Finding Group Differences with Chi-Square when All Your Variables Are Categorical: The Effects of Interaction Feedback on Question Formation and the Choice of Copular Verb in Spanish

> Statistics is a subject of amazingly many uses and surprisingly few effective practitioners.
>
> Bradley Efron and R. J. Tibshirani (1993, p. xiv)

Chi-square is a simple statistical test that is used in a variety of ways. The very variety may lead to some confusion over what the test is good for, but the basic idea behind the test is that you calculate the difference between the scores you observed and the scores you would expect in that situation and then see whether the magnitude of the difference is large or small on the chi-square distribution. Chi-square is a non-parametric test, which means there are fewer assumptions for its use than there are for parametric tests. However, as we will see in this paper, and as noted in Saito (1999) and Hatch and Lazaraton (1991), chi-square is a much-abused test in second language research studies, and often one of its assumptions (that of independence of data) is violated as a matter of course.

This paper will discuss what kinds of statistical tests should be used when you have one or two categorical variables. Chi-square should only be used with ***count data*** (categorical data), not interval-level data. In the case where you have three or more categorical variables, you should use logistic regression, which is not covered in this text.

Recent Examples of Chi-square in the SLA Literature

McDuffie et al. (2013)

This study looked at differences between 4- to 10-year-old boys with fragile X syndrome (who are likely to show signs of autism), autism spectrum disorders, or who show normal development. In the experiment, the participants learned novel words, and one research question was whether there was evidence that individual participants had fast-mapped the novel words. In the experiment the participants had a binary choice between two objects for 4 test items, so a group independence chi-square was performed to see if there were differences in the distribution of results for each group. The results showed that more of the kids who were developing normally learned all 4 words than the kids in the fragile X symdrome or autism spectrum disorders groups. The table below shows the counts in each category.

Number of trials correct:	**Typical development (N=27)**	**Autism spectrum disorder (N=29)**	**Fragile X Syndrome (N=46)**
0	0	1	0
1	0	4	2
2	2	10	20
3	11	9	12
4	14	5	12

Flynn (2012)

This dissertation "examined students' movement in and out of special education and predictors for special education placement" (p. ix). One particular analysis divided students by race (African American or not), looked at whether these students were classified as special education or not, and used a chi-square analysis to determine whether African Americans were disproportionately represented in special education classes. The result of separate group-independence chi-square analyses in Kindergarten, first grade, second and third grade found that African Americans were statistically more likely to be in special education by the end of first grade than non-African Americans, although this imbalance did not continue to second and third grade. Below is a table of the counts of students in each category in first grade.

	In Special Education	Not Special Education
Not African Amerian	15	217
African American	42	282

Oldham, D. (2012)

This study analyzed the effectiveness of a reading intervention program. A chi-square test was used to compare the number of students who met expectations on a criterion referenced competency test (with a reading component) in 4th grade as opposed to the number who met expectations in 5th grade, after they had experienced an intervention designed to boost their reading skills. Among the 4th grade students who were struggling with reading and enrolled in a special program for help, none met expectations for reading on the test, but after the intervention, 10 of the 23 now-5th graders met expectations for the test. The group-independence chi-square test found a statistical difference between the number of students who met expectations in 4th grade and 5th grade.

	Met expectations	Did not meet expectations
4th grade	0	33
5th grade	10	23

Beattie, Webster & Ross (2010)

This study took as a starting point the fact that while most of the time gestures are peripheral to the act of speaking, in some cases gestures seem to be more important and more attended to visually. Using eye-tracking methodology to see which gestures are most attended to, the researchers aimed to see whether there was a relationship between the "level of fixation of gestures and the information uptake from these gestures" (p. 197). A chi-square analysis was performed on 60 gestures that were characterized as C-VPT (meaning they were produced from the viewpoint of the character being talked about; for example, if the speaker says "He ran away," the speaker then pumps her arms as if she is the character himself). The gestures were coded as being fixated upon or not,

and also divided into high span gestures (those that cross the body to a significant extent) and those that are low span (they do not cross any bodily boundaries). The group-independence chi-square test found that the low-span gestures were more likely to be fixated on than the high-span gestures.

	Fixated	Not Fixated
High span	9	21
Low span	17	13

Two Types of Chi-Square Tests

This paper will discuss two main uses of the chi-square test:

- test for goodness of fit of the data
- test for group independence

The first type of test, goodness of fit, is used when we have only **one** categorical variable with two or more levels of choices. The test for group independence is used when there are **two or more** variables, and all of the variables are categorical, with two or more levels of choices. I will illustrate the difference between these two tests in the following sections.

Chi-Square for Goodness of Fit

If we think of the data being partitioned according to its categories, and if we have only one variable, we will think of a one-row table that will list counts. As a concrete example, suppose we looked at a survey question that asked college freshmen what foreign language they wanted to study out of five choices: Chinese, Spanish, French, German, or Japanese. In this study there is only one variable, that of language choice, but it has five levels. Suppose we randomly

surveyed 100 college freshmen at our local Hometown University and this resulted in the data in Table 1.

Chinese	*Spanish*	*French*	*German*	*Japanese*
23	20	15	13	29

Table 1 Fictional data on desired foreign language at One University (Observed Scores).

Our question for this kind of data is whether this distribution is the one we would expect given what we think is the probability of each choice. For simplicity's sake, at this point let us hypothesize that every choice is equally likely, which would mean a 20% chance of choosing each one. Of course, in every sample there is random variation. So do the frequencies we observed match with what we would expect if every chance were equally likely?

To answer this question we will use the chi-square for goodness of fit. This test is used when there is only one categorical variable with two or more levels and we want to measure how good the fit is to the probabilities that we expect. Notice that we surveyed 100 students, and each student falls into only one category or cell in our table.

A chi-square goodness-of-fit test for this data shows that the chi-square statistic (χ^2) is 8.2. The probability of obtaining a statistic this large or larger on the 4 degrees of freedom that we have (note that a df of 4=the number of levels(5) minus 1) is p = .09. If we have set our alpha level to .05, we could not reject the null hypothesis that all choices were equally likely, and we would conclude that no one preferred any one of the language groups over any of the others, but the low p-value might make us think that we were on to something here, and we might seek to improve

the power of the test by gathering additional data.

Chi-Square for Testing Group Independence

You may easily imagine the case where you would have not just one but two categorical variables. To extend the previous example, let us say we randomly survey 100 more freshmen, this time from Big City University, and we now want to know whether there is any difference in the two populations (Hometown University and Big City University) in their choice of preferred language to study. We can construct a table that shows a cross-tabulation of both variables, like the one in Table 2.

		Language					
		Chinese	*Spanish*	*French*	*German*	*Japanese*	*Total*
Population	Hometown U	23	20	15	13	29	100
	Big City U	14	25	10	26	25	100
Total		37	45	25	39	54	200

Table 2 Fictional data on desired foreign language at Two Universities (Observed scores).

Table 2 is called a **contingency table**, because it shows all the events that could happen (Crawley, 2002, p. 180). Notice that 200 people were surveyed, and we find 200 points of data in the table. Each person's data goes into only one cell of the table.

The question that is being asked here is whether there is any association between the two variables, in this case a university population and their choice of language to study.

The chi-square test will test the hypothesis that students at both universities choose to enroll in different language classes randomly. In other words, the null hypothesis says that there is no

difference between enrollment in the various language courses, and no difference between universities in how students choose to enroll in classes. The chi-square test calculates the difference between the scores you recorded (the **observed scores**, listed in Table 2) and the scores you would expect if the null hypothesis of no difference were true (the **expected scores**, listed in Table 3).

	Chinese	*Spanish*	*French*	*German*	*Japanese*
Expected frequencies:					
Hometown University	$\frac{100*37}{200} = 18.5$	$\frac{100*45}{200} = 22.5$	$\frac{100*25}{200} = 12.5$	$\frac{100*39}{200} = 19.5$	$\frac{100*54}{200} = 27$
Big City University	$\frac{100*37}{200} = 18.5$	$\frac{100*45}{200} = 22.5$	$\frac{100*25}{200} = 12.5$	$\frac{100*39}{200} = 19.5$	$\frac{100*54}{200} = 27$

Table 3 Observed and Expected Frequencies for the Foreign Language Survey.

I will just quickly walk you through the process of determining the chi-square statistic for a chi-square test of group independence. My reason for showing this to you, even though you do not have to calculate it by hand, is to demonstrate how simple this test really is. Sometimes you can start thinking that statistics is magic because you don't see the calculations behind the output. Since chi-square is so simple, it is easy to see what is going on in the calculations.

Expected scores are determined by multiplying the total row score by the total column score of any given cell and then dividing by the total number of observations. Thus, the expected score for Chinese at Hometown University is 100 (the total row score for Hometown University) times 37 (the total column score for Chinese) divided by 200 (the total number of observations). Notice in Table 3 that the expected frequencies for both universities in each language category are the same because the row totals in Table 2 are exactly the same in this experiment (exactly 100 students were surveyed at each school).

The chi-square value is now calculated by summing up the difference between the observed (O) and expected (E) score (and squaring it so no negative numbers arise) and then dividing by the expected score:

$$\chi^2 = \sum \frac{(O - E)^2}{E}.$$

This will give the following calculation:

$$\frac{(23 - 18.5)^2}{18.5} + \frac{(20 - 22.5)^2}{22.5} + \frac{(15 - 12.5)^2}{12.5} + \frac{(13 - 19.5)^2}{19.5} + \frac{(29 - 27)^2}{27} +$$
$$\frac{(14 - 18.5)^2}{18.5} + \frac{(25 - 22.5)^2}{22.5} + \frac{(10 - 12.5)^2}{12.5} + \frac{(26 - 18.5)^2}{18.5} + \frac{(25 - 27)^2}{27} = 8.374$$

This chi-square statistic is then checked against the chi-square distribution (with 4 degrees of freedom) to determine the probability of obtaining a chi-square value that is as large as or larger than 8.374. The probability is $p = .07$.

Although technically this goes above $p = .05$, it is quite low, and as an author I might argue that it shows evidence that there is a difference between choice of language at both universities, evidence that could be bolstered by gathering additional data.

Other Situations that May Look like Chi-Square

There may be times when you want to measure the same person more than once on a categorical variable. In this situation, the data may look like it should be analyzed with a chi-square test, but since that would violate the assumptions of chi-square, we should proceed with caution. The following are some scenarios that look like a chi-square analysis would be warranted but which violate the assumption of independence. As Saito (1999) points out, if you perform a chi-square analysis when the data are dependent in this way, your *p*-value will be positively biased, meaning more of the tests will be statistically significant (confidence intervals not going through zero or *p*-values < .05) than really should be.

Scenario One: Case Study with Categorical Data, Only One Participant

First, let's imagine that we have data from one person in a case study on a grammaticality judgment test (GJT). Our GJT is formulated so that the respondent labels each sentence as either grammatical or ungrammatical, and we want to examine the difference in how the respondent performed on the grammatical and ungrammatical items. Our question is whether there is any association between the number of correct responses and the fact that the item was grammatical or ungrammatical. There were 200 items on the test, with the fictitious data in Table 4.

	Grammatical	*Ungrammatical*	*Total*
Correct	80	56	136
Incorrect	20	44	64
Total	100	100	200

Table 4 Data from a Fictitious Case Study with One Participant.

There are 200 responses in the table but they were all made by the same person. Therefore, although this looks very much like Table 3 where we used the chi-square for group

independence, the data are not independent of one another (one of the requirements of chi-square). If you think about it, parametric statistics are about generalizing from the sample statistics to the population. We cannot generalize to an entire population from just one person's data. Generally however, what the researcher wants to test in this situation is just whether the one participant's own choices show any signs of a real pattern that is not due to chance.

In the first version of this book I recommended the binomial test in this situation. In fact, however, the binomial test has exactly the same independence assumption that the chi-square test has, so its usage would also technically violate the independence assumption because each decision is dependent on the same test-taker. However, Howell (2010, pp. 131–2) gives an example of a situation where one judge views 8 stimuli and makes a binary judgment on those 8 items. Howell assumes that the probability of judging each individual item is 50%, and uses the binomial distribution to ask whether the probability of getting 7 out of 8 correct is better than chance. Because the probability of this pattern is only $p = .035$, he concludes that we can reject the null hypothesis that the results are due to chance, and conclude that there is some kind of deliberate pattern in the results. It thus appears that statistical researchers consider the use of a statistical test in this case to be legitimate. We could consider that each choice is independent of the other ones, although in a situation like that represented in Table 4 we might wonder if each choice is independent in the same way that a coin toss result is independent.

Since I do not know of a test that can deal with this data without violating the assumption of independence, one way to proceed might be to just point out the pattern you see and hope it is apparent enough that the reader can also see it. In this case, no statistical test is needed.

However, it appears that using a binomial test or a chi-square test would also be a common choice, although technically not correct. You could note this when giving the results of the test, and note Saito's warning that your results may look better than they really are. In SPSS, access the binary test by going to ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > BINOMIAL. Move your variables to the "Test Variable List" and put in the "Test Proportion" that you want to test (the default is .50). In R, R Commander does not have this command so you will need to use the binom.test( ) command. Simply enter in the number of successes in the first argument, the number of trials in the second argument, and the hypothesized probability of success. You can change this command to test a one-sided hypothesis too by adding the argument alternative=c("less") or alternative=c("greater"). So for the data in Table 4 we'd run the test this way:

```
> binom.test(136, 200, p=.5)

        Exact binomial test

data:  136 and 200
number of successes = 136, number of trials = 200, p-value = 3.904e-07
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.6105301 0.7440388
sample estimates:
probability of success
                  0.68
```

The output shows that the hypothesis that someone would get 136 out of 200 correct if the probability of success is 50% has a very low *p*-value ($p<.0005$) so we can reject the null hypothesis and assume that the person did not respond simply by chance.

Scenario Two: Binary Choice, Only One Variable with Exactly Two Levels

Second, let's imagine a case where we had 30 participants who were tested on their ability to correctly form the past tense on 15 verbs that were presented as 15 separate picture description

tasks. The research question is whether the proportion of successes is statistically higher than one would expect by chance, given the null hypothesis of both choices being equal. We sum up over all 15 items and obtain the results in Table 5.

Correct formation	*Incorrect formation*
339	111

Table 5 Data from a Test with Only Two Choices (Binary Choice).

Here we might assume that we could use a goodness-of-fit chi-square to test the assumption that the participants formed the verbs correctly more often than we would expect if each choice were equally likely. The problem is that we have only 30 participants but we have 450 responses. Each participant contributed more than once to each cell, and participants differ in how much they contributed to each cell—in other words, some participants contributed more to the "Correct formation" cell than others, who might have contributed more to the "Incorrect formation" cell.

In the first edition I recommended the binomial test in this case, but as noted in Scenario One, the binomial test has the same independence assumption that a chi-square test has, and is in fact exactly the same test when there is only one variable with two choices, so that is not a solution to the problem of dependence. Saito (1999) also does not appear to offer a solution to this situation except to say that researchers "should be aware of the statistical assumption that is violated" (p. 467). He points out that the assumption that the data are randomly sampled is frequently violated in SLA research but we continue on with our tests, just realizing that generalizations may be limited, so one approach may be to simply use a chi-square goodness-of-fit here but note that the

p-value may be inflated.

If you were interested in answering the question of whether participants have mastered the verbal system, you could reduce your data (thus losing data) into one outcome only for each participant. In this example that would mean that you could set up a criterion of "success" if the score is more than 50% correct or "failure" if the score was less than 50% correct, and then you would reduce the data to only 30 data points. At this point you could safely use the binomial test to test the hypothesis that the number of participants who achieve success is greater than what would be expected randomly (thanks to Yves Bestgen for this idea). But you can see that this type of analysis may not really answer the question that you wanted.

Scenario Three: Matched Pairs with Categorical Outcome

Let's imagine a third scenario that uses repeated measures with categorical data. Let's imagine a researcher is interested in motivation, and asks 500 incoming Spanish-major freshmen just one question on the very first day of class—"Are you excited about learning Spanish, yes or no?" as a measure of motivation. These students are distributed into the classes of five different teachers, and on the last day of class the researcher asks the students the same question. Table 6 gives a made-up reporting of the data.

	Excited—yes		*Excited—no*		
	First day	*Last day*	*First day*	*Last day*	*Total*
Teacher A	95	88	5	12	200
Teacher B	97	54	3	46	200
Teacher C	87	89	13	11	200
Teacher D	89	63	11	37	200
Teacher E	91	99	9	1	200
Total:	459	393	41	107	1,000

Table 6 Data from a fictitious motivation survey.

Our question would be whether the variable of teacher influenced the excitement students had for Spanish. We can imagine that, if we only wanted to look at the first day or the last day separately, this would very much fit our profile for a chi-square for independence of groups. With separate chi-squares, each person would fit into only one cell of the table. However, this is repeated measures (matched pairs), since each person was surveyed twice. In the case of repeated measures with two categorical variables, the McNemar test can be used.

In SPSS, the McNemar test is accessed by using the ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS menu and then opening the Statistics button and ticking off the McNemar test. For my three variables, I entered them in this order: FIRST in Row, LAST in Column, and TEACHER in Layer 1. With this order, I got the output in Tables 7 through 9.

Case Processing Summary

	Cases					
	Valid		*Missing*		*Total*	
	N	*Percent*	*N*	*Percent*	*N*	*Percent*
first* last* teacher	500	100.0%	0	.0%	500	100.0%

Table 7 Motivation experiment summary with the McNemar Procedure.

Chi-Square Tests

teacher		*Value*	*Exact Sig. (2-sided)*
1	McNemar Test	100	.016[a]
	N of Valid Cases		
2	McNemar Test	100	.000[a]
	N of Valid Cases		
3	McNemar Test	100	.500[a]
	N of Valid Cases		
4	McNemar Test	100	.000[a]
	N of Valid Cases		
5	McNemar Test	100	.008[a]
	N of Valid Cases		

a. Binomial distribution used.

