)
```

A summary( ) of this model reveals that only Language and the covariate of MAT pretest are statistical. That's good as it means we do not violate the assumption of ANCOVA that the covariates are independent influences on the DV. Therefore, use update( ) to remove interaction terms, and anova( ) to compare the differences between models.

```
ModelB= update(ModelA,~.-Condition:Language:MAT.pre.EN, data=LQB)
ModelC= update(ModelB,~.-Condition: MAT.pre.EN, data=LQB)
ModelD= update(ModelC,~.- Language:MAT.pre.EN, data=LQB)
ModelE= update(ModelD,~.- Condition:Language, data=LQB)
anova(ModelA, ModelB, ModelC, ModelD, ModelE)
```

The ANOVA results show no statistical differences between models (the last change from Model D to Model E has a low *p*-value of p = .07, however) so we will keep Model E, which is the simplest model so far. An ANOVA summary of Model E shows the results:

```
> summary(ModelE)
            Df Sum Sq Mean Sq F value  Pr(>F)
Condition    1    462     462   1.583 0.21318
Language     2   3496    1748   5.988 0.00426 **
MAT.pre.EN   1   7172    7172  24.565 6.2e-06 ***
Residuals   60  17517     292
```

The summary shows this is not the minimal adequate model yet though, since there is one non-statistical term (Condition). So just as in the section called "Performing a One-way ANCOVA with One Covariate in R" we might report on this model or take the analysis one last step to delete the non-statistical term of Condition.

```
ModelF= update(ModelE,~.- Condition, data=LQB)
```

As in that previous section with just one covariate, a regression-type summary (summary.lm( )) can get you the R^2 effect size for the entire model that you decide to keep. A tapply( ) command can get descriptive statistics, and the effect( ) command (from the effects package) can call for adjusted means.

For effect sizes, use the relaimpo package:

```
calc.relimp(ModelF)
```

```
Response variable: MAT.post.EN
Total response variance: 447.625
Analysis based on 65 observations

3 Regressors:
Some regressors combined in groups:
        Group  Language : Language[T.French dominant] Language[T.Bilingual]

 Relative importance of 2 (groups of) regressors assessed:
 Language MAT.pre.EN

Proportion of variance explained by model: 38.74%
Metrics are not normalized (rela=FALSE).

Relative importance metrics:

                  lmg
Language   0.05716336
MAT.pre.EN 0.33019874
```

This output shows that the variable of MAT pretest is relatively much more important (explaining 33% of the 39% of the variance that the total model explains) than the Language variable, which explains only 6%. You can always call for partial eta-squared too:

```
> etasq(ModelF)
           Partial eta^2
Language       0.01397256
MAT.pre.EN     0.31499448
Residuals              NA
```

Because the factor of Language dominance is statistical and has more than two levels, let's try the glht( ) command on ModelE (from the multcomp package), specifying that we want comparisons on the Language variable:

```
summary(glht(ModelE, linfct=mcp(Language="Tukey")))

