Two-Way Factorial ANOVA with SPSS

This section will illustrate a factorial ANOVA where there are more than two levels within a variable. The data I will be using in this section are adapted from a dataset called "ChickWeight" from the R statistical program built-in package. These data provide the complex analysis that I want to show here, but I have renamed the dataset in order to better help you understand the types of analyses that we would do in second language research. I call this dataset "Writing.txt," and we will pretend that this data describes an experiment that investigated the role of L1 background and experimental condition on scores on a writing sample. The file is a .csv file, not an SPSS one.

The dependent variable in this dataset is the score on the writing assessment, which ranges from 35 to 373 (pretend this is an aggregated score from four separate judges who each rated the writing samples on a 100-point score). The independent variables are L1 (four L1s: Arabic, Japanese, Russian, and Spanish) and Condition. There were three conditions that students were asked to write their essays in—"correctAll," which means they were told their teachers would correct all of their errors; "correctTarget," which means the writers were told only specific targeted errors would be corrected; and "noCorrect," in which nothing about correction was mentioned to the students.

First of all, we want to examine the data before running any tests, so Table 1 gives a numerical summary of the data. The very highest scores within each L1 were obtained in the condition where no corrections were made, and the lowest scores in the condition where writers were told everything would be corrected. Standard deviations certainly have a large amount of variation

within each L1. The numbers of participants in each group are unbalanced but there are a large number of them at least.

L1	*Condition*	*N*	*Mean*	*SD*
Arabic	CorrectAll	59	48.2	7.1
	CorrectTarget	76	87.0	26.1
	NoCorrect	85	154.4	52.1
Japanese	CorrectAll	30	50.0	8.2
	CorrectTarget	40	101.7	30.9
	NoCorrect	50	182.9	66.0
Russian	CorrectAll	30	52.4	9.9
	CorrectTarget	40	116.7	27.4
	NoCorrect	48	202.5	42.7
Spanish	CorrectAll	30	51.1	9.2
	CorrectTarget	40	109.5	30.3
	NoCorrect	50	224.8	67.0

Table 1 Descriptive statistics for the Writing dataset.

Next, graphics will help in visually getting a feel for the multivariate data. Let's look at data using both of the graphics plots that I introduced in this chapter (I used R for this, so if you want to see the code, look at the document "Two-Way Factorial ANOVA_R"). The top right panel of the interaction plot quickly shows shows that all of the L1s performed similarly in the different conditions (lines are mostly parallel), with the Russian and Spanish speakers performing with the highest scores, and the Arabic speakers with the lowest scores. The bottom left panel shows that there may be an interaction between L1 and condition, because while in the correctAll condition there was little difference between L1s, in the correctAll condition there seemed to be great differences between L1s. It also shows in the main effect panels (top left and bottom right) that Spanish speakers did best overall, and that that those who were not told anything about correction got the highest overall scores (noCorrect) and those who were told everything would be corrected (correctAll) got the most depressed scores.