Table 8 Motivation Experiment results with the McNemar Test.

first* last* teacher Crosstabulation

Count

teacher			last		Total
			yes	no	
1	first	yes	88	7	95
		no	0	5	5
	Total		88	12	100
2	first	yes	54	43	97
		no	0	3	3
	Total		54	46	100
3	first	yes	87	0	87
		no	2	11	13
	Total		89	11	100
4	first	yes	63	26	89
		no	0	11	11
	Total		63	37	100
5	first	yes	91	0	91
		no	8	1	9
	Total		99	1	100

Table 9 Motivation experiment crosstabs with the McNemar procedure.

In R Commander, the test can be run on a table of counts. Follow the sequence STATISTICS > CONTINGENCY TABLES > ENTER AND ANALYZE TWO-WAY TABLE (you could use the choice CONTINGENCY TABLES > TWO-WAY TABLE if you wanted to use the raw data instead of a summary count). Put in the totals for only Teacher A. Click the Statistics tab and tick on "Chi-square test of independence," then press OK and you will get the results of the chi-squared test.

In this case you are doing multiple chi-square tests so you might want to adjust p-values for

multiple tests.

For SPSS data results, Table 7 shows the order that the variables were put in, and correctly shows that there were only 500 participants, although there are 1,000 pieces of data. Table 8 shows the results of the statistical test, with the last box giving the *p*-value for each test (five McNemar tests are run, one for each of the teachers). For four out of the five teachers, there was a statistical difference in the answers that students gave from the first day to the last, since *p*-values are below $p = .05$. Only for Teacher 3 did students not change from the beginning to the end of the semester in answer to the question of whether they were excited about learning Spanish (since the probability of seeing the scores on the first and last days given the hypothesis that there was no difference between those days is 50%). However, just knowing the statistical results does not give us a lot of information. We don't know, for example, whether scores were high and remained high for Teacher 3's students, or maybe they started low and remained low.

To get answers to what was happening, whether it was statistical or not, we need to look at the patterns of data in crosstabs, shown in Table 9. This table divides the data into tables of scores for the individual teachers (each teacher is in a separate row). Looking at the crosstabs, Teacher 3 had no students who were in the first–yes, last–no box, meaning no students had decreased motivation with this teacher, but quite a few were first–no and last–no, meaning the teacher had not had success in increasing motivation for some students. Teacher 5 had the highest number of students whose motivation increased (first–no, last–yes), while Teacher 2 had 43 students whose motivation decreased (first–yes, last–no).

The McNemar test can thus answer the question of whether repeated data that represents pairs of

related points (beginning and end, test 1 and test 2) measured with a categorical scale (yes/no, high/low) are different. Note that the McNemar test is not for all types of repeated measures, just paired measures.

Scenario Four: Summary over a Number of Similar Items by the Same Participants (Counting Cases)

Last, let's imagine a case like the one in Scenario Two where participants were scored on whether they had correctly formed the past tense, but with the addition of the variable of regular or irregular verbs. Let's say there were 45 participants. We'll pretend our theory predicted that participants would be much less likely to provide the correct formation on irregular verbs than regular ones. Let's say we had data like that found in Table 10.

	Correct formation	*Incorrect formation*	*Total*
Regular verbs	144	81	225
Irregular verbs	176	49	225
Total	320	130	450

Table 10 Data from a fictitious study of post-tense formation.

Our question here is whether there is any association between the type of verb and the way the participant formed the verb. Here we might assume, since we have two categorical variables, that we can perform a chi-square independent-group comparison. This would be in violation of the assumptions for chi-square, however, since each participant contributes more than once to each cell. One solution for this kind of data would be to examine each of the 15 items in the test with a separate chi-square group-independence test. This would probably not be the best solution, because it inflates the overall error rate and, as Saito (1999) points out, it does not give the researcher any chance of checking for interactions. Saito (1999) recommends using logistic regression on this kind of data and including a participant variable to see if there is an interaction

effect, but notes that such an approach would still not correct the fact that the data are not independent. In this experiment there was a total of 450 pieces of information, but there were not 450 different participants, all tested on only one item. In fact, there were only 45 participants, and they were each tested on ten items. This means that some people may influence one cell of the contingency table more than others, leading to dependence and an increased chance of a Type I error (rejecting the null hypothesis when it is in fact true: in other words, finding more "statistical" results than you should).

A better approach to this type of data is beyond the scope of this book to show but I want to present a possible solution because I believe this type of data is common in our field. My friend Richard Herrington, a statistical consultant, suggests that this is in fact a linear mixed-effects design with nested categorical data (count data). A mixed-effects design has both fixed and random factors. Because the dataset here aggregates answers over the individuals, there is a random effect due to participant. This random effect is nested within the "treatment" factor of regular versus irregular verb (in other words, each participant has multiple data for both regular and irregular verbs). The mixed model can take account of the correlations that occur in the data because the same participants are tested more than once (Crawley, 2007).

The linear mixed-effects design to use here will need to be a generalized mixed-effects model which will let us specify the distribution. We will use the binomial distribution since the dependent variable has only two choices (correct or incorrect formation of the verb) and because there is a fixed number of observations (in other words, it is not count data where the number of trials is not fixed; in that case a Poisson distribution would be used).

In SPSS the way to choose such an analysis would be to use ANALYZE > GENERALIZED LINEAR MODELS > GENERALIZED LINEAR MODELS. When specifying the model you would use the Binary Logistic, which specifies the Binomial as the distribution and the Logit as the link function. In R Commander you would use STATISTICS > FIT MODELS > GENERALIZED LINEAR MODEL. For "Family" choose binomial, and for "Link function" choose logit. For more understanding of how to set up a linear mixed-effects model and interpret its output you can read the document from Chapter 11, "Performing an RM ANOVA the Mixed-Effects Way". I would also recommend additional reading, such as the chapter by Cunnings and Finlayson (2015).

Application Activity: Choosing a Test with Categorical Data

Read the following descriptions of possible linguistic experiments. Decide whether you would use a goodness-of-fit chi-square, independent-group chi-square, or some other test (fill in the name of the test if you can).

1 Native speaker friends. An instructor in the foreign language department noticed that the students who performed best in her Spanish classes often had friends who were native speakers of Spanish. She surveyed her students to determine whether they had friends who spoke Spanish natively or not (yes/no answer), and then grouped her students into those who received As and Bs in class (successful), those who received Cs (moderate), and those who scored lower grades (unsuccessful). The research question is whether having a friend who speaks the language is related to success in the classroom.

Choose one: goodness-of-fit group independence other _____

2 Bilingualism and language dominance. Dewaele and Pavlenko (2001–2003) conducted an online Bilingualism and Emotion Questionnaire (BEQ). All respondents had at least two

languages, and 31% of the sample had five languages (the maximum investigated in the questionnaire). They also asked respondents which language they considered to be their dominant language, and answers were coded as YES (dominant in L1), NO (dominant in another language besides L1) or YESPLUS (dominant in more than one language). Investigate the possibility that the number of languages a person speaks has a relationship to their answer on the dominance question.

Choose one: goodness-of-fit group independence other _____

3 Self-ratings of proficiency. One thousand students at a university in France were surveyed and asked to self-rate their proficiency in English as Poor, Fair, or Good. Researchers wanted to know whether each of these choices was equally likely in a large sample.

Choose one: goodness-of-fit group independence other _____

4 Extroversion and proficiency. The same researchers wanted to know whether extroversion had any relationship to self-perceived proficiency ratings. A subset of the 1,000 students took the EPI (Eysenck Personality Inventory), and were rated as either Extroverted or Introverted. The researchers investigated whether personality and proficiency ratings were related.

Choose one: goodness-of-fit group independence other _____

5 Lexical storage. In order to investigate how the lexicon is constructed, 50 native speakers of English were provided with 12 pictures representing actions involving verbs that participate in the dative alternation and asked to describe the picture in real time (using the present progressive). The utterances were then classified as being either DO or IO based on whether

the direct object or indirect object directly followed the verb (Example: “He’s giving his girlfriend a gift” versus “He’s giving a gift to his girlfriend”). The research question was whether speakers preferred one order over the other for each verb.

Choose one: goodness-of-fit group independence other _____

6 L1 background and success in ELI. Your university’s English Language Institute wants to know whether the number of students who pass the exit exam come from a balanced mix of the L1 of the students who enter. The enrollment last year was composed of students who speak as their L1: French, 30; Spanish, 20; Arabic, 35; Japanese, 60; Korean, 43; Mandarin, 16. The students who passed the exit exam included: French, 23; Spanish, 19; Arabic, 12; Japanese, 47; Korean, 40; Mandarin, 16. You want to investigate whether the proportion who passed the exam from each L1 is approximately the same, assuming that all groups should have an equal chance of passing.

Choose one: goodness-of-fit group independence other _____

7 Foreign accent and study abroad I. Two English teachers at a Japanese university suspect that the foreign accent of their students is greatly improved by a study abroad session. The teachers survey all of the English majors in their school and categorize students into those who have done a study abroad and those who have not. The researchers also ask the teachers to listen to their students read a paragraph and rate the accent of their students as Poor, Fair, or Excellent.

Choose one: goodness-of-fit group independence other _____

8 Foreign accent and study abroad II. The researchers who looked at foreign accent and study abroad wrote a paper and were roundly criticized for their unreliable measure of foreign accent. The researchers went back and gathered recordings of the students reading a paragraph and then asked five native speakers of English to rate each sample as Poor, Fair, or Excellent. The researchers then again investigated the relationship between study abroad and foreign accent.

Choose one: goodness-of-fit group independence other _____

Answers to Application Activity: Choosing a Test with Categorical Data

1 **Native speaker friends**. Use group independence chi-square.

2 **Bilingualism and language dominance.** Use group independence chi-square.

3 **Self-ratings of proficiency**. Use goodness-of-fit chi-square.

4 **Extroversion and proficiency.** Use group independence chi-square (although a better approach might be to use the actual numbers from the EPI instead of collapsing them into a category!).

5 **Lexical storage**. Since the researchers wanted to examine each verb separately, a goodness-of-fit chi-square could be conducted for each of the 12 items. If the researcher thought all the verbs were equivalent and added the items together, they would not be able to use a chi-square (repeated measures here with 12 items), but could possibly use either a binomial test or treat the data as interval-level. The binomial test, like the chi-square test, tests the observed values against the expected values, but when there are only

two categories the binomial test is an exact test while the chi-square is only an approximation.

6 **L1 background and success in ELI**. The McNemar test, which matches up repeated data, would seem to be a good choice here.

7 **Foreign accent and study abroad I.** Use group independence chi-square.

8 **Foreign accent and study abroad II.** This is problematic and probably not a great research method! One way to approach this would be to average the foreign accent ratings of the five judges, but the question is how to average categorical ratings! One could give each rating a number value (Poor=1, Fair=2, Excellent=3) and thus average them, but this is not an interval scale so it is questionable whether averaging should take place that way! Better that the researcher had used a larger scale that could be considered an interval-level variable. If results were averaged one could proceed with the group independence chi-square. McNemar is not appropriate as there are more than two ratings. The best approach would actually be a linear mixed-effect model with nested categorical data, but that is not treated in this book!

Data Inspection: Tables and Crosstabs

With categorical data, tables are a useful tool for seeing a summary of the data and getting a quick view of the patterns that may be present. In the case of goodness-of-fit data, because there is only one variable, a simple tabular summary is the best way to examine the data. For group comparison data, variables can be cross-tabulated to produce a contingency table of the data.

Summary Tables for Goodness-of-Fit Data

In producing a table to summarize patterns with a goodness-of-fit situation, data from Geeslin and Guijarro-Fuentes (2006) will be used. The authors wanted to know whether Spanish speakers (both L1 and L2), in a given context, preferred the Spanish verb *ser, estar*, or the use of both. The dependent variable was the choice of verb, which is categorical and has no inherent ranking or value. Note that the dependent variable is a count of how many times each of the three possibilities of verbs was chosen. The independent variable was also categorical and consisted of membership in one of the three populations (Spanish L1 speaker, Portuguese L1 speaker, or Portuguese L1 learner of Spanish L2). The dataset you will use is one I calculated by using the report for percentage use of the verbs in Appendix B of Geeslin and Guijarro-Fuentes's paper and knowing how many participants were in each group. Note that a summary of the data over all of the items would result in a situation like that of Scenario Four (in the section titled "Other Situations that May Look like Chi-Square"), so only the responses of native speakers of Spanish from Item 3 will be examined (this portion can be found in the GeeslinGF3_5.sav file).

Summary Tables for Goodness-of-Fit Data in SPSS

To make a frequency table, open ANALYZE > DESCRIPTIVE STATISTICS > FREQUENCIES. When you do you will see a dialogue box like that in Figure 1. All you need to do is move your variables to the VARIABLE(S) box.

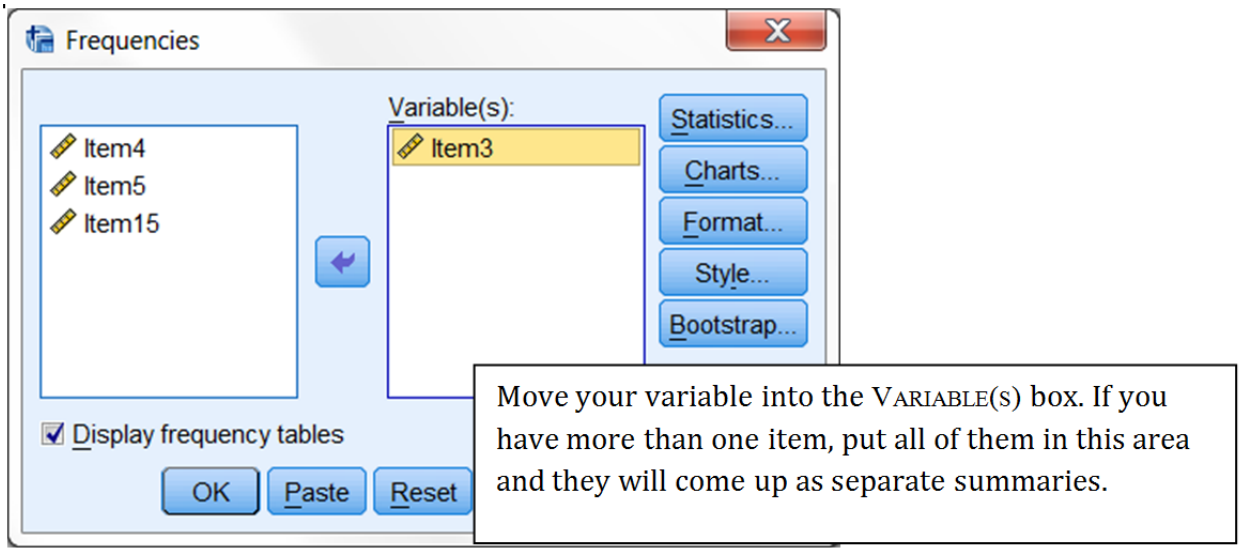

Figure 1 Frequency Table dialogue box in SPSS.

The main output will be labeled with your variable name as shown for Item 3 in Table 11.

Item 3

		Frequency	*Percent (%)*	*Valid Percent (%)*	*Cumulative Percent (%)*
Valid	Estar	13	68.4	68.4	68.4
	Ser	4	21.1	21.1	89.5
	Both	2	10.5	10.5	100.0
	Total	19	100.0	100.0	

Table 11 Output from Goodness-of-Fit summary.

The output in Table 11 shows that, although the majority of native speakers chose to use *estar* in the situation for Item 3 (68.4%), there was some variation with the other two choices as well.

Summary: Creating a Frequency Table with One Categorical Variable in SPSS

1. On the drop-down menu, choose ANALYZE > DESCRIPTIVE STATISTICS > FREQUENCIES.
2. Move the variable or variables you want into the box labeled VARIABLE(S). Press OK.

Summary Tables for Goodness-of-Fit Data in R

To make a frequency table when you have raw count data, in R commander choose STATISTICS > SUMMARIES > FREQUENCY DISTRIBUTIONS. I imported the Geeslin and Guijarro-Fuentes (2006) data as GGF3. Choose the variables you want to see summarized. The additional box you can check on the dialogue box (shown in Figure 2) will conduct a chi-square goodness-of-fit test, which we will examine later in the paper, so for now don't check it.

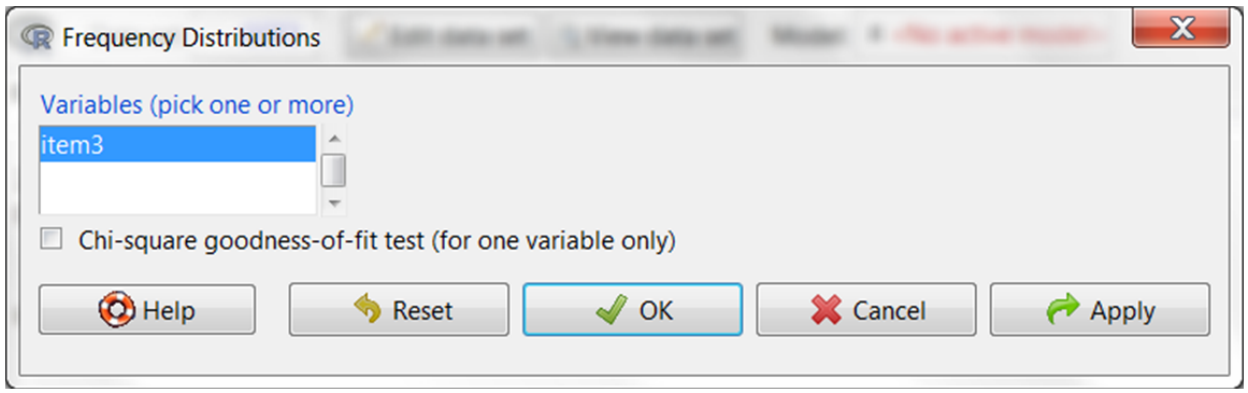

Figure 2 Frequency Table dialogue box in R Commander,

The output is a simple table that counts the number of responses in each category and a table of the percentage of the counts:

```
> .Table  # counts for item3

Estar   Ser  Both
   13     4     2

> round(100*.Table/sum(.Table), 2)  # percentages for item3

Estar   Ser  Both
68.42 21.05 10.53
```

The output shows that, although the majority of native speakers chose to use *estar* in the situation for item 3 (68.4%), there was some variation with the other two choices as well.