confint(glht(ModelE, linfct=mcp(Language="Tukey")))
```

The results show no differences between groups. The difference between the French-dominant and English-dominant group is 5.5, 95% CI [-19.8, 8.7], the difference between the French-dominant and Bilingual group is 3.0, CI [-9.6, 15.5], and the difference between the English-dominant and Bilingual group is -2.5, CI [-15.2, 10.1]. This phenomenon, where the omnibus test is statistical but the individual tests is not, is known to occur (Wilcox, 2011; Fairley, 1986) and could be a result of having the bulk of the joint data concentrated in a very narrow region.

Performing a Two-way ANCOVA with Two Covariates in SPSS

This section assumes you have already read the sections of this paper that describe how to call for a one-way ANCOVA with one covariate in SPSS. To look at what might be different when we have two covariates along with two IVs in an ANCOVA, use the SPSS file LarsonHall2008.sav. In this study I looked at whether exposure in childhood to formal English lessons resulted in Japanese college students being able to perform better on a test of the English R/L/W contrast and on a grammaticality judgment test in English. Since Japanese learners of English who began studying English at a younger age might have more hours of exposure, I wanted to use this variable as a covariate. In other words, I wanted to factor this out of the comparison between the group that had had early exposure and the group that hadn't. I also thought it was possible language aptitude might be involved, so I measured that in order to statistically take that factor out of the equation.

We'll look at the question of scores on the grammaticality judgment test. My research question was whether the group that had early exposure differs from the group which did not, and I named this factor ErlyExp (for Early Exposure). To create an analysis with two IVs, I added the factor of Sex as well.

First, I checked additional assumptions for ANCOVA that there was not a strong correlation between covariates. I checked the correlation between the two variables of language Aptitude and total amount of input in English (Totalhrs) by simply following the menu ANALYZE > CORRELATE > BIVARIATE and choosing those two variables. The correlation between these two covariates is $r = 0.08$. This is not cause for worry at all.

I also wanted to check to see that there are strong correlations between the DV and the covariates, so I checked the correlation of the DV (GJTscore) with Totalhrs and Aptitude. There is a weak correlation between GJTscore and Aptitude (r = .08) and a stronger correlation between GJTscore and Totalhrs (r=.18). This is not ideal for an ANCOVA analysis, as I'd like to have stronger correlations between my covariates and my DV, but to illustrate the analysis I will continue.

Follow the same directions for calling for the ANCOVA as were given in previous SPSS sections of this chapter. For this analysis, put GJTscore in the "Dependent Variable" box. Put Sex and ErlyExp in the "Fixed Factor(s)" box, and put Totalhrs (amount of input) and Aptscore (aptitude test) into the "Covariate(s)" box. Open the MODEL button and test out the CUSTOM MODEL with 4 interactions: all the two-way combinations between Aptscore and the IVs (Sex and ErlyExp) and all the two-way combinations between Totalhrs and the IVs. Press CONTINUE and run the ANCOVA (perhaps check off Bootstrap to make the analysis run faster while you are checking the assumptions right now). Look for the box labeled "Tests of Between-Subjects Effects" and you will see that the *p*-values for the interactions are above p=.05, so we may

assume there are no interactions between the covariate and the IVs.

Open up the same analysis and change the Model back to full factorial. Since Sex and ErlyExp only have 2 levels each, there is no need to move them over to call for comparisons. Tick on bootstrapping again if you unticked it for the test of assumptions. The full factorial model in SPSS will call for any interactions between IVs but not interactions between covariates and IVs (if it did, that would make this a four-way ANOVA, which would be quite complicated!). The main results are found in Table 11.

Table 11 Main results for testing a two-way ANCOVA with two covariates in SPSS.

Tests of Between-Subjects Effects

Dependent Variable: gjtscore

Source	Type II Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1665.998[a]	5	333.200	3.316	.007
Intercept	36910.253	1	36910.253	367.326	.000
aptscore	47.849	1	47.849	.476	.491
totalhrs	412.677	1	412.677	4.107	.044
sex	687.874	1	687.874	6.846	.010
erlyexp	44.609	1	44.609	.444	.506
sex * erlyexp	.606	1	.606	.006	.938
Error	19493.822	194	100.484		
Total	2576316.000	200			
Corrected Total	21159.820	199			

a. R Squared = .079 (Adjusted R Squared = .055)

One could summarize the results as follows:

A two-way ANCOVA showed that there was no effect for the interaction between sex (male vs. female) and early experience ($F_{1,194} = .006$, $p = .94$, partial eta-squared = .00) nor for the simple main effect of early experience ($F_{1,194} = .44$, $p = .51$, partial eta-squared = .002), but there was a

very small main effect for sex (gender) ($F_{1,194} = 6.8$, $p = .01$, partial eta-squared = .03). The covariate of total hours of input was statistical ($F_{1,194} = 4.1$, $p = .04$, partial eta-squared = .02), although again with a small effect size. Note that a statistical effect for this variable means scores would be adjusted to take the effect of hours of study into account. The effect of the covariate of aptitude was not statistical ($F_{1,194} = .48$, $p = .49$, partial eta-squared = .002), meaning there was not enough influence from this variable to affect adjusted scores. Overall, the model accounted for very little of the variance in GJT scores ($R^2 = 8\%$, adjusted $R^2 = 6\%$).

Performing a Two-way ANCOVA with Two Covariates in R

This section assumes you have already read the previous section of this paper describing how to call for a one-way ANCOVA with one covariate in R. To look at what might be different when we have two covariates along with two IVs in an ANCOVA, use the SPSS file LarsonHall2008.sav, imported into R as larsonhall2008. In this study I looked at whether exposure in childhood to formal English lessons resulted in Japanese college students being able to perform better on a test of the English R/L/W contrast and on a grammaticality judgment test in English. Since Japanese learners of English who began studying English at a younger age might have more hours of exposure, I wanted to use this variable as a covariate. In other words, I wanted to factor this out of the comparison between the group that had had early exposure and the group that hadn't. I also thought it was possible language aptitude might be involved, so I measured that in order to statistically take that factor out of the equation.

We'll look at the question of scores on the grammaticality judgment test. My research question was whether the group that had had early exposure differs from the group that did not, and I

named this factor ErlyExp (for Early Exposure). To create an analysis with two IVs, I added the factor of Sex as well.

First, I checked additional assumptions for ANCOVA that there was not a strong correlation between covariates. I checked the correlation between the two variables of language aptitude (aptscore) and total amount of input in English (totalhrs) by simply following the menu STATISTICS > SUMMARIES > CORRELATION MATRIX in R Commander and choosing those two variables. The correlation between these two covariates is $r = 0.08$. This is not cause for worry at all.

I also wanted to check to see that there are strong correlations between the DV and the covariates, so I checked the correlation of the DV (gjtscore) with totalhrs and Aptitude (aptscore). There is a weak correlation between GJTscore and Aptitude ($r = .08$) and a stronger correlation between GJTscore and Totalhrs ($r = .18$). This is not ideal for an ANCOVA analysis, as I'd like to have stronger correlations between my covariates and my DV, but to illustrate the analysis I will continue.

For the assumption that the regression slopes are equal we can test for the presence of an interaction between the covariates and the grouping variable (here, erlyexp and sex). We'll test for an interaction between each covariate (aptscore and totalhrs) and the grouping variable one at a time.