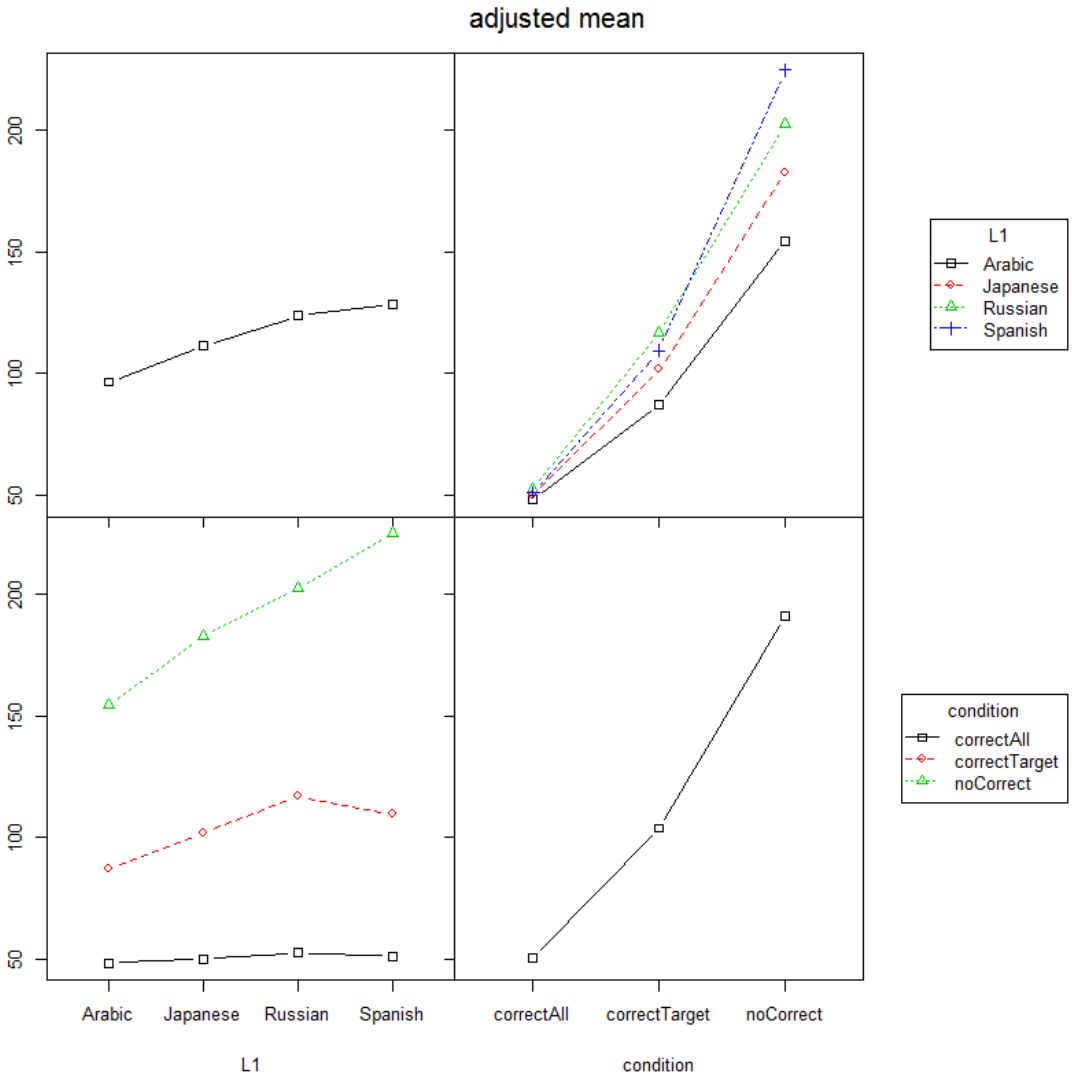

Figure 1 Interaction plot from phia package with Writing dataset.

Next, let's look at an interaction plot with boxplots for main effects:

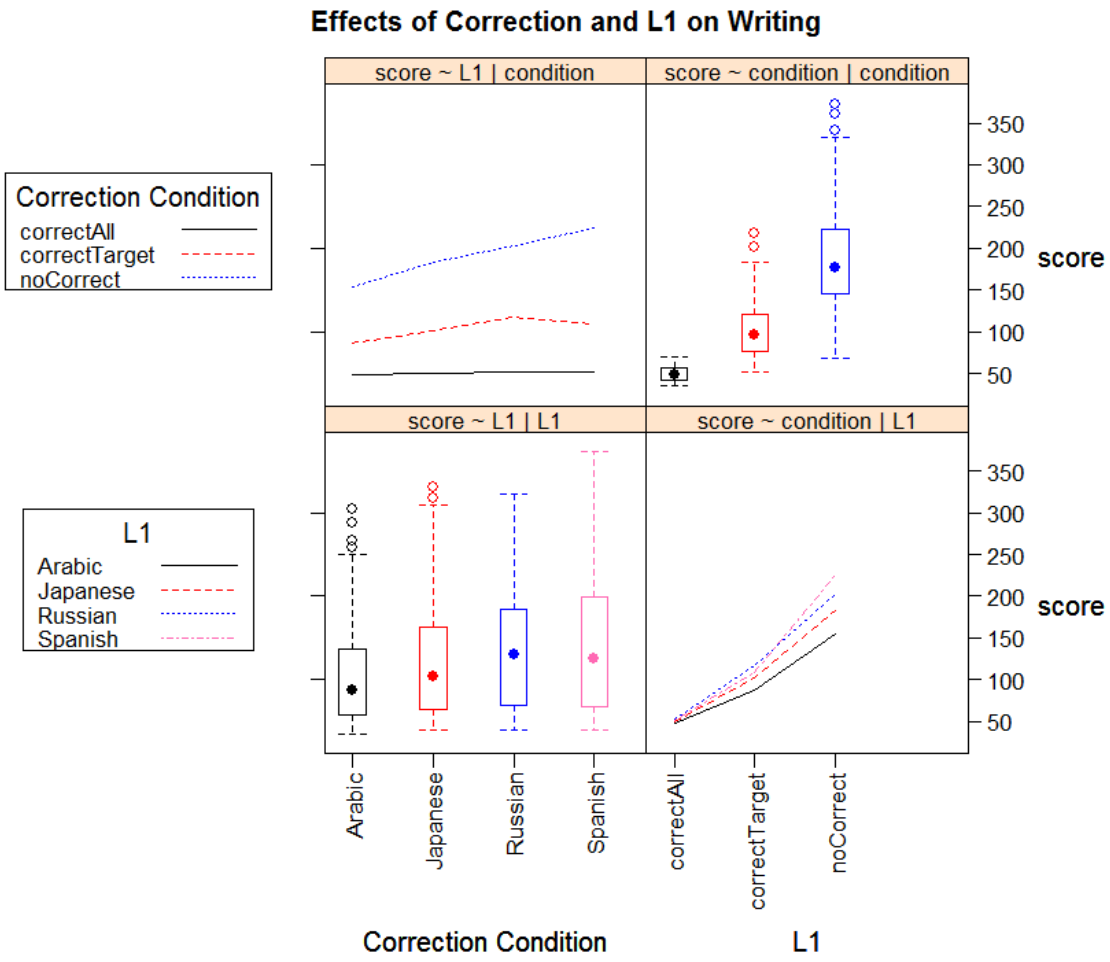

Figure 2 Interaction and boxplot from HH package with Writing dataset.

The interaction plots in Figure 2 don't tell us any more than the ones in Figure 1, but I like the boxplots because we can see that the data are not normally distributed (there are outliers, represented by dots outside the maximum lines of the boxplots, and the distribution is not symmetrical) and variances are certainly not equal for Condition, as the amount of variation for

the correctAll condition is quite low while for the noCorrect it is quite large.

In the first edition of the book I published a barplot for these same data. This is reproduced now in Figure 3. You can see that a barplot is not nearly as informative as either of these graphics for this data.