The R code for this action in R Commander has several steps because it wants to produce both a raw count table and a percentage table, as I've shown above:

.Table <- table(GGF3$item3) #puts the table into an object that can be called again

.Table # counts for Item3

round(100*.Table/sum(.Table), 2) # percentages for Item3, rounded to 2 decimal points

remove(.Table)

Summary: Creating a Frequency Table with One Variable in R

1. In R commander choose STATISTICS > SUMMARIES > FREQUENCY DISTRIBUTIONS and choose a categorical variable (it must be listed as a "factor").
2. The R code for counts and percentages is the following (N.B. items in red should be replaced with your own data name):

```
.Table <- table(GGF3$item3)
.Table
round(100*.Table/sum(.Table), 2)
remove(.Table)
```

Summary Tables for Group-Independence Data (Crosstabs)

When you have two or more categorical variables, a simple frequency table will not be adequate to capture the complexity of the situation, especially the interactions of the variables. In this case, a cross-tabulation (crosstabs) is used. I will use data from the Dewaele and Pavlenko (2001–2003) Bilingualism and Emotion Questionnaire (BEQ) (to follow along with me, use the BEQ.Dominance file or import it into R and name it beqDom). I will be asking the question of whether the number of languages someone speaks has any relationship to whether they choose their first language as their dominant language, their non-dominant language, or a co-dominant language. In this dataset there are two categorical variables, that of number of languages, and language dominance (CatDominance).

The crosstab data (Table 12) show that although generally the majority of people reply that their L1 is dominant, the number of L1+ answers gets larger as the number of languages goes up, until for those who know five languages this is the answer with the largest count. On the other hand, the number of those who are dominant in a language that is not their L1 is smaller, but this doesn't seem to increase as the number of languages known increases.

Summary Tables for Group-Independence Data (Crosstabs) in SPSS

To inspect the data numerically, use the ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS menu. You'll see the dialogue box in Figure 3. Put one variable each in the ROW(S) and COLUMN(S) boxes. If you have more than two variables, add the additional variables one by one in the LAYER box.

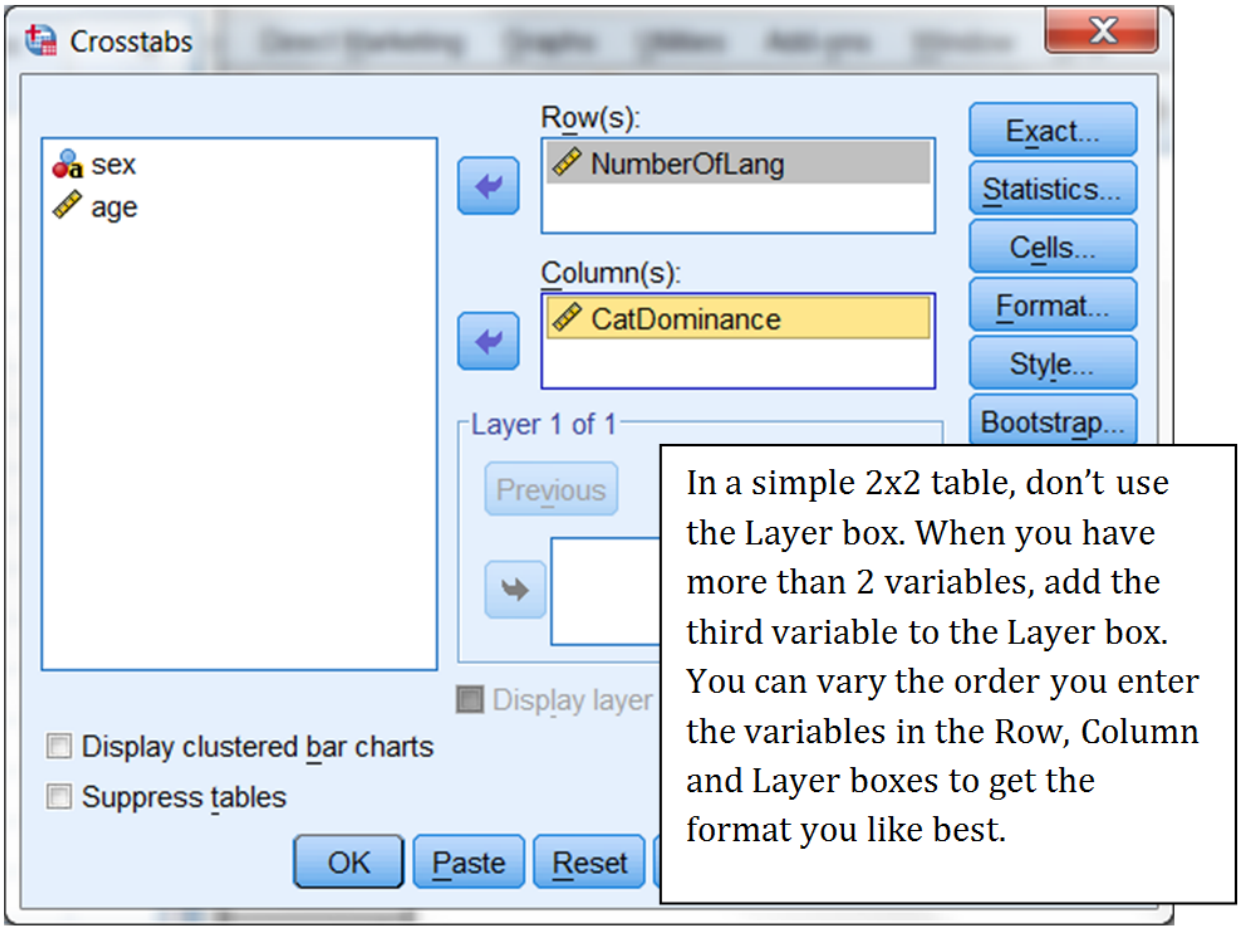

Figure 3 How to make a crosstab in SPSS with two categorical variables.

The main part of the output results in a table such as that seen in Table 12.

Number of languages	L1 Dominant	Other Dominant	L1+ Other(s) Dominant
Two	94	26	17
Three	159	26	83
Four	148	23	110
Five	157	30	163

Table 12 Cross-tabs for number of languages and language dominance with Dewaele and Pavlenko (2001–2003) data.

Creating a Cross-Tabulated Table with Two or More Categorical Variables

1 On the drop-down menu, choose ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS.
2 Move the variable or variables you want into the boxes labeled ROW, COLUMN, and LAYER (use LAYER if you have more than two variables). In order to get the data in a format that is easy to understand you might need to experiment with changing the order of the variables in these areas. Press OK.

Summary Tables for Group-Independence Data (Crosstabs) in R

For the crosstabs in R Commander, pull down STATISTICS > CONTINGENCY TABLES > TWO-WAY TABLE. Note that all the variables need to be classified as "factor" variables in R for the two-way table to be a choice in the dialogue box (see Appendix A if you need help with changing a numeric variable into a categorical "factor" variable; you can use the str( ) command with your data to see the structure of the variables and see if "factor" is listed). This information pertains to the situation where you have raw counts of data, as in the BEQ.Dominance.sav file. Figure 4 shows the dialogue box for the two-way table.

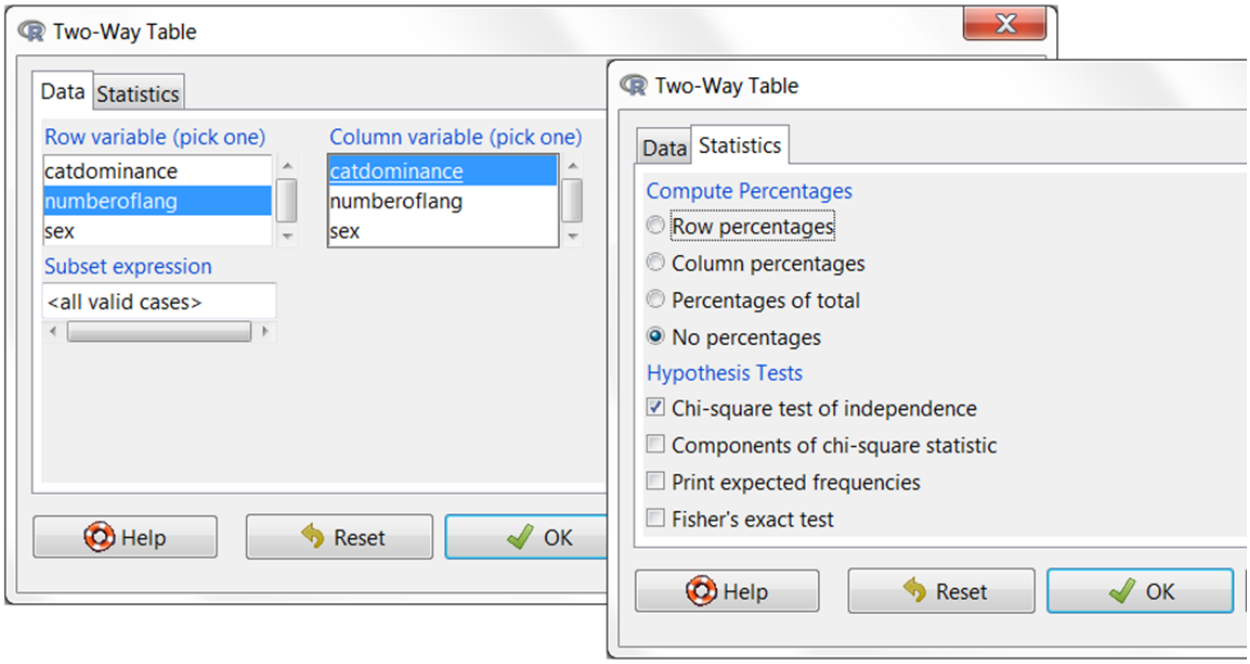

Figure 4 How to make a crosstab in R Commander with two categorical variables.

The output of the two-way table (here done without looking at any percentages) is shown in Table 12 above (with the format improved).

Here is the analysis of the R code that R Commander uses for the crosstab (without any percentages called for) in R:

```
.Table <- xtabs(~numberoflang+catdominance, data=beqDom)
```

.Table <- . . .	The data is put into an object named .Table
.xtabs(~numberoflang+catdominanc	The xtabs() command will build a

e, data=beqDom)	contingency table out of the variables following the tilde. Order affects how contingency tables look—put the row variable first and the column variable second.

After you finish this command you will need to ask R to *show* the data for the table like this:

```
.Table
```

Another situation that might arise is when you have the information from a summary table already (something like what is shown in Table 2). You can then create your own matrix from the data and perform the commands listed below (such as rowPercents( )) on this data as well. The following commands create a 2 × 2 dataset that crosses the use of relative clauses with teaching method:

```
TM<-matrix(c(12,0,18,16),nrow=2,ncol=2,byrow=T, dimnames=list(c("Relative Clauses", "No RCs"), c("Method A", "Method B")))
```

The dimnames( ) command lists the names to give the rows first (Relative clauses, then No RCs), and then the names to give the columns. Type the name of the matrix to see that you have lined everything up correctly:

```
> TM
                  Method A Method B
Relative Clauses        12        0
No RCs                  18       16
```

If you have more than two categorical variables and raw count data, you should choose the MULTI-WAY TABLE option under CONTINGENCY TABLES. The format is the same as for the two-way table except that the Hypothesis Tests are not included. For three-way or higher tables, you may need to experiment to see which order brings out the format of your data that is best suited to your purposes. As a table can have only two dimensions, the control variable is the one which will split your data into separate tables. Just to show you an example of what this would look like, using the same beqDom data I will investigate the question of whether sex interacts with the other variables, and since I have chosen it as the control variable I will get one table with only the data for females and one table with the data for males. Figure 5 shows the multi-way table.

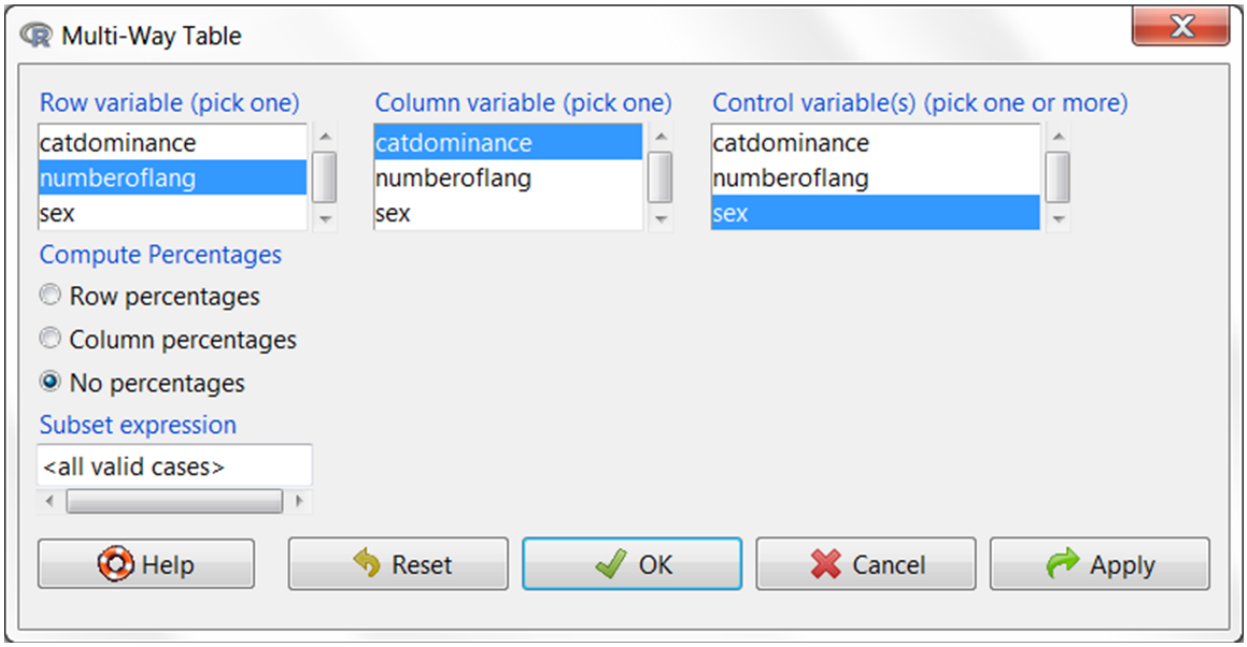

Figure 5 How to make a crosstab with three or more categorical variables in R.

The result is two separate tables, split by sex:

```
, , sex = F

            catdominance
numberoflang YES  NO YESPLUS
       Two    66  19      14
       Three 113  18      61
       Four   95  18      92
       Five   99  23     112
```

```
, , sex = M

            catdominance
numberoflang YES  NO YESPLUS
       Two    28   7       3
       Three  46   8      22
       Four   53   5      18
       Five   58   7      51
```

The R command is very simple—just add the third variable to the end of the equation seen for the two-way table:

```
.Table <- xtabs(~numberoflang+catdominance+sex, data=beqDom)
```

In R you cannot build one table that shows raw counts and percentages in the same count table, but you can easily calculate the percentages and create your own formatted table like this:

```
rowPercents(.Table) #Row Percentages
colPercents(.Table) #Column Percentages
totPercents(.Table) #Percentage of Total; this only works with two-way tables
```

Summary: Creating a Cross-tabulated Table with Two or More Categorical Variables

1. In R Commander, choose STATISTICS > CONTINGENCY TABLES > TWO-WAY TABLE for two variables, or STATISTICS > CONTINGENCY TABLES > MULTI-WAY TABLE for more than two variables. The default is to report counts, but repeat and choose different choices in order to get row percentages, column percentages, or total percentage (two-way table only).
2. The R code for creating cross-tabulated tables is (replace code in red with your own data):

```
.Table <- xtabs(~numberoflang+catdominance+sex, data=beqDom)
rowPercents(.Table) # Row Percentages
colPercents(.Table) # Column Percentages
totPercents(.Table) # Percentage of Total; this only works with two-way tables
remove(.Table)
```

3. In the case where you do not have raw data but only summarized data, you can make a table for use with other statistical commands later in the chapter like this (replace code in red with your own data):

```
TM<-matrix(c(12,0,18,16),nrow=2,ncol=2,byrow=T,
dimnames=list(c("Relative Clauses", "No RCs"), c("Method A", "Method B")))
```

Application Activities with Tables of Categorical Variables

1. Use the Mackey and Silver (2005) dataset and examine the frequency of the differing developmental levels of question formation in the pretest (use the MackeySilver2005.sav file). What is the most frequent developmental level? What is the least frequent?
2. Use the Mackey and Silver (2005) dataset to examine whether there is any difference between experimental groups on the pretest data. From just eyeballing the data, does it appear that students in both experimental groups had roughly the same distribution of developmental levels for question formation?
3. Use the Dewaele and Pavlenko BEQ data (BEQ.Dominance). By eyeballing the crosstab

percentages, examine whether the number of languages someone speaks has any relationship to whether they choose their first language as their dominant language (YES), their non-dominant language (NO), or a co-dominant language (YESPLUS). Use the categorical variables "CatDominance" and "NumberOfLang."

4 Further calculate the three-way intersection of number of languages known, dominance, and sex in the BEQ.Dominance file. Does the pattern noted in Question 3 seem to hold equally well for both males and females?

Answers to Application Activities with Tables of Categorical Variables for SPSS only

1 Mackey and Silver (2005) frequency chart

Choose ANALYZE > DESCRIPTIVE STATISTICS > FREQUENCIES. Move the PRETEST variable into the box and click OK.