```
summary(aov(gjtscore~erlyexp*aptscore, data=larsonhall2008))
```

```
summary(aov(gjtscore~erlyexp* totalhrs, data=larsonhall2008))
summary(aov(gjtscore~sex* aptscore, data=larsonhall2008))
summary(aov(gjtscore~sex* totalhrs, data=larsonhall2008))
```

In no case was the interaction statistical, so I may proceed with the ANCOVA analysis. In starting with the maximal model, I am going to start with interactions between the IVs but not between the covariates and anything else. This is because the model will be so complex with 4 terms that I just don't want to do it. I have already tested to see if there are interactions between each covariate and the IV, so I won't get more complex than that. So here is my first model:

```
LH1=aov(gjtscore~ aptscore + totalhrs + sex*erlyexp, data=larsonhall2008)
```

```
summary(LH1)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
aptscore      1    132   131.8   1.311 0.25355
totalhrs      1    678   677.6   6.743 0.01013 *
sex           1    811   811.5   8.076 0.00497 **
erlyexp       1     45    44.6   0.444 0.50602
sex:erlyexp   1      1     0.6   0.006 0.93816
Residuals   194  19494   100.5
```

The summary shows that there is a statistical main effect for the IV of sex but not for the IV of early exposure, at least when the effects of aptitude and hours of exposure are factored out. We could stop at this step, as this is exactly the model that SPSS automatically creates, but to get to the minimal adequate model, let's try removing the interaction between the two IVs, as it is not statistical.

```
LH2=update(LH1,~.-erlyexp:sex, data=larsonhall2008)
summary(LH2)
```

The summary shows that nothing has changed—the early exposure variable is still not statistical, nor is the covariate of aptitude score. Shall we keep this simpler model?

```
anova(LH1, LH2)
```

```
Model 1: gjtscore ~ erlyexp * sex + totalhrs + aptscore
Model 2: gjtscore ~ erlyexp + sex + totalhrs + aptscore
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1    194 19494
2    195 19494 -1  -0.60642 0.006 0.9382
```

Yes. There is no difference between the models so we will keep the simpler one. At this point I don't want to eliminate the variable of early exposure, however, because my main question is whether this variable is statistical when the effect of aptitude and input is factored out. My analysis has answered that question—it is not.

To check effect sizes I like the relaimpo package:

```
calc.relimp(LH2)
```

```
Response variable: gjtscore
Total response variance: 106.3308
Analysis based on 200 observations

4 Regressors:
erlyexp sex totalhrs aptscore
Proportion of variance explained by model: 7.87%
Metrics are not normalized (rela=FALSE).

Relative importance metrics:

                lmg
erlyexp  0.006462423
sex      0.042156835
totalhrs 0.026105822
aptscore 0.003980297
```

The output shows that only 8% of the variance in the data is accounted for by this four-term model, which is not very much. Of the 8%, the most important factor is sex, accounting for 4% of the variance. Next comes total hours of input, accounting for 3% of the variance, and then Early Experience and Aptitude; both account for only about ½% each. I could also get the partial eta-squared effect sizes:

```
> etasq(LH2)
          Partial eta^2
erlyexp     0.002283071
sex         0.034083030
totalhrs    0.020711780
aptscore    0.002417634
Residuals            NA
```

I could write up my results something like this:

A two-way ANCOVA showed that there was no effect for the interaction between sex (male vs. female) and early experience ($F_{1,194} = .006$, $p = .94$) so I removed it, resulting in a model with

only 4 main effects. In that model the main effect of early experience was not statistical ($F_{1,195}$ = 2.51, p = .11), but there was a very small main effect for sex (gender), ($F_{1,195}$ = 9.37, p = .002). The covariate of total hours of input was statistical ($F_{1,195}$=4.30, p=.04), although again with a small effect size. This meant scores would be adjusted to take the effect of hours of study into account. The effect of the covariate of aptitude was not statistical ($F_{1,195}$=.47, p=.49), meaning there was not enough influence from this variable to affect adjusted scores. Overall, the model accounted for very little of the variance in GJT scores (R^2=8%, adjusted R^2=6%), with sex accounting for 4% of that variance, hours of input accounting for 3%, and Early Experience and Aptitude both accounting for only about ½% each.

Performing a Robust ANCOVA in R

As we have seen throughout the book, there are many techniques that can be used to robustly analyze data. Bootstrapping is one technique that we have used, and SPSS provides for bootstrapping in the ANCOVA dialogue box, but this bootstrapping is only performed on the mean scores and adjusted mean scores ("Estimated marginal means") and then pairwise comparisons between the different levels of one independent variable. The R functions examined here did not provide bootstrapping, but the online document in Chapter 9 called "Bootstrapped One-Way ANOVA in R" provides information about performing robust one-way ANOVA analysis and you could use that method, although it does not provide adjustments for the covariate.

Another technique that is called robust for ANCOVA is an approach where we eliminate the assumption of parametric models that the regression lines of all groups are parallel (Wilcox, 2005). The method I will present in this section does not make any assumptions about the

linearity of regression lines, allows heteroscedasticity, and performs well even if data is non-normal. The general idea of the method is to use a Loess smoother and then make comparisons between groups at specific points along the smooth line. This method also trims the data (20% is recommended) and bootstraps it with the recommended bootstrap-t. If you don't want to trim data or bootstrap you can still use the function but specify that the values for these arguments should be zero. The function will pick 5 arbitrary points at first but a researcher can specify the points themselves as well, and will probably want to do so after looking at scatterplots with smooth lines that will result from the output.

Wilcox's (2005) function ancboot( ) from the WRS package can only be used with one covariate, and it can only examine two groups at a time, and your dataset as is will have to be massaged in order to use it. This function has been shown to have good power when ANCOVA assumptions are not met, but another advantage, Wilcox says, is that "[e]ven when ANCOVA has relatively good power, an advantage of using trimmed means with a running-interval smoother is that the goal is to determine where the regression lines differ and by how much, and this is done without assuming that the regression lines are straight" (2005, p. 526).

To use the ancboot( ) function for Lyster, Quiroga and Ballinger's (2013) one-way ANCOVA, first the data must be subsetted into separate dataframes for each group as shown below. Use the LQB.sav file imported into R as "LQB." To carry out the one-way ANCOVA on the English MAT, let's subset the data into groups for Condition only:

```
LQB.Exp=subset(LQB,subset=Condition=="Experimental")
```

```
LQB.Compare=subset(LQB,subset=Condition=="Comparison")
```

Now we will set up the data into separate variables, one for each group in the DV, and one for each group on the covariate as well:

```
x1=LQB.Exp$MAT.post.EN
y1=LQB.Exp$PA.post.Eng
x2= LQB.Compare$MAT.post.EN
y2= LQB.Compare$PA.post.Eng
```

Now open the package and try the basic command:

```
library(WRS)
ancboot(x1,y1,x2,y2)
```

Here is an analysis of this command:

```
ancboot(x1,y1,x2,y2,fr1=1,fr2=1,tr=.2, nboot=599, plotit=T, pts=NA)
```

ancboot (x1, y1, x2, y2)	Performs an ANCOVA using trimming and a percentile bootstrap method; the data for group 1 are stored in x1 and y1, and the data for group 2 are stored in x2 and y2. I have put the data for the DV in X and the data for the covariate in Y.

fr1=1, fr2=1	The values of the span (f) for the groups, defaults to 1; span refers to how frequently data are sampled by the running-interval smoother, meaning how smooth the smooth line looks
tr=.2	Specifies amount of trimming, default is 20%
nboot=599	Specifies number of bootstrap samples to use; 599 is default
plotit=T	Plots a scatterplot with Loess smoothers drawn in
pts=NA	If not specified, the function will automatically choose 5 points at which to compare groups. The points can be specified by the user, such as pts = c(1,3,5)

The output returns the points where data are compared, confidence intervals for the test that trimmed means of the groups are equivalent at that point, and a *p*-values for the test. Wilcox says that the function determines "a critical value based on the Studentized maximum modulus distribution" (2005, p. 527). The plot that is returned is also very useful. Here is the text output from the command run above:

```
[1] "Note: confidence intervals are adjusted to control FWE"
[1] "But p-values are not adjusted to control FWE"
[1] "Taking bootstrap samples. Please wait."
$output
      X n1 n2       DIF      TEST    ci.low     ci.hi    p.value
[1,] 60 22 12 -2.589286 -0.871904 -11.06941 5.8908410 0.37729549
[2,] 69 26 12 -3.875000 -1.379655 -11.89533 4.1453300 0.18196995
[3,] 78 28 13 -4.222222 -1.626759 -11.63377 3.1893266 0.12020033
[4,] 81 28 13 -5.388889 -2.464565 -11.63270 0.8549263 0.02337229
[5,] 90 28 12 -5.291667 -2.251667 -12.00255 1.4192124 0.03171953