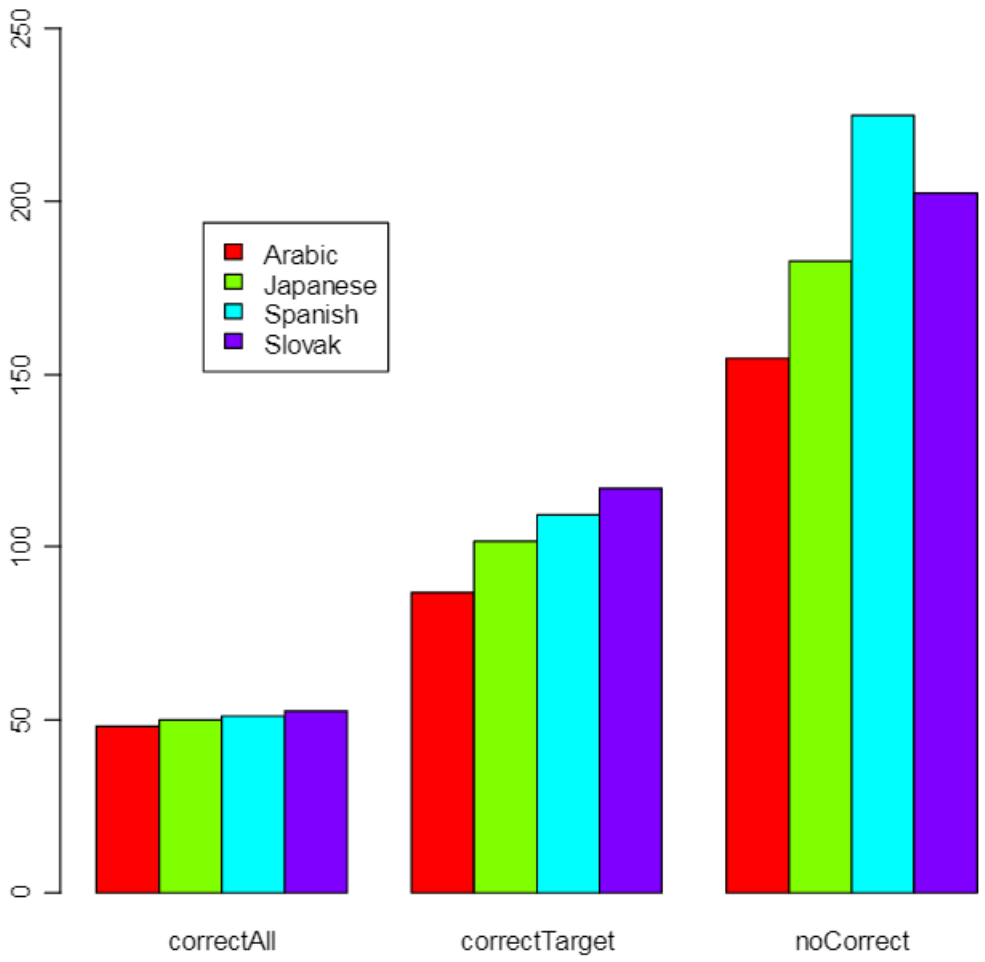

Figure 3 Barplot of Writing dataset

Calling for aTwo-way Factorial ANOVA

Our first step in the analysis is to open data that is not in the SPSS format. To get comma-delimited data into SPSS, simply navigate to where you have the writing.csv file saved, making sure to specify in the dialogue box that you want to see "All Files" in the box labeled "Files of type." Double-click on writing.csv and a Text Import Wizard will automatically open. Press the NEXT button. On Step 2 of 6, change the answer for "Are variable names included at the top of your file?" to Yes. Press the NEXT button again and again until you are on Step 4. Here you should make sure that the answer to the question "Which delimiters appear between variables?" is "Comma." In the "Data preview" box your data should look nicely ordered by this point. Press the NEXT button. In Step 5, you can change the names of the variable if you like, although I won't. Press NEXT, then FINISH, and you will see the data in SPSS.

Going through the steps for a factorial ANOVA, I used the ANALYZE > GENERAL LINEAR MODEL > UNIVARIATE menu. I put SCORE in the Dependent variable box, and L1 and CONDITION in the "Fixed factor(s)" box. In the MODEL button I changed the SS to Type II; both L1 and CONDITION have more than two levels, so in the POSTHOC button I moved both to the right and ticked the LSD and Tukey boxes for comparisons. In the OPTIONS button I moved L1*condition to the right, and also ticked the boxes "Descriptive statistics," "Estimates of effect size," "Homogeneity tests," "Spread vs. Level plot" and "Residual plot." I opened the BOOTSTRAP button, check the "Perform bootstrapping" box, leave the number of samples at 1000, and change the "Confidence Intervals" to BCa. Then I pressed OK and ran the analysis.

In the output, Levene's test is statistical, with $p = .000$ listed in the output. This means the assumption of equal variances is violated, which we pretty much assumed by looking at the

numerical summaries with the widely varying standard deviations (the variance is the square of the standard deviation).

Now going to the test of between-subject effects (Table 2), we see that there are statistical results in all three parts of the ANOVA model: L1, CONDITION, and the interaction between L1 and CONDITION (because the *p*-value located in the "Sig." column is smaller than $p = .05$). The partial eta-squared effect size is largest for CONDITION. The total R^2 for this model is shown at the bottom of the table, and it is $R^2 = .69$, meaning that the combination of these variables accounts for 69% of the variance (or 68% if adjusted for bias).