You should have N = 26, and the most frequent level is 2, while the least frequent is 4.

2 Mackey and Silver (2005) crosstabs

Choose ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS. I put the EXPGROUP variable in the "Row" box and PRETEST in the "Column" box. Click OK.

You should have N = 26. From the numbers it looks like there were more participants in the experimental group who were at lower developmental levels.

3 Dewaele and Pavlenko (2001–2003) data

Choose ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS. To answer the question about the number of languages someone speaks and their reported language dominance, I put NUMBEROFLANG in "Row" and CATDOMINANCE in "Column."

You should have total N = 1036. Just eyeballing the numbers, it looks like a larger percentage of those who speak three languages report co-dominant languages (83 out of 268) than those who speak only two languages (17 out of 137). The trend of more people speaking co-dominant languages seems to hold for those who speak four and five languages as well. So just from looking at the data I would suspect there will be a relationship between number of languages someone speaks and reported language dominance.

4 Dewaele and Pavlenko

Choose ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS. To answer the question about differences between males and females, I put NUMBEROFLANG in "Row" and CATDOMINANCE in "Column," and SEX in "Layer."

You should have N=1036. In looking at males and females separately, the pattern of increasing numbers responding that dominance is among more than one language as number of languages known increases holds approximately true for both males and females, although it seems stronger in females (for females with 5 lges, YES = 99 and YESPLUS = 112, while for males with 5 lges, YES = 58 and YESPLUS = 51).

Visualizing Categorical Data

Categorical data consist of counts of frequencies, so a traditional way of visualizing such data has been with barplots. In this section I will show you how to create barplots, but I'd also like you to be aware that there are new and very exciting ways to visualize categorical data. Friendly (2000) notes that while methods for visualizing quantitative data have a long history and are widely used, methods for visualizing categorical data are quite new. Two plots, the mosaic plot (Meyer, Zeileis & Hornik, 2006) and the doubledecker plot, can offer a lot of information to the reader and give statistical intuitions, but these are only available using R.

Barplots with One Categorical Variable in SPSS

There may not be a lot of reasons to make a barplot when you have only one variable (it does not really provide a lot more information than what you can get from the numbers in a table), but, if you really want to do it you can. I will illustrate how a barplot can be made with just one variable by looking at the Geeslin and Guijarro-Fuentes (2006) data. The variable "Item 3" in the GeeslinGF3_5.sav file records whether participants chose to use the verb *estar*, the verb *ser*, or whether both verbs were equally plausible in that context.

To make a barplot with one variable, choose GRAPHS > LEGACY DIALOGS > BAR (in Version 12.0 the LEGACY DIALOGS step is absent). A dialogue box will appear. Choose SIMPLE and then SUMMARIES FOR GROUPS OF CASES, as shown in Figure 6. Press the DEFINE button. Move one variable into the "Category Axis" box.

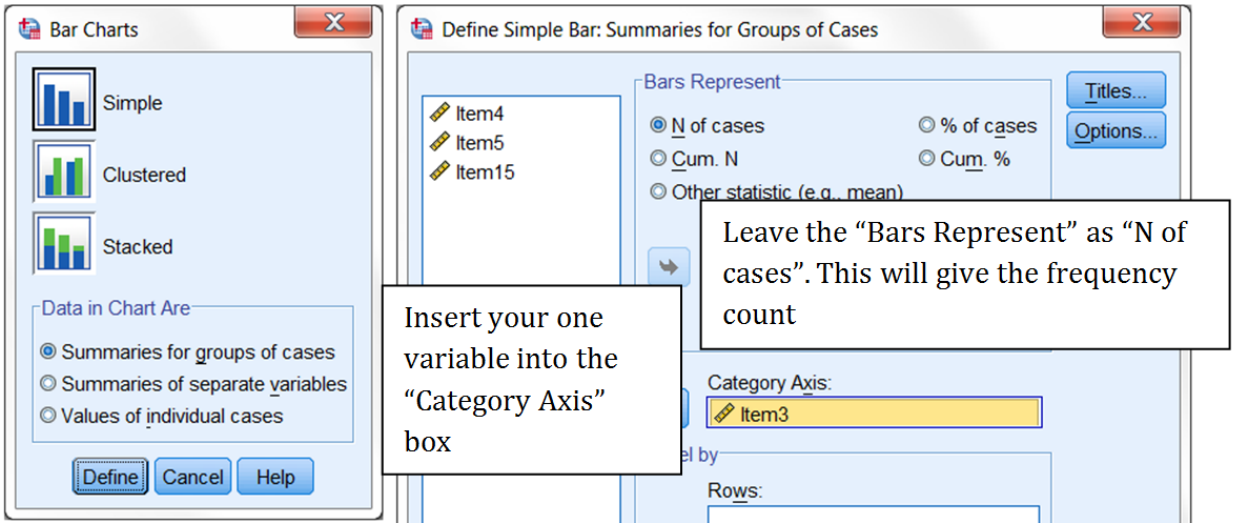

Figure 6 Dialogue box for Bar Charts in SPSS.

The resulting barplot in Figure 7, which just displays a count of the data, shows graphically that *estar* is by far the most frequent response to Item 3.

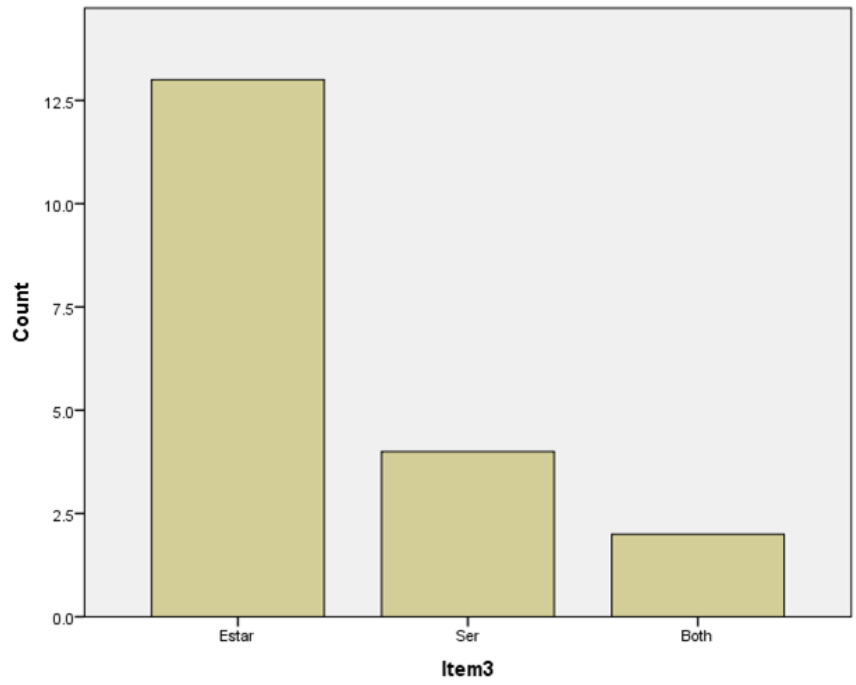

Figure 7 Boxplot with one categorical variable.

Barplots with One Categorical Variable in R

Using R Commander, choose GRAPHS > BAR GRAPH (make sure you have imported the GeeslinGF3.sav file; I named it GGF3). The dialogue box will show any variables that are labeled as a "Factor" (see Appendix A if you need help with changing a numeric variable into a categorical "factor" variable; you can use the str( ) command with your data to see the structure of the variables and see if "factor" is listed). You can add axis or graph labels here as well (see Figure 8).

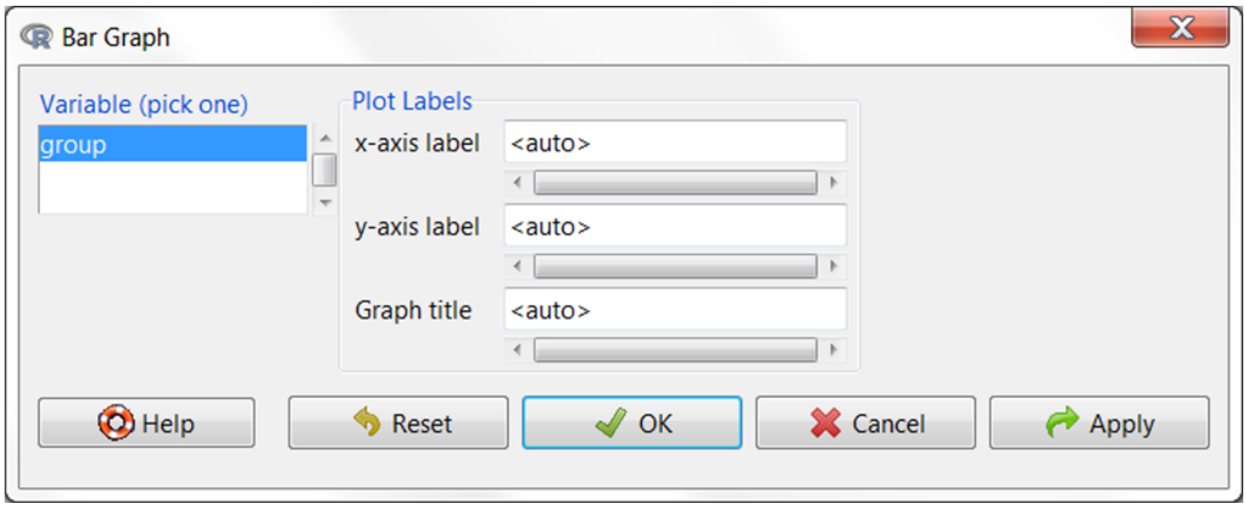

Figure 8 Dialogue box for bar graphs with one categorical variable in R Commander.

The resulting graph will look like Figure 7. The R syntax that generates this figure can be analyzed as follows:

```
barplot(table(GGF3$item3), xlab="item3", ylab="Frequency")
```

barplot( )	The command to make the barplot.
table(GGF3$item3)	Turns geeslin3$Item3 into a contingency

	table (we saw this command earlier in the paper to make a summary count table of one variable).
xlab, ylab	These arguments label the x and y axes respectively.

For this code, if you do not explicitly enter the labels, the labels already affixed to the variables will be automatically inserted.

Barplots with Two Categorical Variables in SPSS

In order to look at barplots with two categorical variables in SPSS I will use the Mackey and Silver (2005) dataset. The authors wanted to know whether being in an experimental group where feedback was given helped the participants improve in their ability to form questions. The authors calculated the pretest question-formation developmental level, and then coded for a categorical variable of improvement or no improvement. The SPSS file is called MackeySilver2005.sav.

For two categorical variables, choose GRAPHS > LEGACY DIALOGS > BAR from the menu. You will get the same dialogue box as seen in Figure 6, but this time change the default SIMPLE box to the CLUSTERED box, and leave the default button on SUMMARIES FOR GROUPS OF CASES. Press the DEFINE button and then enter one variable into the CATEGORY AXIS box. You will enter the other variable into a box labeled DEFINE CLUSTERS BY. The variable that you put into this last box will be the one that is compared in the differing groups. In other words, it is the variable that

contrasts the bars in each group. For example, in Figure 9, because I entered DELPOSTTEST (the Delayed Post Test) into the DEFINE CLUSTERS BY Box, the two outcomes in the test (development or no development) are the two bars that are compared in each group. The variable that is put in the CATEGORY AXIS box will result in separating groups. For Figure 9 I put the EXPGROUP variable (Experimental Group) into the CATEGORY AXIS box.

Because you are working with categorical data where you are simply measuring counts, there is no need to change the BARS REPRESENT button to anything different from "N of cases."

For two categorical variables, the barplot in Figure 9 is obtained (this plot has been modified slightly for format). It shows that the number of students who developed in their question-formation ability (the white box) was greater in the experimental group than in the control group, but in my mind there is still a question as to whether a boxplot is worth the space as it does not provide much information. The 4 numbers represented in Figure 9 could be presented in a concise table with much less space used.

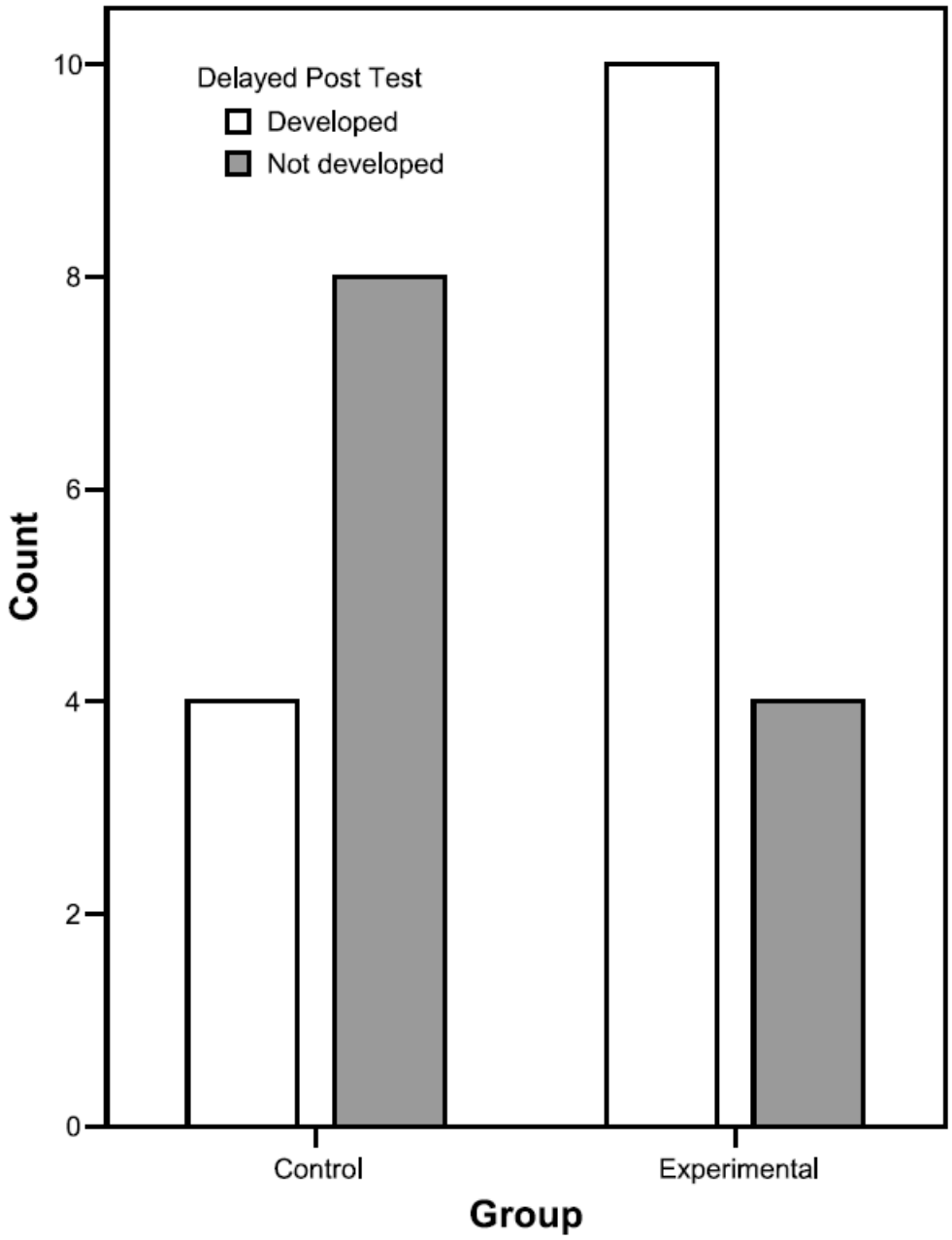

Figure 9 A barplot with two categorical variables using data from Mackey and Silver (2005).

SPSS has no good way to layer three categorical variables in a barplot. Exercise # 2 in the application activities in the section "New Techniques for Visualizing Data" has three variables

but I will ask you to look only at two variables at a time. It would be possible, of course, to put the two barplots side by side for a comparison. The newer graphics I will show you in the three sections called "Association Plots," "Mosaic plots" and "Doubledecker Plots" are able to integrate three or more categorical variables at one time.

Barplots with Two Categorical Variables in R

Perhaps because barplots are not a highly favored graphic by more statistically minded people, they are not very sophisticated in R Commander. You can use R Commander to create a barplot for one categorical variable but to make more sophisticated barplots with two variables, R code will need to be used. Here I'll use the Dewaele & Pavlenko BEQ data (import the BEQ.Dominance.sav file into R as beqDom). We will look at the relationship between the number of languages a person knows and which language they say is their dominant language. Look at a barplot of this data in Figure 10.