$crit
[1] 2.855558
```

The results show that the Experimental and Comparison groups were tested at five points on the MAT posttest (we set it up so that the MAT was on the x-axis and the PA on the y-axis): 60, 69, 78, 81, and 90 (we know this because they are listed in the column with "X"). The n1 and n2 specify how many subjects were tested at each of those points. Notice that n1, the Experimental group, is always larger than n2, the Comparison group. If we decide we want to test at more intervals, we will find that the number of participants in those points changes.

The "DIF" column gives the difference in scores at that point, and the "TEST" gives the test statistic for the hypothesis that the mean of one group is equal to the mean of the other at that point. The "p.value" column gives the *p*-value for the test. The "ci.low" and "ci.hi" columns give the lower and upper ends of the 95% confidence interval for the difference between the groups. The last part of the text output lists the critical value used to evaluate whether the test is statistical. If the value in the "TEST" column is *below* this critical value, the test will not be statistical. Wilcox notes that this critical value will be adjusted if fewer than 5 points appropriate for testing can be found. I'm not sure why the statisticality of the test differs here between the CIs (where it passes through zero for point 81 and as such should not be statistical) and the *p*-

value (which is below .05 for point 81 and should thus be statistical, although the test statistic is lower than the critical value, which means it should not be statistical!).

The plot that is called for is extremely helpful in understanding what is happening, and shows the points with the smooth lines on them (Figure 13).

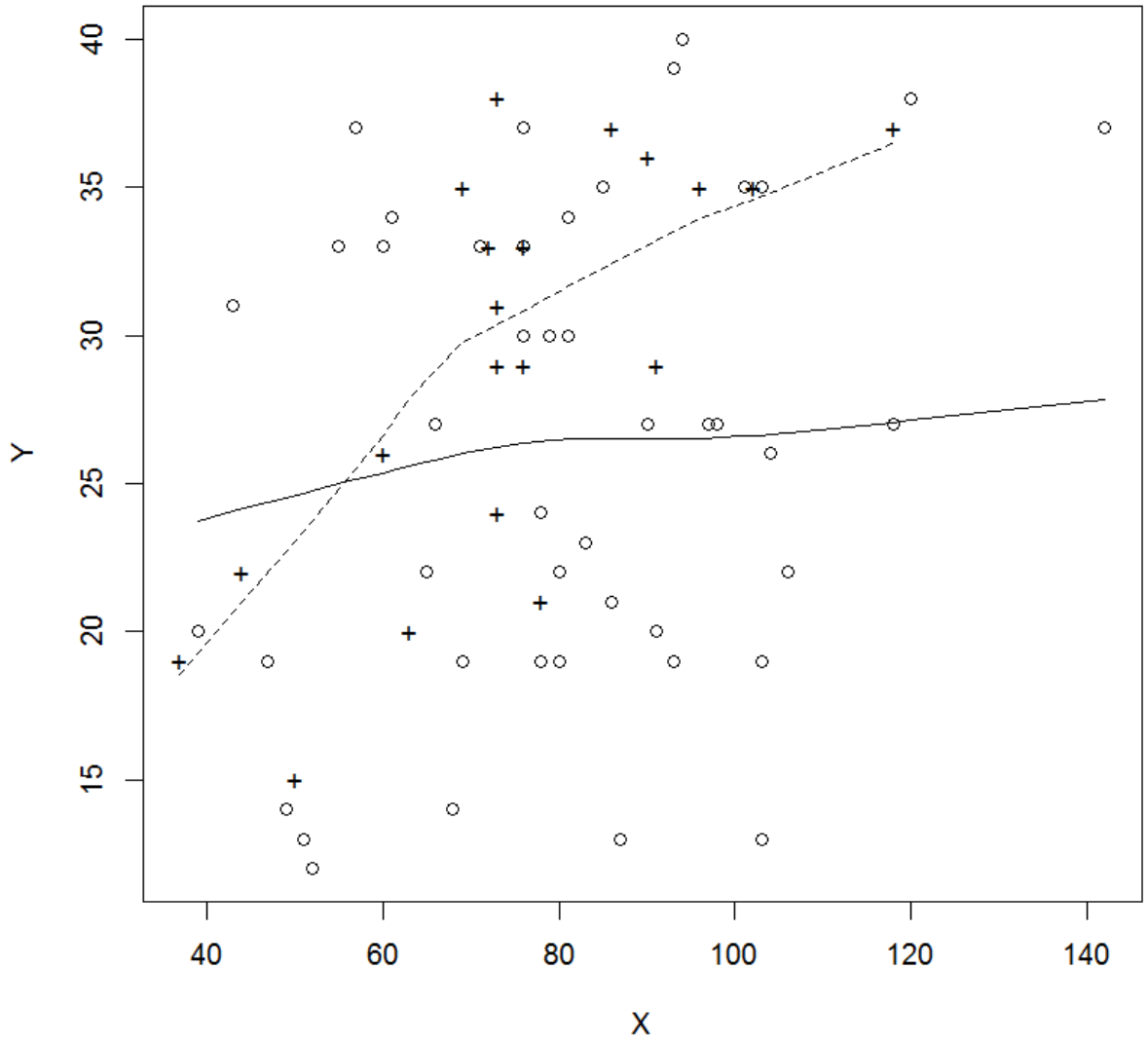

Figure 13 Output from Wilcox's ancboot function showing Loess smooth lines for robust ANCOVA comparison .

This scatterplot does not come with a legend, but Wilcox (2012) writes that Group 2 is indicated with a "+" and the dashed line for a smoother result, so we know that the open dots and the smooth Loess line represent Group 1, the Experimental group, and the plus signs and dashed

Loess represent Group 2, the Comparison group. The confidence intervals indicated no differences at any point, while the *p*-values for the test indicated that there were statistical differences at 81 and 90. The relationship between the Comparison group and PA scores is quite strong, indicating that those who scored higher on the MAT posttest were those who had higher PA scores on the posttest, while the Loess line for the Experimental group shows that PA was not very important to higher scores on the MAT. This makes sense logically, as the Comparison group did not get the same explanation that the Experimental group got, so those participants who were already high in analytical abilities (as evidenced by higher PA scores) performed better on the MAT (at least, at points 81 and 90).

We might be interested to look at other points along the range of scores, although we will probably find the numbers are not so evenly divided as they were for the points the computer chose. But let's say we decide to look at differences between the groups at points 45, 65, 85, and 105. We would just enter these into the pts=c( ) argument like this:

```
ancboot(x1,y1,x2,y2, pts=c(45, 65, 85, 105))
```

```
$output
       X n1 n2       DIF       TEST       se    ci.low      ci.hi     p.value
[1,]  45 11  4  4.071429  0.8582706 4.743759 -11.76866 19.9115192 0.405676127
[2,]  65 25 12 -1.783333 -0.5827985 3.059949 -12.00094  8.4342740 0.537562604
[3,]  85 31 14 -6.500000 -3.0542831 2.128159 -13.60623  0.6062279 0.005008347
[4,] 105 18  5 -8.583333 -3.5380208 2.426027 -16.68418 -0.4824819 0.005008347

$crit
[1] 3.339143
```

The program made the comparisons at the points I chose, but notice that the critical value got larger. Still there is apparently a difference between the groups not only at 85 but also at 105, farther out in the data. If you ask me, this type of information is much more interesting than the results of the parametric ANCOVA, where we found that "the participants did not differ in their performance by the experimental group they belonged to, when scores were adjusted for posttest Phonological Awareness, ($F_{1,62} = 1.18$, $p = .28$)." The parametric analysis forced us to consider parallel regression lines, and found they were not different enough in intercept to say that the groups were different. The non-parametric ANCOVA lets us look at the way the data actually runs, and get a better sense of what is happening, which is an interaction between the DV and the covariate at higher levels of scores.

Tip: Sometimes when you run this program you may get an error message like this:

```
Warning in min(sub[vecn >= 12]) :
  no non-missing arguments to min; returning Inf
Warning in max(sub[vecn >= 12]) :
  no non-missing arguments to max; returning -Inf
Warning in near(x1, x1[isub[i]], fr1) : NAs introduced by coercion
Warning in near(x2, x1[isub[i]], fr2) : NAs introduced by coercion
Error in var(y) : 'x' is empty
```

I found the solution to this error was to pick a smaller number of points to test at where I had a larger number of participants. Wilcox says "the points among the covariates at which the groups will be compared are determined by the function; it finds a point among the x1 values that has the deepest halfspace depth, plus the points on the .5 depth contour, and the groups are compared at these points, provided that the corresponding sample sizes are at least 10" (2005, p. 533), so this error appears to generate when there are too few participants to test at a certain point.

Now let's see what happens when we want to do a two-way ANCOVA with one covariate, as in Lyster, Quiroga and Ballinger's (2013) two-way ANCOVA with the English MAT posttest as

the DV, the English MAT pretest as the covariate, and both Condition and Language dominance as the IVs. We can use both of these IVs by simply splitting the data by both of the variables, although in some cases this will result in very small samples. We have files already split by Condition, so now let's split those by Language dominance:

```
LQB.Exp.EN=subset(LQB.Exp,subset=Language=="English dominant")
LQB.Exp.FR=subset(LQB.Exp,subset=Language=="French dominant")
LQB.Exp.BI=subset(LQB.Exp,subset=Language=="Bilingual")
```

Do the same with the LQB.Compare file. Now we can only compare two groups at a time, so the following code shows how we could compare the English-dominant Experimental group with the French-dominant Experimental group.