Table 2 Univariate ANOVA output for Writing: Between-Subjects Effects.

Tests of Between-Subjects Effects

Dependent Variable: score

Source	Type II Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2003816.04[a]	11	182165.095	113.211	.000	.688
Intercept	8577351.074	1	8577351.074	5330.590	.000	.904
L1	131252.065	3	43750.688	27.190	.000	.126
condition	1777365.323	2	888682.662	552.292	.000	.661
L1 * condition	70588.062	6	11764.677	7.311	.000	.072
Error	910739.882	566	1609.081			
Total	11491907.00	578				
Corrected Total	2914555.926	577				

a. R Squared = .688 (Adjusted R Squared = .681)

Now we know that participants performed differently based on their L1, but we want to know whether there are any statistical differences among the speakers and the conditions seem to be highly different from one another, but we should check this statistically as well. The post-hocs that we called for when running the ANOVA can answer these questions. However, the interaction between L1 and CONDITION is statistical, and I have previously stated that simple (or

main) effects may not be of interest when interactions are taking place. So we really want to know whether the combinations of L1 and CONDITION are statistically different from each other. We will use the approach used previously in Section 10.5.1 of the book, where we added to the syntax of SPSS.

Return to your factorial ANOVA command (ANALYZE > GENERAL LINEAR MODEL > UNIVARIATE), and open the PASTE button. The Syntax Editor comes up. Add these two lines and RUN > RUN ALL:

/EMMEANS = TABLES(L1*CONDITION)COMPARE(L1)

/EMMEANS = TABLES(L1*CONDITION)COMPARE(CONDITION)

After your analysis has run (and with the Bootstrap, it takes a while!), you'll be looking for a title in the output that says L1*Condition (actually, you'll have both), and under that "Pairwise Comparisons."

We can now use our estimation approach to analyze this data. We will look at the confidence intervals, either in the "Pairwise Comparisons" box, or in the "Bootstrap for Pairwise Comparisons" box if we want to use bootstrapping, and look at whether there are effects for comparisons and how large the effects are. Table 3 shows the beginning and of the bootstrapped output for the situation where Condition is the conditioning factor.

Table 3 Main output for testing the two-way Interaction Comparisons (bootstrapped).

Bootstrap for Pairwise Comparisons

Dependent Variable: score

condition	(I) L1	(J) L1	Mean Difference (I-J)	Bootstrap[a]				
				Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
							Lower	Upper
correctAll	Arabic	Japane	-1.729	.013	1.807	.342	-5.030	1.622
		Russia	-4.196	.008	2.019	.035	-8.189	-.278
		Spanis	-2.896	.023	1.981	.127	-6.646	1.018
	Japane	Arabic	1.729	-.013	1.807	.342	-1.759	5.370
noCorrect	Arabic	Japane	-28.532	.051	10.695	.008	-49.852	-7.529
		Russia	-48.091	-.483	8.275	.001	-63.555	-33.276
		Spanis	-70.452	.168	11.294	.001	-92.631	-49.023
	Japane	Arabic	28.532	-.051	10.695	.008	6.990	50.066
		Russia	-19.559	-.534	10.629	.064	-40.227	-.337
		Spanis	-41.920	.117	13.134	.002	-67.871	-16.195
	Russia	Arabic	48.091	.483	8.275	.001	31.828	65.490
		Japane	19.559	.534	10.629	.064	-1.505	42.551
		Spanis	-22.361	.651	11.785	.066	-46.218	3.400
	Spanis	Arabic	70.452	-.168	11.294	.001	47.788	94.221
		Japane	41.920	-.117	13.134	.002	16.339	67.814
		Russia	22.361	-.651	11.785	.066	-.807	43.679

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

For the comparisons where the L1 is the conditioning factor and then the conditions are compared, there are 36 comparisons all together, but only half of them are unique for each comparison, so there are 18 comparisons. There is another way to compare the data, which is using Condition as the conditioning factor and then comparing the L1s within that (you'll find this in another table with the same label in the output). There are 6 comparisons of L1 within each condition that are unique, again making for 18 comparisons. This is a large number of comparisons, but we won't worry about it for the moment. Instead, take some time to look at the confidence intervals and see what they look like.