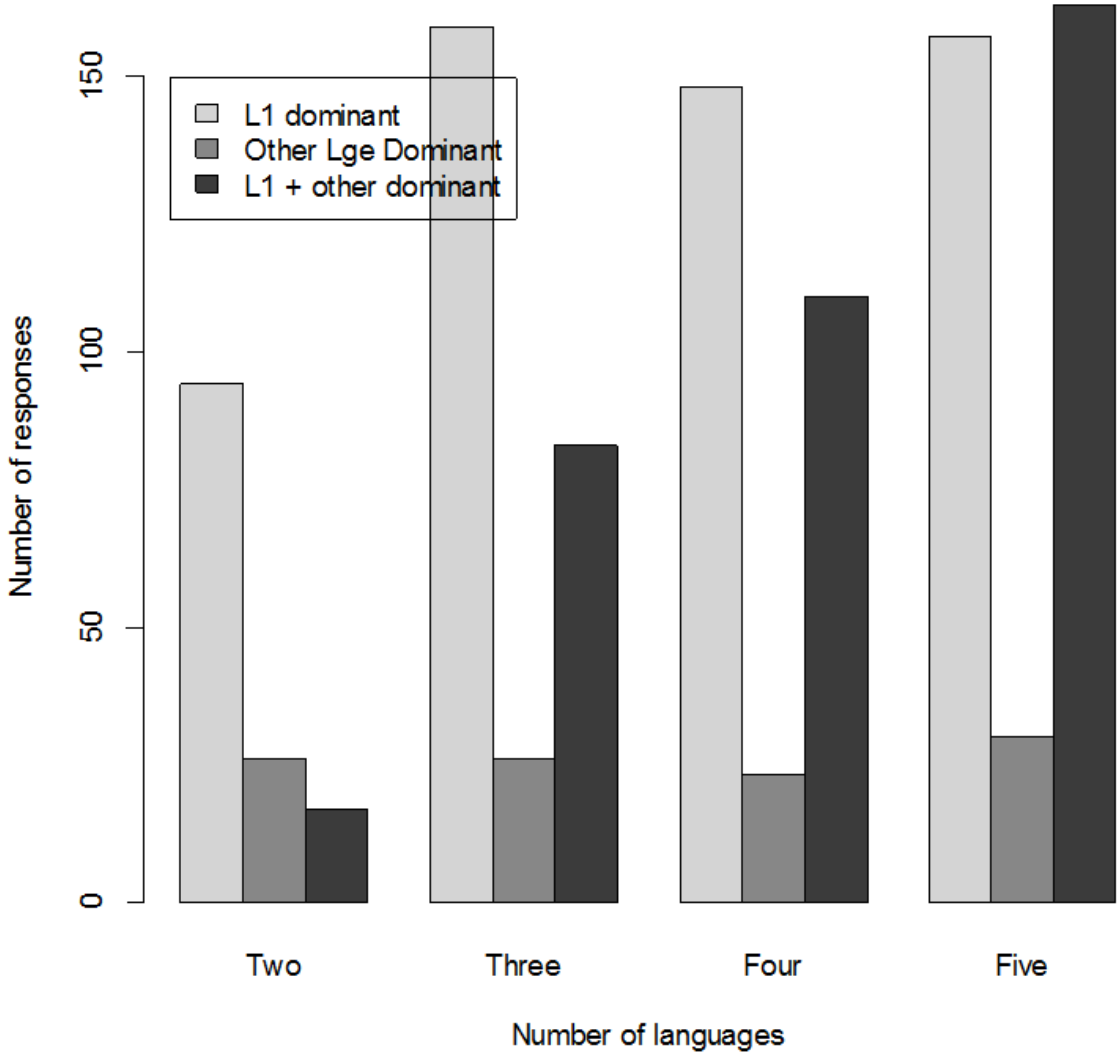

Figure 10 Barplot of two categorical variables in R from Dewaele and Pavlenko (2001–2003) data.

The code for this barplot is long but gives you a lot of control over colors, placement of legend, etc. Use the epitools package (Aragon, 2012).

```
library(epitools) #use install.packages("epitools") if you do not already have this package
colors.plot(TRUE) #use this to pick out three colors you want in the boxplot columns
```

#when you are finished, pull down the one menu choice of STOP and choose STOP LOCATOR

```
x y color.names
1 14 12 grey83
2 14 11 grey53
3 14 10 grey23 #returns names of the colors you picked
attach(beqDom) #by doing this, we can just specify variable names
barplot(tapply(catdominance, list(catdominance, numberoflang),length),
col=c("grey83", "grey53", "grey23"), beside=T, ylab="Number of responses",
xlab="Number of languages")
locator() #allows you to pick out a point on graph to locate the legend
$x $y
[1] 0.8272034 158.2098 #returns the position you chose on the graph
```

#when you have clicked a spot on the graph,

#pull down the one menu choice of STOP and choose STOP LOCATOR

```
legend(.827,149.7, legend=c("L1 dominant", "Other Lge Dominant", "L1 + other
dominant"),
fill=c("grey83", "grey53", "grey23")) #use the color and location data you got previously
detach(beqDom)
```

The analysis of the barplot command is:

```
barplot(tapply(catdominance, list(catdominance, numberoflang),length),
col=c("grey83", "grey53", "grey23"), beside=T, ylab="Number of responses",
xlab="Number of languages")
```

---

barplot( )	The command to make the barplot.
tapply (x, index, function)	Applies a function to the array in x using a list of factors in the index argument.
x= catdominance	This array for the tapply command is the CATDOMINANCE factor. Choose the dependent variable. I want my barplot to show how many people are in each of the categories of dominance (there are three: L1, LX, and L1PLUS). The variable you enter here is the one that is compared in the differing groups. In other words, it is the variable whose bars contrast in each group (here, category dominance).
index=list(catdominance, numberoflang)	This index for the tapply command gives a list of two factors. If I put in only numberoflang, I would end up with only one bar for each factor in numberoflang. In order to split these up into three bars for each factor (one for each of the dominance categories), I need to put both factors that I want in this index list. The second variable is the one that determines how to split the data into groups (so in the barplot there are groups for the number of languages spoken).
function=length	This function for tapply is just the number of responses, so I use the function length. Other possible functions you might want to use include: mean, median, sum, range, quantile.

col=c(. . .)	Specifies the colors to fill in the boxes with; I used the epitools package, color.plot=TRUE command to bring up a graphic containing all of R's colors (see code above). I then clicked on the three colors I wanted, and the print-out returned the names of the colors.
beside=T	Causes the bars of the barplot to stand upright.
ylab="", xlab=""	Gives labels to the x- and y-axes.

The analysis of the legend command is:

```
legend(.827,149.7, legend=c("L1 dominant", "Other Lge Dominant", "L1 + other dominant"),
fill=c("grey83", "grey53", "grey23"))
```

legend (x, y, legend, fill)	The command to plot a legend at point x, y, and to add a legend and fill colors for the legend.
.827, 149.7	I obtained this value for x and y by previously using the locatorcommand.
legend=c("L1 dominant", "Other Lge Dominant", "L1 + other dominant")	The part inside the c() specifies what characters to use to label the legend
fill=c("grey83", . . .)	If you want to have boxes that show colors matching your barplot boxes, add this argument.

Creating Barplots Summary

Creating a Barplot in SPSS

1. On the drop-down menu, choose GRAPHS > LEGACY DIALOGS > BAR.
2. If you have one categorical variable, choose SIMPLE for the type of barplot, and SUMMARIES FOR GROUPS OF CASES. Put your variable into the "Category Axis" box.
3. If you have two categorical variables, choose CLUSTERED for the type of barplot, and SUMMARIES FOR GROUPS OF CASES. Put the variable that will define clusters on the x-axis in the "Category Axis" box and the variable that will contrast bars in the "Define clusters by" box.

Creating a Barplot in R

1. If you have one variable, you can use R Commander's drop-down menu. Choose GRAPHS > BAR GRAPH. Pick your one variable. The R syntax for one variable is (replace items in red with your own data names):

```
barplot(table(GGF3$Item3))
```

2. If you have two variables, use the R syntax, listing first the variable that is compared in the differing groups:

```
barplot(tapply(catdominance, list(catdominance,
numberoflang),length),
col=c("grey83", "grey53", "grey23"), beside=T, ylab="Number
of responses", xlab="Number of languages")
legend(.827,149.7, legend=c("L1 dominant", "Other Lge
Dominant", "L1 + other dominant"),
fill=c("grey83", "grey53", "grey23"))
```

Tip: Although barplots can be useful to help visualize categorical data, they are an information-poor graphic and do not provide much data. There is usually not adequate justification for the space they take up. A leading name in statistical graphics, Edward Tufte, has said that "[t]he bar chart wastes space; you could show at least 100 numbers in the space that now shows 1 number" (third comment under the Sparklines: Intense, simple, word-sized graphics heading). Instead, consider using a more information-rich graphic, such as those described in the following sections.

Tip: In this chapter I have showed you how to make barplots for categorical data. I do not recommend barplots for graphing interval data because boxplots are much more information-rich and useful to your reader (see Larson-Hall & Herrington, 2009 for more information). A giant in information design, Wainer (1996, p. 105) says, "[i]t is certainly profligate to use an entire bar when all of the information about the mean is contained in the location of the top line; the rest is chartjunk" (p. 105). However, if you do insist on using a barplot for interval data you ought to put error bars on it (the frequency counts shown here cannot have error bars because categorical data is summarized using counts and percentages, not measures of dispersion) and label them as to what kind of error bars they are. Do not use error bars if you have repeated measure data, however (see Belia, Fidler, Williams & Cumming, 2005 for more information).

If you do insist on making a barplot using interval-level data where bars represent mean scores instead of counts, in SPSS choose CLUSTERED and SUMMARIES OF SEPARATE VARIABLES in the BAR CHARTS dialogue box. Now any variables you enter into the BARS REPRESENT area will show their mean score. Use the CATEGORY AXIS box to split the data into different groups. Add error bars through the OPTIONS button.

Instructions on how to write code that will produce error bars in R are given in Crawley (2007) starting on page 56. You should also choose mean for the function instead of length as was shown here for count data. You can also see the barplot and code in Chapter 11 for the Writing.txt file.

Application Activities with Barplots

1 Using the fabricated dataset of LanguageChoice.sav, create a barplot that shows the distribution of which language students choose to study, based on which university (POPULATION) they come from. Comment on big differences in language preference between the two universities.

2 Using the fabricated dataset Motivation.sav, create one barplot that shows the distribution of YES and NO responses for the five teachers at the beginning of the semester (FIRST). Then create another barplot that shows the distribution of responses at the end of the semester (LAST). What do you notice from the visual data?

3 Use the Mackey and Silver (2005) dataset and create a barplot that examines the development of questions on the immediate posttest (DEVELOPPOST), categorized by experimental group (GROUP). We saw in Figure 9 from the delayed posttest that more students in the experimental group developed than students in the control group. Does that pattern hold true for the immediate posttest?

4 Use the Dewaele and Pavlenko BEQ data (BEQ.Dominance). Graphically explore the question of whether the number of languages someone speaks has any relationship to whether they choose their first language as their dominant language (YES), their non-dominant language (NO), or a co-dominant language (YESPLUS). What do you conclude, from looking at the data?

Answers to Application Activities with Barplots (Conclusions only)

1 LanguageChoice.sav

Students at Hometown U. seemed much more interested in Chinese than students at Big City U., while students at Big City U. were much more interested in German than students at Hometown U.

2 Motivation.sav

What I notice here is that students in all classes were excited at the beginning of the semester, but only students with teachers 1, 3 and 5 remained highly excited about learning Spanish by the end of the semester.

3 Mackey and Silver (2005)

The graph shows that in the immediate posttest, the pattern seen in Figure 9 did NOT hold. For the immediate posttest, both groups had more students who developed than who didn't, although the number who developed is less for the control group than for the experimental group (and the

number who didn't develop is roughly equal in both experimental conditions). It seems that the treatment was more effective in the long run, not immediately afterwards.

4 Dewaele and Pavlenko (2003) data

When I looked at the graphic I wanted to know what percentage of the people chose each response, not the actual number of participants. The graph clearly showed that the percentage of people who claimed to have more than one dominant language was much larger with people who knew three or more languages than with those who only knew two.

New Techniques for Visualizing Data

The next part of this document will contain ways of visualizing categorical data as flat contingency tables. Meyer, Zeileis, and Hornik (2007) explain that, in these kinds of plots, "the variables are nested into rows and columns using recursive conditional splits. . . . The result is a 'flat' representation that can be visualized in ways similar to a two-dimensional table" (p. 5). The three plots that I like best are association plots, mosaic plots (Meyer, Zeileis & Hornik, 2006), and doubledecker plots. Although commands for association plots and mosaic plots exist in the base R system (assocplot, mosaicplot), I have found that it can be difficult to get data into the correct format to work with these commands. A better choice is the vcd package (Meyer, Zeileis & Hornik, 2014), which provides an easy way to format data through the structable( ) command, and provides more ways to manipulate graphic parameters (see Meyer, Zeileis, & Hornik, 2007 for examples). The vcd package also provides other plots for visualizing categorical data, such as the sieve plot, cotab plot, and pairs plot, although these plots will not be explored here.

Association Plots

The Cohen–Friendly association plot in the vcd package (using the command assoc( )) shows visually where the counts in the data exceed the expected frequency and where they show less than the expected frequency. More technically, it shows the "residuals of an independence model for a contingency table" (Meyer, Zeileis, & Hornik, 2007). In this way, the association plot can help the reader quickly ascertain which interactions are different from what is expected by the assumption that the two variables are independent. Let's examine the Dewaele and Pavlenko (beqDom) data for the relationship between number of languages and language dominance to illustrate this graphic.

In Figure 11 you can see that all of the boxes above the line are blue and have solid lines and those that are below the line are red and have dotted lines. What we can see here is that there are more persons (solid lines) with two languages who say their L1 is their dominant language (YES) or is not their dominant language (NO) than expected, while there are fewer persons with two languages (dotted lines) who say they are dominant in more than one language (YESPLUS) than we would expect if there were no relationship between number of languages and language dominance (the null hypothesis). The departure from expected is not so striking for those with three and four languages, but we again see some striking differences for persons with five languages. There are fewer persons (dotted lines) with five languages who say their L1 is their dominant language (YES), and there are more persons (solid lines) with five languages who say they are dominant in more than one language (YESPLUS) than would be expected if the variables were independent.

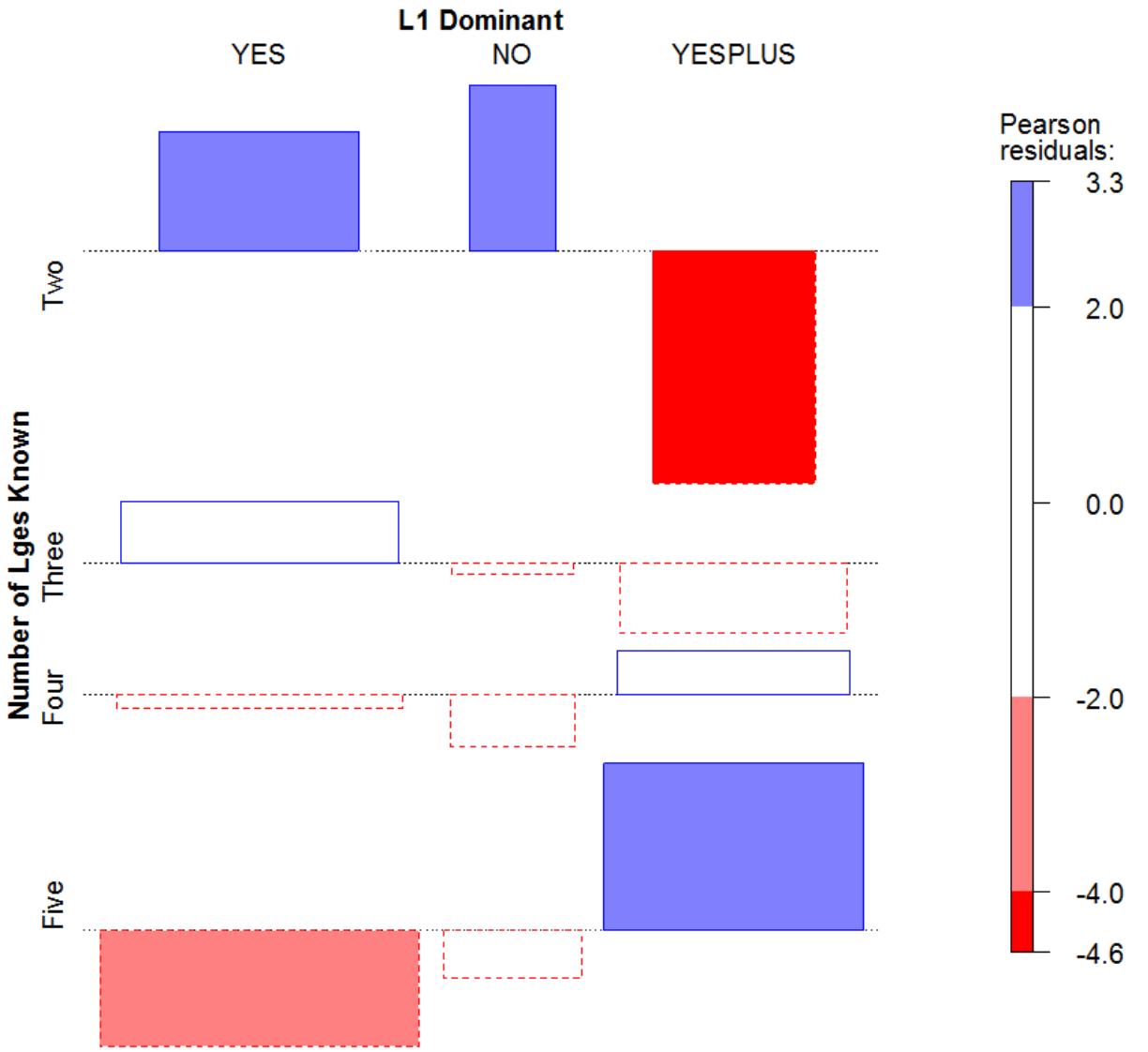

Figure 11 Association plot with two categorical variables (Dewaele and Pavlenko data).

The Pearson residuals plot to the right uses both color and saturation to indicate inferences that can be made from the data. The most saturated hue (the one below −4.00) for the red (dotted line) allows you to identify areas where the null hypothesis that the variables are independent can be rejected (Meyer, Zeileis, & Hornik, 2007). Residuals above 4 or below −4 indicate a difference that means the null hypothesis can be rejected. Residuals between 2 and 4 (and −2 and −4) are medium-sized and do not indicate a statistical rejection of the null hypothesis (Meyer,

Zeileis, & Hornik, 2007). Therefore, in Figure 11, the only cell that we can see is individually statistical is the one between persons with two languages and L1 dominance in more than one language (YESPLUS).