```
x1=LQB.Exp.EN$MAT.post.EN
y1= LQB.Exp.EN$MAT.pre.EN
x2= LQB.Exp.FR$MAT.post.EN
y2= LQB.Exp.FR$MAT.pre.EN
ancboot(x1,y1,x2,y2, tr=0, pts=c(60)) #I changed trimming to zero since there are so few
#variables, and tried to just choose one point
```

I kept getting a warning error about the degrees of freedom. I could not get this analysis to work, so my suspicion is that there were just too few participants in each area. With larger numbers in each group this would probably work, but because only two groups can be compared at one time

on two variables, this robust ANCOVA is more limited than the parametric version, but I think that when it can work, it is an informative function. For another example of the robust ANCOVA, see my results in Larson-Hall (2008).

Performing a Robust ANCOVA

1 Wilcox (2012) lists several possible commands for performing robust ANCOVAs. In this section I focus on just one, ancboot( ), which does not assume that groups have parallel regression lines, uses means trimming and also a bootstrap-t procedure.
2 Data need to be arranged so that data for only one group are in a data frame. The subset( ) command is useful for this.
3 The basic ANCOVA command is:
ancboot(x1,y1,x2,y2,fr1=1,fr2=1,tr=.2, nboot=599, plotit=T, pts=NA)
where x1, y1, x2 and y2 must have your data for your two groups for the DV and the covariate. The rest of the specifications are default but can be changed.
4 Errors might arise if there are too few data points at specific points of comparison, but you can try to control this via the pts=c( ) term of the command. A scatterplot called from this command helps show visually what points of comparison are of interest.

Application Activities for ANCOVA for R (No Answers Given)

1 Class Time. Use the dataset I have called ClassTime.sav (import into R as "classtime"; this dataset was taken from Howell (2002, p. 629), but I adapted it to reflect a design that will be associated with the second language research field). Let's pretend that a researcher who is in charge of teaching Arabic at the university level notices that there seems to be a difference in how students in her 8 a.m. class respond to her teaching versus how students in the later classes respond. At the start of a new school year she gives them an initial test of their enthusiasm and motivation for learning Arabic. There are 30 items that contain a ten-point Likert scale, where a higher score is more positive about the class. The researcher averages their answers together for a score out of 10. She then administers the same test at the end of the semester. The researcher has five classes,

one at 8 a.m., one at 10 a.m., one at 11 a.m., one at 1 p.m., and one at 2 p.m. This study could be analyzed with an RM ANOVA (if the data were arranged in the "wide" format) but the researcher decides to analyze it with an ANCOVA using the pretest scores as a covariate so that any differences among the posttest scores due to variability in pretest scores will be controlled. Use PreTestScores as the covariate, PostTestScores as the dependent variable, and TimeOfClass as the independent variable. First check the special assumptions for ANCOVA. Even if the data violates the assumptions, go ahead and perform the ANCOVA. What are the results of the parametric ANCOVA?

2 Lyster, Quiroga and Ballinger (2013). Use the data for the French posttest MAT as the DV and run a one-way ANCOVA with Condition as the IV. First check the special assumptions for ANCOVA. Even if the data violates the assumptions, go ahead and perform the ANCOVA. What are the results of the parametric ANCOVA?

3 Using the same data as in #2, run a robust ANCOVA for the French posttest MAT. Explain the results of your tests and the scatterplot. Compare to the robust one-way ANCOVA for the English posttest MAT as explained in the section of this chapter called "Performing a robust ANCOVA in R"—are results similar?

4 Lyster, Quiroga and Ballinger (2013). Use the data for the French posttest MAT as the DV and run a two-way ANCOVA with Condition and Language as the IVs. First check the special assumptions for ANCOVA. Even if the data violates the assumptions, go ahead and perform the ANCOVA. What are the results of the parametric ANCOVA?

5 The first edition of the book featured data from Lyster's (2004) object identification task. Lyster (2004) investigated the question of whether conditions involving the provision of form-focused instruction had differing effects on the post-task results of participants

taking an object identification task. Because groups were found to differ statistically on the pretest, the pretest was used as a covariate in an ANCOVA analysis. Use the Lyster.Oral.sav file (import as lysterO into R) and run a one-way ANCOVA with PostObjectID as the DV, Condition as the IV, and PreObjectID as the covariate. First check the special assumptions for ANCOVA. Even if the data violates the assumptions, go ahead and perform the ANCOVA. What are the results of the parametric ANCOVA?

6 Using the same data as in #5, run a robust ANCOVA for the PostObjectID. Explain the results of your tests and the scatterplot. Compare to the results of the parametric one-way ANCOVA in #5.

7 Larson-Hall (2008). Use the SPSS file LarsonHall2008.sav (import into R as larsonhall2008). In previous sections of this document you saw an analysis with two covariates (aptscore and totalhrs) with the dependent variable of grammaticality judgment test scores. Perform the same analysis using the dependent variable of the phonemic discrimination test scores (rlwscore). Start by seeing whether the model satisfies the special ANCOVA assumptions. Even if the data violates the assumptions, go ahead and perform the ANCOVA. Is early exposure a statistical factor for GJT scores when the effects of aptitude and input are factored out?

Reporting the Results of an ANCOVA