The basic conclusion we can make is that in the “noCorrect” condition L1 was very important. Every L1 group scored differently from the others in this condition, with the mean scores showing that, from the best performers to the worst, we have Spanish L1 > Russian L1 > Japanese L1 > Arabic L1. The bootstrapped confidence intervals were quite wide in all cases, however, and for some comparisons in the “noCorrect” condition we could not be very sure of a strong effect since the lower range of the CI was very close to zero, such as for the comparison between Arabic L1 and Japanese L1 speakers, with a mean difference of 28.5 points and CI of [-49.85, -7.5], with mean scores showing the Japanese did better than the Arabic speakers. In other cases, there were clearer differences with the lower range of the CI quite far away from zero, as with the comparison between Arabic and Spanish speaker, with a with a mean difference of 70.5 points and CI of [-92.63, -49.02], with mean scores showing the Spanish speakers did considerably better than the Arabic speakers.

Of course, this is not nearly the end of the analysis one could do, and in reporting on the results I would want to present descriptive statistics of mean scores, estimates of mean differences between groups, and confidence intervals or bootstrapped confidence intervals for a full complement of comparisons, not just the ones where there was an effect. However, it would probably be redundant to give all 36 results. Eighteen would be enough to establish your story, and my story would be that L1 made a large difference in the “noCorrect” condition, some difference in the “correctTarget” condition, and absolutely none in the “correctAll” condition. Everyone performed very poorly when they thought all of their mistakes were going to be graded.

Note that some readers of your work would certainly object if you reported this many unadjusted

confidence intervals. You could quote Cumming (2011), who I cited above as saying that unadjusted CI were better than any adjusted *p*-values, or you could satisfy your critics by giving *p*-values along with your CIs, as SPSS could easily supply them if you added the term "ADJ(BONFERRONI)" or "ADJ(SIDAK)" to the end of the syntax line that starts with /EMMEANS.

What if you decided that you really didn't need to run all of the comparisons, and instead decided to do planned comparisons to cut down on the number of comparisons you made? This is a good way to cut down on comparisons, although remember from Chapter 9 that I could not find a way to get confidence intervals from these planned contrasts in SPSS, so if you did this in SPSS you would only get *p*-values (in R I was able to get confidence intervals, so there must be a way to do it, but I don't know how in SPSS). See Sections 10.5.3 and 10.5.4 to walk through how to do planned comparisons for this dataset.

Doing Multiple Comparisons on a Two-Way Factorial ANOVA

1. ANALYZE > GENERAL LINEAR MODEL > UNIVARIATE.
2. Open the OPTIONS button. In the box under "Estimated Marginal Means: Factor(s) and Factor Interactions" move everything on the left to the right-hand box. Press CONTINUE. In the UNIVARIATE dialogue box open the PASTE button, which brings up the the SYNTAX EDITOR. Insert syntax that calls for a comparison of the interaction, like this (N.B. items in red should be replaced with your own data name):
   /EMMEANS = TABLES(L1*CONDITION)COMPARE(L1)
   Repeat the syntax line with the other parts of the interaction at the end (this will just change the order that the pairwise comparisons are done in, so in the output choose the one that tells your story best and ignore the other one), like this:
   /EMMEANS = TABLES(L1*CONDITION)COMPARE(CONDITION)
   In the Syntax editor choose RUN > RUN ALL. Then look for results in the part of the output that shows the interaction (like L1*Condition) and find the box that says "Pairwise Comparisons." Look at the confidence intervals that are returned.
3. If you want bootstrapped confidence intervals, open the Bootstrap button and check the "Perform bootstrapping" box, and change the CIs to BCa. A separate box labeled "Bootstrap for Pairwise Comparisons" will show the output.
4. If you'd like to do planned comparisons, you'll need to create a new variable that combines the levels of your two independent variables into one, return to the One-Way ANOVA command, and enter contrasts into the CONTRAST button of that dialogue box.

Bibliography

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge: New York.