In order to create an association plot, the data need to be in the form of a contingency table. The beqDom file is a data frame, so a new object that is a contingency table must be created. The variables that will be used are catdominance (the answer for which language is dominant) and numberoflang (the number of languages a person knows). The commands I used to generate this plot are below.

```
install.packages(vcd)
library(vcd)
(DOM=structable(catdominance~numberoflang,data=beqDom))
assoc(DOM, gp=shading_Friendly, labeling_args=list(set_varnames=
c(catdominance ="L1 Dominant", numberoflang ="Number of Lges Known")))
```

Here is an analysis of these commands. The first step is to get the data into the correct format:

Here is the analysis for getting the data into the correct format:

```
(DOM=structable(catdominance ~ numberoflang,data=beqDom))
```

(. . .)	Putting the parentheses around the entire command will cause the table to appear without having to call the object; I have named the object DOM but it could be named anything—"x," "blah", etc.

structable( )	This vcd command creates a structured contingency table that automatically removes NA values and allows you to specify exactly which columns you want to use out of a certain data frame
catdominance ~ numberoflang	The formula inside the structable( ) command is exactly like the regression equation (see Chapter 7 in the book). Here it means that the outcome, dominance in a language, is modeled by the predictor variable, number of languages spoken
data=beqDom	Specifies the dataframe to be used

This command is the one that produces the actual graphic:

```
assoc(DOM, gp=shading_Friendly, labeling_args=list(set_varnames=
c(catdominance ="L1 Dominant", numberoflang ="Number of Lges Known")))
```

assoc ( )	The command for an association plot in the vcd package
gp=shading_Friendly	Specifies the type of shading on the plot; if left off, the association plot will consist of equally colored grey boxes. I like the Friendly shading because it distinguishes positive and negative with solid and dotted lines, making it better for black and white production. One other nice shading option is gp=shading_hcl
labeling_args=list(set_ varnames= c(. . .))	This argument allows you to specify exactly what names you want to give to your variables. Another useful labeling argument for renaming factor levels could be added with a

	comma after the argument to the left (still in the labeling_args argument though): set_labels=list(CatDominance =c("L1", "LX", "L1+LX")

Note that the mosaic plot could be used as well with three variables. The object created by structable would just add one more variable. The variable of sex is contained in the beqDom dataframe, so it is a simple matter to create a new object with three variables:

```
DOM3=structable(catdominance ~ numberoflang +sex,data=beqDom)
assoc(DOM3)
```

However, I think the mosaic plot or doubledecker plot is a better visual for this more complex situation, and so I will recommend you use those plots when you have three categorical variables.

Summary Creating association plots in R

1. Open the vcd library:
   library(vcd)
2. Put your data in the correct form:
   (DOM=structable(catdominance ~ numberoflang,data=beqDom))
3. Call the plot:
   assoc(DOM, gp=shading_Friendly, labeling_args=list(set_varnames=
   c(catdominance ="L1 Dominant", numberoflang ="Number of Lges Known")))

Mosaic Plots

Another plot that is quite similar to the association plot is the mosaic plot. It comes from the same package (vcd) as the association plot. In order to compare plots, we will use the same Dewaele and Pavlenko data in this section (beqDom).

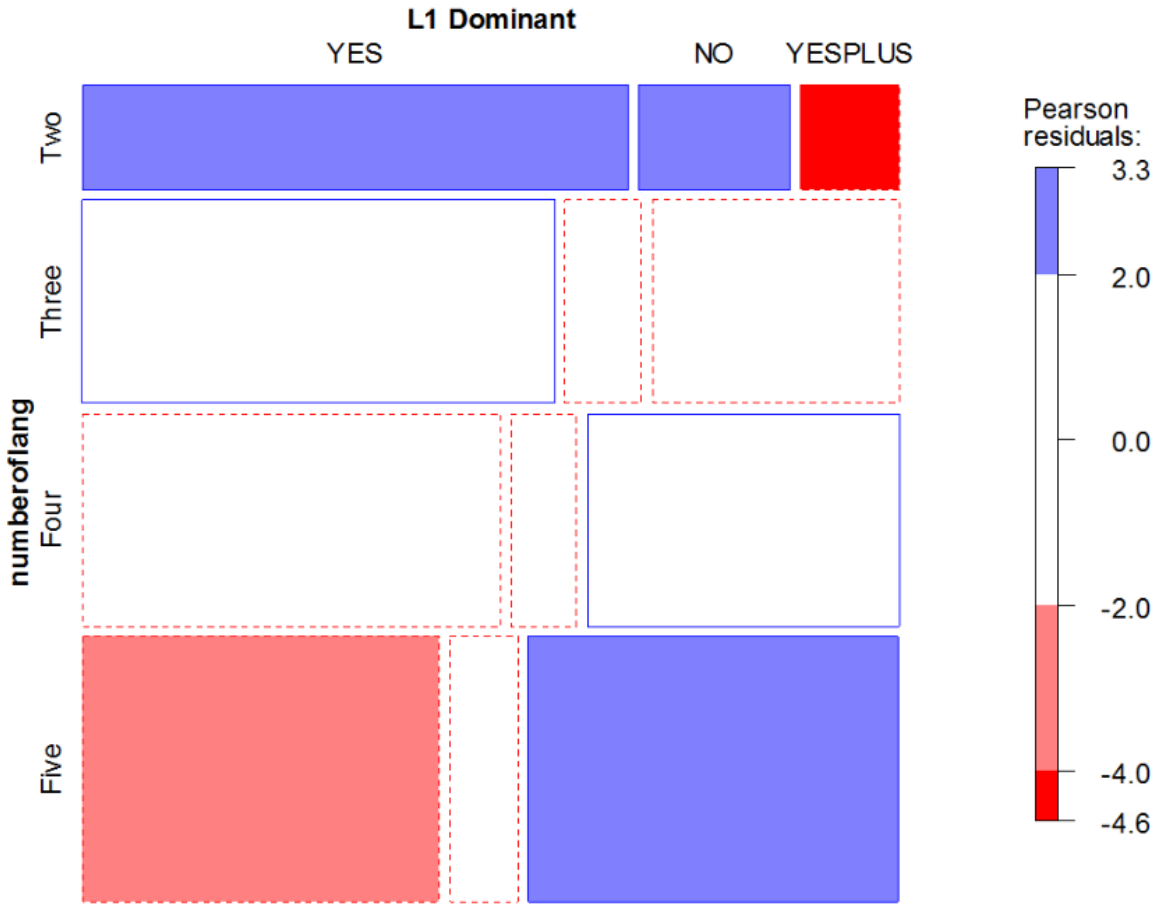

Figure 12 Mosaic plot with two variables (Dewaele and Pavlenko data).

The mosaic plot uses the same kind of reasoning as the association plot, with red-shaded areas (with dotted lines) indicating values that are less than expected, and blue-shaded areas (with solid lines) indicating values that are more than expected. The area of the boxes also gives an indication of its proportion to the whole, which the association plot did relative to the row but not to the whole dataset. Using the same values of shading as in the association plot (from Friendly), individual cells that violate the assumption of independence are more deeply colored.

The command needed to create this plot is mosaic ( ), and uses exactly the same arguments as were needed for the association plot, including the object I've named DOM, which has the data structured already:

```
library(vcd) #use this if you have not already opened the vcd package
mosaic(DOM, gp=shading_Friendly, labeling_args=list(set_varnames=
c(catdominance="L1 Dominant", numberoflangs="Number of Lges Known")))
```

See the section called "Mosaic plots" for more detailed information about each part of this command, as it is identical to the association plot syntax except for the name of the command (mosaic( )).

Mosaic plots are also good at representing the intersection of three categorical variables. Let's look at the Mackey and Silver (2005) data again (we saw it previously in this chapter in the section called "Barplots with Two Categorical Variables in SPSS"). Import the SPSS file into R with the name mackey. We will examine the relationship between experimental group and absence or presence of development in question formation on the delayed posttest as well as consider how pretest developmental level interacted with the other factors.

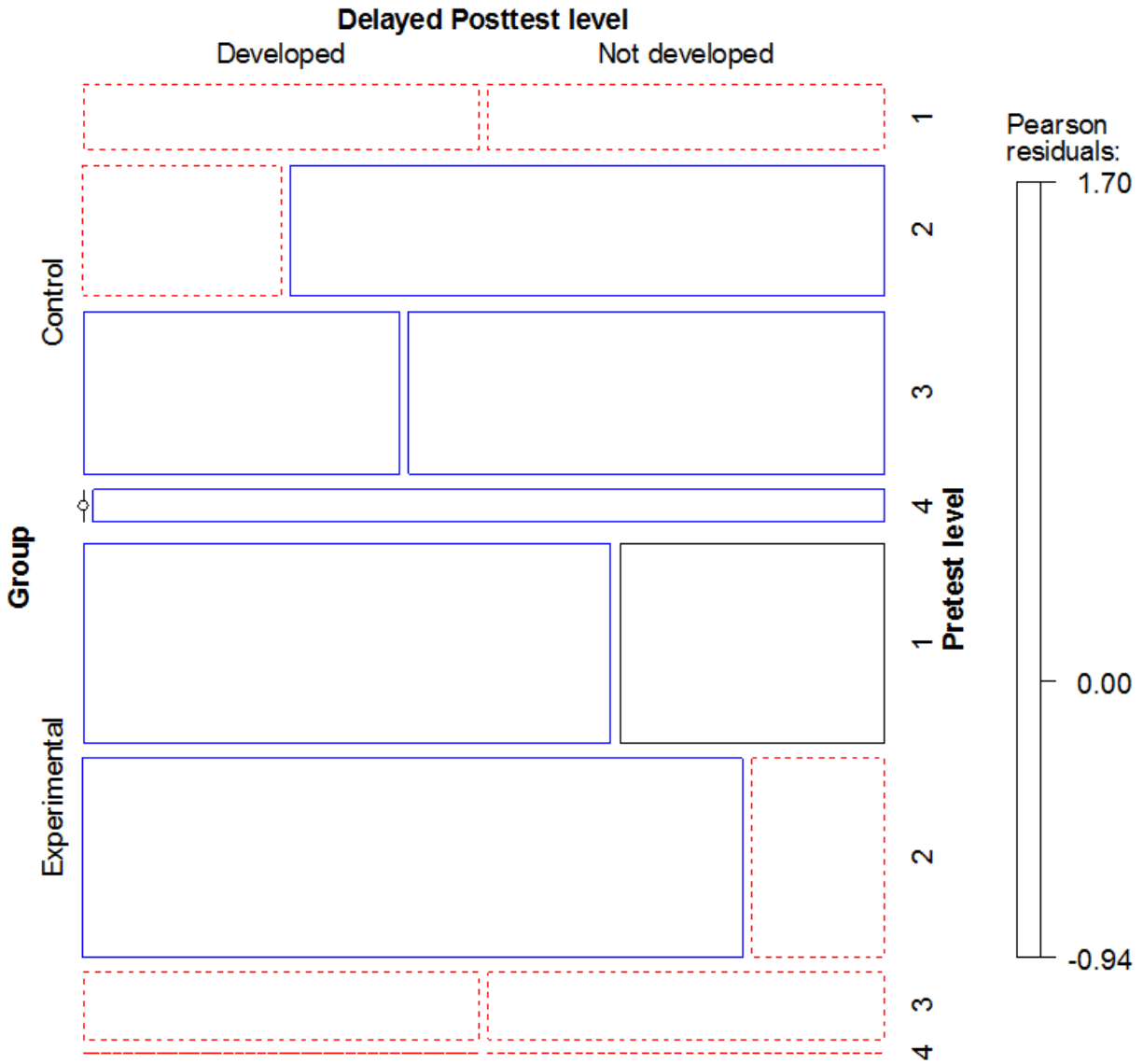

Figure 13 Mosaic plot with three variables (Mackey and Silver data).

Here is the R code that I used to make the three-variable association plot in Figure 13:

```
(DEV=structable(developdelpost ~ group + pretest, data=mackey))

mosaic(DEV, gp=shading_Friendly, labeling_args=list(set_varnames=

c(developdelpost="Delayed Posttest level", group="Group", pretest="Pretest level")))
```

Just looking at the pattern of results in Figure 13 with dotted lines being places where results are less than expected and solid lines those where results are more than expected, those in the control group who were in the lowest level of the pretest had fewer results than expected in both the "developed" and "not developed" categories, whereas those in the highest levels (3 and 4) of the experimental group had fewer results in both the "developed" and "not developed" categories than expected. Notice that the boxes for level 4 are both quite skinny and the one under the control condition has a dot over it. This means there were not enough participants in this category to evaluate it. The Friendly shading shows us that none of the trends were large enough to merit much attention, and we would expect no statistical differences here from the null hypothesis that all of the variables are independent.

Summary Creating Mosaic Plots in R

1. Open the vcd library:
   library(vcd)
2. Put your data in the correct form (enter your own data names where the words are in red):
   (DOM=structable(catdominance ~ numberoflang,data=beqDom))
3. Call the plot:
   mosaic(DOM, gp=shading_Friendly, labeling_args=list(set_varnames= c(catdominance ="L1 Dominant", numberoflang ="Number of Lges Known")))

Doubledecker Plots

The doubledecker plot is a very nice visual when you have three or more categorical variables and you want to visualize conditional independent structures. For this section I will illustrate with data for the Titanic (the dataset is called Titanic and is in R's base dataset so you should be

able to perform all of the commands in this section without importing any special dataset). This data, of course, has nothing to do with linguistics, but understanding this table in the context of something you may already understand may help with figuring out how this table works. The table looks at what factors influenced whether passengers on the Titanic survived. The Titanic dataset is a table and if you type Titanic into R, you will see that there are three predictor factors that are included in the Titanic survival tables: sex (male or female), age (child or adult) and class (1st, 2nd, 3rd or crew). With three predictors and one response variable (survival), we *can* include all of the variables in a mosaic or association plot,[1] but to my way of thinking the doubledecker plot is a more elegant solution.

The doubledecker plot in Figure 14 gives labels in layers underneath the table. The darker shadings are for those who survived (the skinny key that shows this is on the far right of the graph).

[1] If you want to see a mosaic plot utilizing all of the Titanic variables, run the following code (from p. 31 of Meyer, Zeileis & Hornik, 2007):
mosaic(Titanic, labeling_args=list(rep=c(Survived=F,Age=F)))

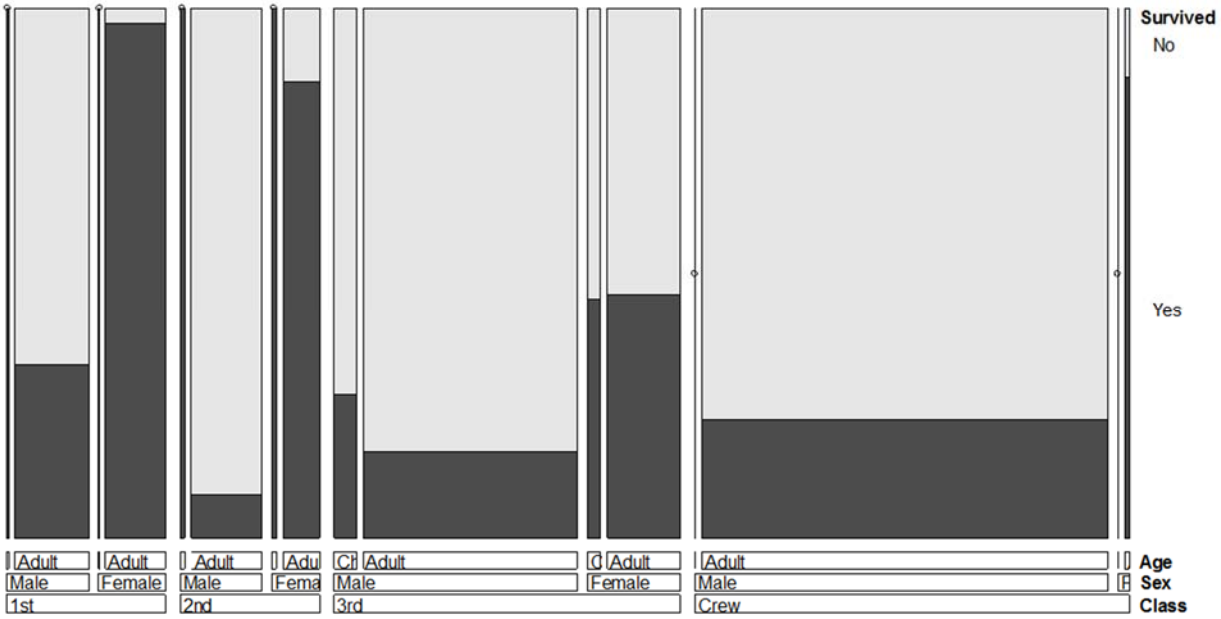

Figure 14 Doubledecker plot using 4 categorical variables (Titanic data).

As your eye moves from left to right over Figure 14, first you can see that proportionally more people in first class survived than in third class or crew (width of boxes represents proportions of numbers). Among first class passengers (almost all of whom were adults; the children are a slim black line to the left of the adults in first class), proportionally more females than males survived.

The command for this table is (after the vcd package is opened):

```
doubledecker(Survived~Class+Sex+Age, data=Titanic)
```

The order of the arguments will change the order in the table, so if you create your own doubledecker plot you might want to experiment with the order.

Summary Creating doubledecker plots in R

1 Open the vcd library:
library(vcd)
2 Call the plot:
doubledecker(Survived~Class+Sex+Age, data=Titanic)

Application Activity for New Techniques for Visualizing Data

1 Using the Dewaele & Pavlenko's BEQ.Swear.sav file (import as beqSwear; notice this is different file from the beqDom file we have been working with in the chi-square documents), create a mosaic plot examining the relationship between language dominance (L1dominance) as an outcome variable, and educational level (degree), and number of languages (numberoflanguages) as predictor variables. Educational level (degree) has 4 levels: PhD, MA, BA and A level. In order not to make the mosaic plot too messy, don't worry about changing the default labels on the graph, and instead add this argument to your mosaic plot: labeling_args=list(rep=c(degree=F) (it cuts out repetition of the educational level labels). Do you see any patterns to remark upon?