In a parametric ANCOVA you'll want to report specifically about whether the assumptions of ANCOVA were satisfied. You should then report the results of the test and be sure to include the F-value, the *p*-value, the degrees of freedom and the effect size (post-hoc power is unnecessary; see Section 4.3.4 of the book for more information). You will probably want to report adjusted means and standard errors for your variables. If post-hoc comparisons are made you can then

include confidence intervals. Whether you report about the covariate depends on whether you are just trying to factor its influence out or whether you are still interested in its effect on the dependent variable. Make comments about what it means to have a statistical or non-statistical covariate, and about the size of the effect sizes. In each section I have included some reporting of results so I won't report any specific result here.

For a non-parametric ANCOVA, report the level of trimming and what type of bootstrap was used (bootstrap-t for the Wilcox ancboot function), then report at what points comparisons between the DV and the covariate were made and whether they were statistical (use *p*-values and/or confidence intervals). Show the scatterplot and explain the results.

Summary

Use an analysis of covariance when you want to control for the effect of some variable. Your covariate will most likely be a continuous variable. Ones we saw in this chapter used in second language research designs were pretest scores, language proficiency, intelligence, amount of input in the L2, and reading ability. When you factor the effects of these variables out you will then be able to test for the effect of other independent variables, disregarding the effects of the covariate. Remember that basically the covariate is just another independent variable. In some cases it may be of interest to report whether the covariate was statistical, meaning that it had a statistical effect on the dependent variable. In other cases, the goal may just be to factor the effects of that variable out of the equation, although your reader may be interested to know whether mean scores are being adjusted, which they are if the covariate is statistical. ANCOVA can be used with any of the ANOVA research designs, including one-way, factorial, and RM ANOVA. ANCOVA should be used with caution, however, as it contains even more

assumptions than a regular ANOVA, and these assumptions may not accurately describe the data.

In the case where your ANCOVA violates the assumption that there is no interaction between the dependent/response variable and the covariate, you could use the robust ANCOVA shown here to model your data. Even when data conform to the assumptions of ANCOVA, this robust ANCOVA that does not assume parallel lines for the different levels of the IV can add interesting information to your analysis.

Howell (2002) warns against a use of the ANCOVA when it would result in a situation that would go against logic or common sense. If controlling for your covariate results in a design that does not exist in reality, then it doesn't make much sense to test for it statistically. For example, you probably wouldn't want to factor age of acquisition out of a research design involving early and late bilinguals. Would you really want to examine, say, context of acquisition (naturalistic, instructed, or both) while ignoring the effects of age? Age is an important factor and it would be silly to ignore it while examining the effects of a different variable.

Bibliography

Beech, J. R., & Beauvois, M. W. (2005). Early experience of sex hormones as a predictor of reading, phonology and auditory perception. *Brain and Language*, *96*(1), 49–58.

Clark, M. (2014). Tests of equivalence (PPT). Retrieved from RSS Notes Online website: www.unt.edu/rss/class/mike/5700/Equivalence%20testing.ppt.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*(6), 426–443.

Crawley, M. J. (2007). *The R book*. New York: Wiley.

Culatta, B., Reese, M., & Setzer, L. A. (2006). Early literacy instruction in a dual-language (Spanish–English) kindergarten. *Communication Disorders Quarterly, 27*(2), 67–82.

Cumming, G. and S. Finch. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychology*, 60, 170–180.

Fairley, D. (1986). Cherry trees with cones? *The American Statistician*, 40(2), 138–139.

Fraser, C. A. (2007). Reading rate in L1 Mandarin Chinese and L2 English across five reading tasks. *The Modern Language Journal*, *91*(3), 372–394.

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

Larkin, J. H. & Simon, H. A. (1987). Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science*, 11(1), 65–100.

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research, 24*(1), 35–63.

Lee, J. H. & Macaro, E. (2013). Investigating age in the use of L1 or English-only instruction: Vocabulary acquisition by Korean EFL learners. *The Modern Language Journal*, 97(4), 887–901.

Lim, K.-M., & Hui Zhong, S. (2006). Integration of computers into an EFL reading classroom. *ReCALL, 18*(2), 212–229.

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition, 26*(3), 399–432.

Lyster, R., Quiroga, J., & Ballinger, S. (2013). The effects of biliteracy instruction on morphological awareness. *Journal of Immersion and Content-Based Language Education, 1* (2), 169–197.

Miranda Casas, A., Soriano Ferrer, M. & Baixauli Fortea, I. (2013). Written composition performance of students with attention-deficit/hyperactivity disorder. *Applied Psycholinguistics*, 34, 443–460.

Peters, E., Hulstijn, J. H., Sercu, L. & Lutjeharms, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. *Language Learning*, 59(1), 113–151.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn & Bacon.

Tufte, E. (2001). *Envisioning Information*. Cheshire, CT: Graphics Press.

Van Beuningen, C. G., De Jong, N. H. & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning*, 62(1), 1–41.

Wilcox, R. (2005). *Introduction to robust estimation and hypothesis testing*. San Francisco: Elsevier.

Wilcox, R. R. (2011). *Modern statistics for the social and behavioral sciences: A practical introduction*. New York: Chapman & Hall/CRC Press.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). San Diego, CA: Academic Press.