2 Using the same beqSwear file, create a doubledecker plot with educational level and number of languages as the predictor variables and language dominance as the response variable. You might want to play around with the order of the two predictor variables to see which seems better. Do any patterns stand out? Compare how this same data looks in the mosaic plot versus the doubledecker plot. Which do you prefer?

3 Use the Mackey and Silver (2005) dataset (mackey) and create an association plot that examines the development of questions on the delayed posttest (DevelopDelPost) as a function of development on the immediate posttest (DevelopPost) and experimental

group (Group). What stands out? Additionally, create a mosaic plot of the same data. Which do you prefer for this dataset—the association plot or the mosaic plot?

4 Use the fabricated dataset Motivation.sav (import as motivation). This dataset shows the distribution of YES and NO responses as to whether students are motivated to learn a second language (the class they are enrolled in) at the beginning (first) and end (last) of the semester. Data is further divided into responses given in each of five teachers' classes. Study this data and try to come up with the best visual plot you can for the data. What would be most appropriate—an association plot, a mosaic plot, or a doubledecker plot? Try all of them and see what looks best to you. Furthermore, you will have to decide which of the variables you call the outcome and which of the others you call the predictor variables.

Assumptions of Chi-Square

Chi-square is a non-parametric test, so it has fewer assumptions than parametric tests. However, chi-square does have some basic assumptions that must be met in order to make the results valid.

Meeting chi-square assumptions		
1 Independence of observations	Required?	Yes
	How to test assumption?	Make sure that number of participants will be equal to the number of observations in the contingency table (so there are no repeated measures)
	What if assumption not met?	1) If data is not independent, use a different statistical test (see earlier section for ideas of other tests that could be used) or rearrange the data into a format that is independent; 2) if you have paired data (before/after treatment measurement) you can use the McNemar test

2	Nominal data (no inherent rank or order)	Required?	Not strictly; interval or ordinal data may be collapsed into categories but this results in data loss and so cannot be recommended
		What to do if assumption is not met?	If you have ordinal data, then the ordering of the rows and columns will make a difference to the outcome; Howell (2002, pp. 311–312) says one approach to this type of data is to use the linear-by-linear association reported in the chi-square output. This number is more reliable as a chi-square measure with ordinal data than the Pearson chi-square
3	Data are normally distributed (this is the requirement that there are at least five cases in every cell, since normal distribution in the regular sense of the word can't happen with categorical data)	Required?	Yes
		How to test assumption?	1) Look for a violation of the assumption of expected frequencies. There should be at least five cases for each cell of the contingency table. This will be tested automatically by the computer software. Some authors say that in larger tables up to 20% of expected frequencies can be less than 5 (Field, 2005)
		What if assumption not met?	1) Don't worry too much about this (Howell, 2002, p. 159). With small sample sizes, your problem is likely to be a loss of power, so if you find a statistical difference then you don't need to worry about this. If you do not find a difference, then be aware that it may be because your sample sizes are too small and thus you lack power to find real differences; 2) use Fisher's exact test (Brace, Kemp, & Snelgar, 2003); 3) collect more data or collapse two categories into one in order to have enough counts in each cell

4 Non-occurrences must be included as well as occurrences	Required?	Yes; an example of this mistake would be counting the number of students with high motivation levels who earned a 2 (advanced) on the Oral Proficiency Interview scale in French, Russian, and Japanese, and then conducting an independent-groups chi-square on whether there was a relationship between motivation and high achievement, with a contingency table like this:

	French	Russian	Japanese	Total
Obs.	15	6	9	30
Exp.	10	10	10	30

	In counting only those students with high motivation, we would leave out the students we surveyed with non-high levels of motivation
How to test assumption?	Make sure that the number of participants in the study is equal to the grand total of observations in the contingency table
What if assumption not met?	Include the non-occurrences as well as the occurrences; in our example above, we'll need to know how many people in total from each language group were surveyed:

	French	Russian	Japanese	Total
High	15	6	9	30
Non-high	63	6	19	88
Total	78	12	28	118

With only occurrences included, it seemed that more students of French who were highly motivated became advanced speakers, but that was before we knew how many speakers in total were tested. In fact, 50% of the Russian speakers with high motivation scored highly, while only 19% of the French speakers with high motivation scored highly.

I would like to emphasize one more time that the first assumption of chi-square listed here is often violated in research studies I have seen from applied linguistics and sociolinguistics. To give another example of this type of violation, assume that we do a study on the number of times our participant supplies a third person singular –s on a verb appropriately. For example, say we have five participants and they were observed for how many times they put a third person –s on the 20 obligatory instances in a cloze test. We want to know if our participants supplied the –s more times than would be predicted by chance (so we are interested in a goodness-of-fit chi-square). We might have the following type of table:

	Supplied –s	*Omitted –s*	*Total*
Participant 1	15	5	20
Participant 2	7	13	20
Participant 3	12	8	20
Participant 4	3	17	20
Participant 5	10	10	20
Total	47	53	100

Table 14 Contingency table for made-up data regarding suppliance of third person singular –s on verbs.

If we now performed a chi-square for goodness of fit, assuming that each choice is equally likely, $\chi^2 = .36$ on 1 degree of freedom, which has an associated *p*-value of .55, meaning there is no evidence here that our participants performed any differently from chance. However, we can clearly see that some participants appeared to do much better or worse than chance (such as Participants 1 and 4). The problem with using chi-square in this case is that we have 100 pieces of data in our contingency table but only five participants. Each person has contributed more than once to a cell and thus the data are not independent. This type of data analysis is similar to that shown in Scenario Two of the section at the beginning of this paper entitled "Other Situations that May Look like Chi-Square."

Chi-Square Statistical Test

There are two types of chi-square statistical tests. One is used when you have only one variable, and want to examine whether the distribution of the data is what is expected. This is called the

one-way goodness-of-fit chi-square.

The other type of chi-square is used when you have two variables, with two or more levels each. This is used when you want to examine whether there is a relationship between the variables. This is called the two-way group-independence chi-square.

To look at a goodness-of-fit chi-square test I will use the dataset extrapolated from the appendix of Geeslin and Guijarro-Fuentes (2006), although I am not answering the questions they asked in their study. Instead, I will look only at three items to determine whether the choices (of the verbs *ser* and *estar* or both verbs equally) are distributed equally. If you have worked through the previous sections of the paper you will know that this study involved asking participants which verb they would use in various situations.

One-Way Goodness-of-Fit Chi-Square in SPSS

Use the dataset GeeslinGF3_5.sav and open the menu sequence ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > CHISQUARE. You will see a dialogue box like that in Figure 15. To run the chi-square test, move the variables you'd like to analyze over to the TEST VARIABLE LIST box and press OK. I looked at the test for Items 3, 4 and 5.

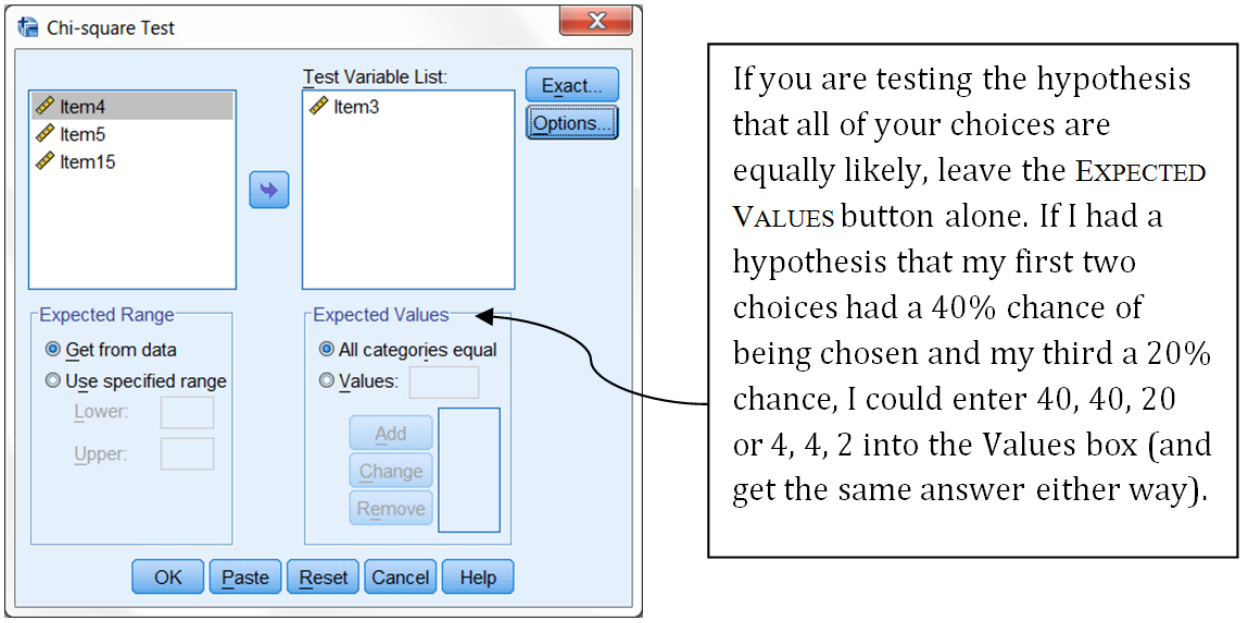

Figure 15 Dialogue box for a one-way goodness-of-fit chi-square test in SPSS.

The first part of the print-out from the one-way chi-square is seen in Table 13. First we see the summary of observed and expected counts for Items 3 through 5. The chi-square was done using the hypothesis that all categories were equally likely to have been chosen, as can be seen in the "Expected N" column (all of the choices have identical expected N). Notice that for Item 4, where there was no variation (all of the speakers chose the verb *ser* for this situation), a chi-square cannot be performed.

Item3

	Observed N	Expected N	Residual
Estar	13	6.3	6.7
Ser	4	6.3	-2.3
Both	2	6.3	-4.3
Total	19		

Item4

	Observed N	Expected N	Residual
1	19	19.0	.0
Total	19[a]		

a. This variable is constant. Chi-Square Test cannot be performed.

Item5

	Observed N	Expected N	Residual
1	13	6.3	6.7
2	5	6.3	-1.3
3	1	6.3	-5.3
Total	19		

Table 13 Output on Descriptives from a one-way goodness-of-fit chi-square test.

The next part of the printout (Table 14) shows the results of the goodness-of-fit chi-square test.

The "Test Statistics" box shows that the chi-square statistic for Item 3 is 10.8 at 2 degrees of freedom, which gives a very low probability (p = .004). Item 5 likewise has a very low p-value. The null hypothesis is that all choices are equally likely, so with a p-value below p = .05, we can conclude that the choices are not all equally distributed. Notice that the chi-square does *not* tell us whether two of the choices (of the verb *ser, estar*, or both) are the same and one is different, or if all three are different. However, in this case for all three items the number of native speakers of Spanish who chose *estar* is quite large (*estar* = 13 for both Item 3 and 5), so if Items 3 and 5 are statistical this means that statistically native speakers are more likely to choose *estar* than any other answer for these items (for Item 4 this cannot be tested statistically because it is the only choice, but logically this must be true for Item 4 as well).

Test Statistics

	Item3	Item5
Chi-Square[a]	10.842	11.789
df	2	2
Asymp. Sig.	.004	.003

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.3.

Table 14 Chi-square test results output from a one-way goodness-of-fit chi-square test.

Notice that this type of testing does not give us a confidence interval, which is what we'd rather have than a p-value. For the hypothesis that all choices are equally likely (or any other type of configuration that we'd like to propose) I do not know of a way to get SPSS to return a confidence interval for testing this hypothesis.

This is the way to run the test if you have all of the original data typed in. But if you know the summary numbers and don't want to have to type everything in, there is another way to analyze the data, and this is by telling SPSS that the number in the dataset is a summary number, not the score for a specific individual on that row. Look at Figure 16, which shows an SPSS dataset with summarized data, and the dialogue box for the menu sequence DATA > WEIGHT CASES. By weighting cases in SPSS you tell the statistical program to use the number 13, say, as a summary and not as the result of the individual in Row 1.

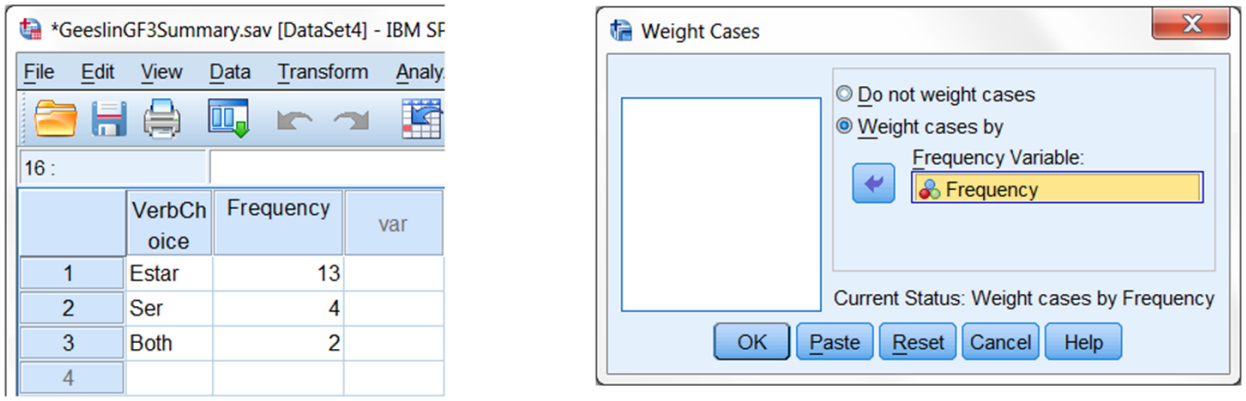

Figure 16 Using summary data instead of each row representing one individual's data in SPSS, and weighting the data.

After weighting the data the chi-square analysis can be done as usual and as noted previously, will give the exact same results as seen in Table 14.

Summary: Performing a One-Way Goodness-of-Fit Chi-Square in SPSS

1 Choose ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > CHISQUARE.
2 Put variable in TEST VARIABLE LIST box and press OK. No further adjustments are made if you are testing the hypothesis that all of the choices are equally likely.

One-Way Goodness-of-Fit Chi-Square in R

In this section I use the dataset GeeslinGF3_5.sav, imported into R as GGF3. To call for the goodness-of-fit chi-square test in R Commander, you follow the same sequence as you would to get a numerical summary: STATISTICS > SUMMARIES > FREQUENCY DISTRIBUTIONS. Pick the variable you want to test and tick the box that says "Chi-square goodness-of-fit test." Note that the only variable that is available is Item 3, because in SPSS I made the variable a factor by adding values for the 3 choices in the Variable View. In a moment, I will show the way you could still perform the chi-square test even if it does not show up as a factor in R Commander. Back to the original dialogue box, after you press OK, an additional box will pop up, as shown in Figure 17.

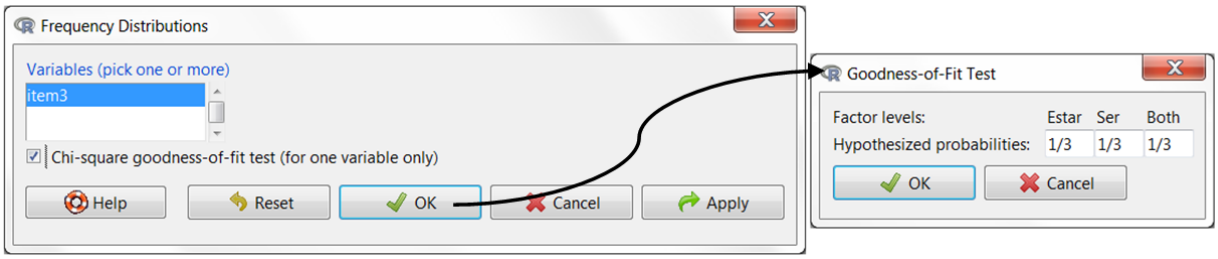

Figure 17 Dialogue boxes for a one-way goodness-of-fit chi-square test in R Commander.

The null hypothesis that will be tested is that every category is equally likely. Since there are 3 categories for this dataset, that means each category has a 1/3 chance of being selected. The

"Goodness-of-fit" test dialogue box automatically assumes that each level has an equal chance unless you change this ratio manually.

The output for this test is shown here:

```
        Chi-squared test for given probabilities

data:  .Table
X-squared = 10.8421, df = 2, p-value = 0.004422
```

The chi-square test shows that the probability of getting this distribution of answers is highly unlikely (p = .004) if the null hypothesis is true. We therefore conclude that the L1 speakers of Spanish had a preference for some of the answers over others, although the chi-square test cannot tell us which answer was preferred. The chi-square does not tell us whether two of the choices (of the verb *ser, estar,* or both) are the same and one is different, or if all three are different. For that, however, look to the counts. They clearly show that the native speakers favored the verb *estar* in this particular situation.

```
Estar   Ser  Both
   13     4     2
```

The R code for the chi-square test is:

```
.Table <- table(GGF3$item3)
.Probs <- c(0.3,0.3,0.3)
chisq.test(.Table, p=.Probs, correct=TRUE)
```

.Table <-	Creates a table of counts for the choices

table(GGF3$item3)	
.Probs <- c(0.3, 0.3, 0.3)	This sets up the probability for the choices
chisq.test( )	Performs a chi-square test for group independence as well as goodness-of-fit
.Table	The contingency table with one variable is the source of the data for the chisquare test (xtabs is normally used if there is more than one variable)
p=.Probs	Tests the hypothesis that the distribution of the data in .Table is the same as the distribution created in the .Probs variable.
correct=TRUE	Specifies that the continuity correction should not be applied. Change to FALSE if you have a 2x2 table.

By default, the probability of each choice is set to be equal, which is what I wanted in this case. However, if you would like to specify different probabilities using R code, you will need to create a new object where the probabilities sum to 1, and then include that object in the arguments to the chi-square command. For example, to test the hypothesis that there was a 40% chance of picking *estar* or *ser* and only a 20% chance of picking both, you would specify your *p*-value like this:

```
.Probs=c(.4,.4,.2)
```

The R Commander dialogue box didn't show the data for Items 4 and 5 because they weren't factors. We could make these variables factors (see Appendix A for information about creating a factor), make tables of them, and then run the chi-square test on them. For Item 4, however, the only answer chosen was the first one, *estar*. Running the chi-square statistic will result in an error message that 'x' (your vector of numbers) must have at least 2 elements. However, it is hardly necessary to run a formal chi-square for this item, since it is clear there is little probability that all participants would choose *estar* if each of the 3 choices were equally likely. For item 5, here is how you could change it into a factor and then run the chi-square test if the . Probs file is not removed:

```
Item5=factor(rep(1:3, c(13,5,1))) #Create a factor where choice #1=13 times, #2=5 times, #3=once
.Table<-table(Item5)
chisq.test(.Table, p=.Probs)
```

The result is that $\chi^2 = 11.79$, df = 2 (because there are 3 choices), and $p = .0028$. Here the native speakers of Spanish chose *estar* 15 times, so it appears that in all three items (3 through 5), the native speakers statistically preferred the choice of *estar* over the other choices.

If, instead of a dataset that contained the choices of each individual person in a separate row you simply knew the summary data of how many times each possibility was selected, you could also easily create a table from the summary data. For example, we know that for Item 3 *estar* was

chosen 13 times, *ser* 4 times and both 2 times. To make a table directly out of this summary data, type:

```
.Table <- table(rep(1:3, c(13,4,2)))
```

This doesn't provide nice labels but will work for the analysis.

Summary Performing a One-way Goodness-of-fit Chi-square

1. In R Commander, choose STATISTICS > SUMMARIES > FREQUENCY DISTRIBUTIONS. Choose your variables and tick the box that says "Chi-square goodness-of-fit test." If all choices are equally likely, leave the default probabilities in the next dialogue box alone. If not, type in your hypothesized probabilities.
2. In R, use this code (add in your own data where the typing is red):

```
.Table <- table(GGF3$item3)
.Probs <- c(0.3, 0.3, 0.3)
chisq.test(.Table, p=.Probs, correct=TRUE)
```

If desired you can specify the probabilities of each choice, making sure the numbers sum to 1: `.Probs<-c(.4, .4, .2)`

Two-Way Group-Independence Chi-Square in SPSS

The chi-square test for independence uses two variables and tests the null hypothesis that there is no relationship between the two variables. In order to illustrate a group-independence chi-square test (this is also called a multidimensional chi-square, although note that only two variables can be tested at one time) I will use data presented in Mackey and Silver (2005). In doing a group-independence chi-square, the null hypothesis for this dataset would be that there is no relationship between group membership (experimental or control group) and improvement in question-formation level.

To test the null hypothesis, choose ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS. This is the same procedure that was used in the section called "One-Way Goodness-of-Fit Chi-Square in R" with different data in order to obtain the crosstabs tables—the only additional step is to open up the STATISTICS button. In the Crosstabs dialogue box seen in Figure 18, I put the EXPGROUP (splitting) variable in the "Column(s)" box, and the delayed posttest developmental categorization (DEVELOPDELPOST) into the "Row(s)" box. I also ticked the "Display clustered bar charts" box to get barplots to appear in the output. In the dialogue box shown when the STATISTICS button is pressed (also in Figure 18), I ticked the box labeled "Chi-square" to call for the test, and then the box labeled "Phi and Cramer's V" in order to ask for an effect size. Note that, if you wanted to do a McNemar test (when you have paired data), this box would be the place to ask for it.

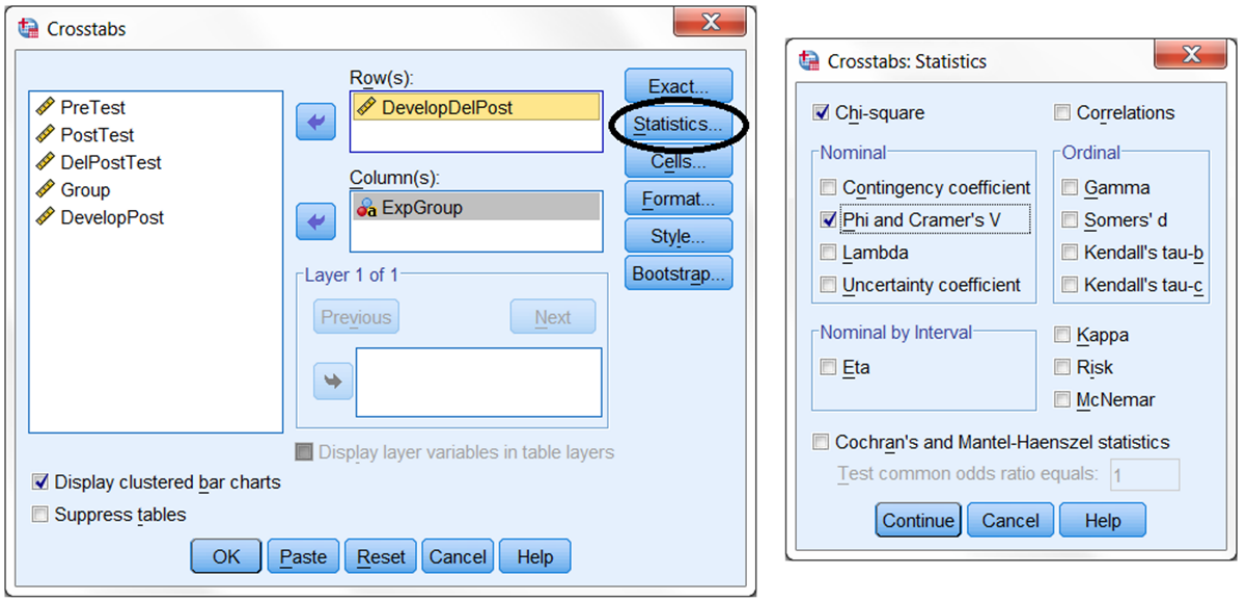

Figure 18 Dialogue boxes for a two-way group-independence chi-square in SPSS.

Also open the CELLS button from the main Crosstabs dialogue box and click the boxes that I have clicked on in Figure 19. SPSS also provides the possibility of calculating bootstrapped confidence intervals for the effect sizes, although not for the chi-square test itself, so open the BOOTSTRAP button and make those choices shown in Figure 19.

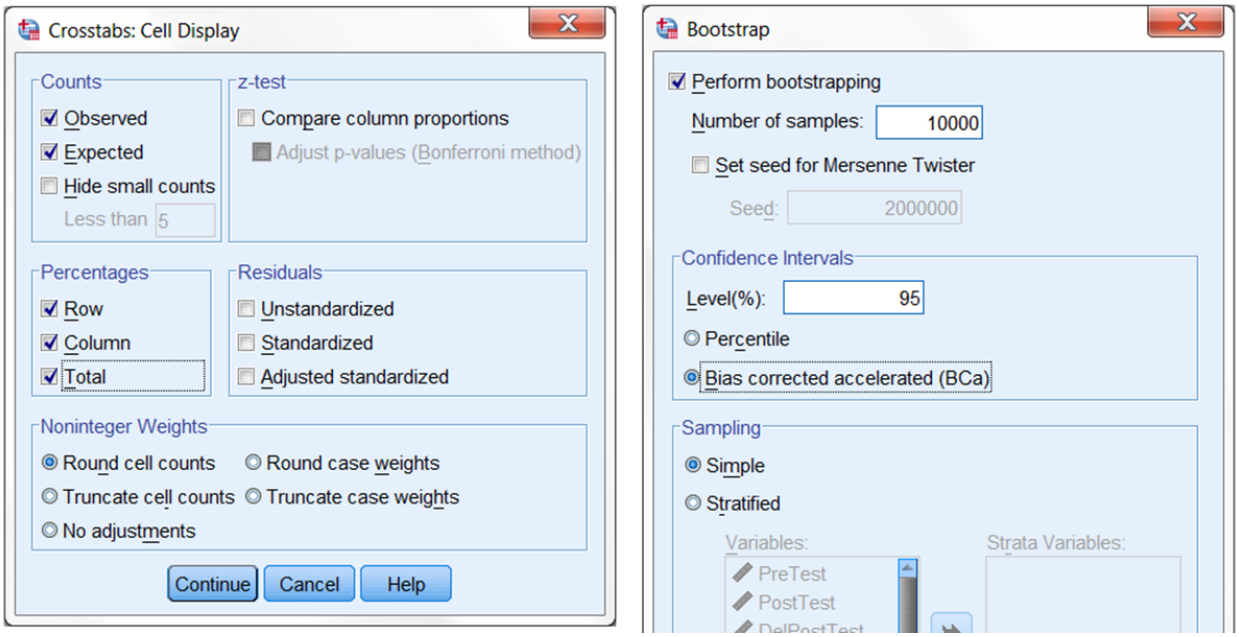

Figure 19 Dialogue boxes for CELLS and BOOTSTRAP buttons in the group-independence chi-square in SPSS.

A button we are not opening is the EXACT button (its availability depends on whether you have purchased the Exact Tests option). The Exact method of calculating the *p*-value of the chi-square test can be used when you have less than 5 cases in every cell, as noted in the section on assumptions of chi-square if you assume that data are normally distributed. This button also provides a Monte Carlo method for calculating an exact *p*-value.

The first part of the output, shown in Table 15, gives a summary of how many cases are included. This table shows that no cases are missing, which is good. If you did have missing data, you could impute your data (see Section 1.5 in the book).

Table 15 Case Processing Summary Table in Chi-Square Output.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
DevelopDelPost * ExpGroup	26	100.0%	0	0.0%	26	100.0%

The next part of chi-square test output gives the crosstab table with expected counts and all of the row and column percentages. Table 16 shows just the first part of this crosstab, since the section titled "Summary Tables for Group-Independence Data (Crosstabs) in SPSS" in this document already examined these types of summary data, although not for this dataset. Do not ignore your summary data though—you should be thoroughly acquainted with the details of your data, not just the results of statistical tests. The two pieces of information should make sense together, and often you can get a good sense just from the summary data as to how your hypothesis will be answered. For example, Table 17 shows that the count of those who developed in the experimental group (N = 10 out of 14 in the experimental group) is much greater than the count of those who were in the control group and who developed (N = 4 out of 12 possible). From just the summary data we might suspect that the chi-square test will be statistically significant and show that there was an interaction between group membership and

development in question formation.

Table 16 Crosstab Summary in Chi-Square Output.

DevelopDelPost * ExpGroup Crosstabulation

			ExpGroup		
			Con	Exp	Total
DevelopDelPost	Developed	Count	4	10	14
		Expected Count	6.5	7.5	14.0
		% within DevelopDelPost	28.6%	71.4%	100.0%
		% within ExpGroup	33.3%	71.4%	53.8%
		% of Total	15.4%	38.5%	53.8%
	Not developed	Count	8	4	12

The next part of the output, shown in Table 17, gives the chi-square tests. The statistic normally reported for the chi-square test is the first row of Table 17, the Pearson chi-square. The χ^2 value is 3.77 with 1 degree of freedom (because there are two variables), with a *p*-value just barely above the .05 statistical level ($p = .052$). Since we know that *p*-values are notoriously unreliable (see Section 4.1.4 of the book for more information), the fact that this *p*-value is not extremely small will make us wary of trusting this result one way or the other. It will be more important for us to look at the effect size.

Table 17 Chi-Square Test Results in Chi-Square Output.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.773[a]	1	.052		
Continuity Correction[b]	2.396	1	.122		
Likelihood Ratio	3.862	1	.049		
Fisher's Exact Test				.113	.060
N of Valid Cases	26				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.54.

b. Computed only for a 2x2 table

Footnote a is important to notice at once, because it informs us whether expected counts are less than 5, which would be problematic. None are in this case. The likelihood ratio is an alternative test to the chi-square, which also uses the χ^2 distribution (it is also called a *g*-test). Howell (2002) asserts that this test is gaining in popularity and that its results should be equivalent to the chi-square when sample sizes are large. The linear-by-linear association test assumes that both variables are ordinal. Report this if your variables have inherent rank.

> Tip: Because this is a 2 × 2 table, meaning that each variable only has 2 levels, the chi-square output shown in Table 17 contains two more lines than you would see in any other case. One is the Continuity Correction, which some authors argue should be used for a 2 × 2 contingency table. Howell (2002) gives good arguments against it, and I would recommend against it as well. The other line is Fisher's Exact Test. This can be used when cells do not have the minimum number of counts.

The next part of the output is listed under "Symmetric Measures" and gives the effect sizes. Howell (2002) notes that good measures of effect size to use for the chi-square test are phi (φ), which is used for 2 × 2 contingency tables with only two levels per each variable, and Cramer's V, which is used for tables larger than 2 × 2 (when there are more than two levels per variable). I will not reproduce the output here, but if you have followed along with me you will see that Phi = -.381 and Cramer's V = .381. **Phi** (or **Cramer's V**) is a percentage variance effect size (from the *r*-family) and would indicate that the experimental treatment accounted for 38% of the variance in the data. Phi can only be used as an effect size when there is a 2 × 2 contingency table, but Cramer's V is an extension for tables with more levels. 38% is a modest effect size but does show some effect, but there is a wrinkle in straightforwardly concluding that the experimental group helped students progress more than the control group. The bootstrapped results for the effect size, shown in Table 18, show that the effect size is not very reliable. The

95% confidence interval shows that the size of the effect could be as low as .01 or as high as .70 (ignoring the negative sign on the Phi line). Although this is not zero, this is a very wide interval and shows us that the effect size of .38 cannot be assumed to be correct.

Table 18 Bootstrap for effect size in Chi-Square Output.

Bootstrap for Symmetric Measures

			Bootstrap[a]			
					BCa 95% Confidence Interval	
		Value	Bias	Std. Error	Lower	Upper
Nominal by Nominal	Phi	-.381	.001	.188	-.701	-.006
	Cramer's V	.381	.003	.178	.031	.728
N of Valid Cases		26	0	0	.	.

a. Unless otherwise noted, bootstrap results are based on 10000 bootstrap samples

Although the descriptive statistics led us to believe that we would find an association between group membership and development in question formation, the statistical analysis here shows us that this study would not be sufficient to conclude with any confidence that this was true. Because the confidence interval is so wide it is hard to say from this study what the true effect of group membership was, and so I would conclude in this case that it is not possible to confidently say that group membership is related to development in question formation.

Summary: Performing a Two-Way Group-Independence Chi-Square

1. ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS. Tick the "Display clustered bar charts" box if you want to see a barplot.
2. Open the Statistics button and tick the "Chi-square" and "Phi and Cramer's V" boxes.
3. Open the Cells button and tick the "Expected values" and all of the boxes under "Percentages."
4. Open the Bootstrap button and tick the "Perform bootstrapping" box. Change the "Number of samples" to 10,000 and the confidence intervals to "BCa."

Two-Way Group-Independence Chi-Square in R

The chi-square test for independence uses two variables and tests the null hypothesis that there is no relationship between the two variables. To illustrate the use of this test I will use the Dewaele and Pavlenko data from their Bilingual Emotion Questionnaire (BEQ.Dominance, imported into R as beqDom). Dewaele and Pavlenko received answers from 1578 multilinguals to the question as to what language they considered their dominant language. They categorized their answers into yes (L1-dominant), no (L1 not dominant), or yesplus (L1 plus more dominant). In doing a group-independence chi-square, the null hypothesis for this dataset would be that there is no relationship between the number of languages that a person speaks and their language dominance.

To perform the two-way chi-square in R Commander, choose STATISTICS from the drop-down menu, then CONTINGENCY TABLES as shown in Figure 20. At this point, you will need to decide which choice in this menu is appropriate for your data. Use the TWO-WAY TABLE choice when your data are in raw input, not summary form. This is when you have as many rows as you do participants or items. Use the CONTINGENCY TABLES > ENTER AND ANALYZE TWO-WAY TABLE choice when you have summary data. In both cases, you can only run the chi-square test on two variables. When you have more than two variables, you can use the MULTI-WAY TABLE choice,

but you cannot perform any statistical tests on the data; instead, you will see descriptive statistics of two tables split by the third variable that you specify, as was shown in the section "Summary Tables for Group-Independence Data (Crosstabs) in SPSS." If you have three variables you could experiment with order to find what two variables you want to test at a time.

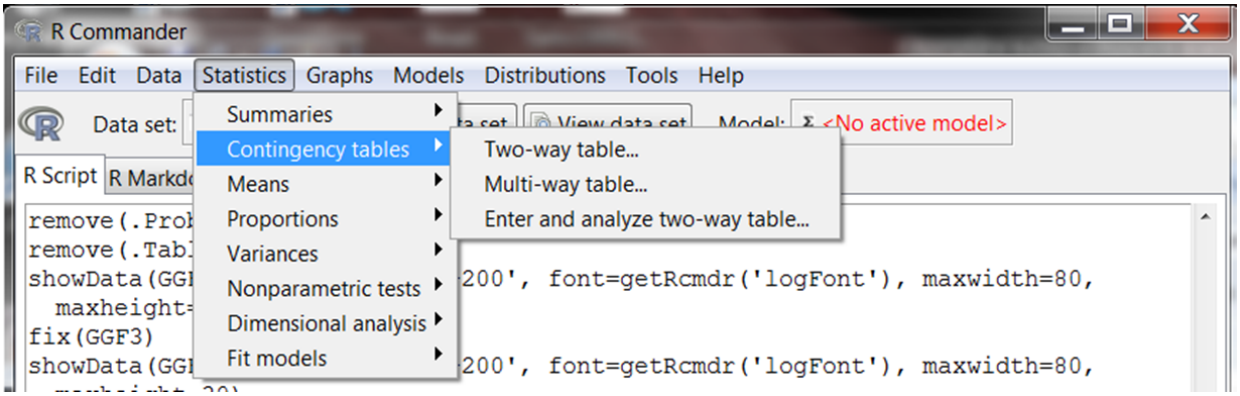

Figure 20 Opening up the two-way group independence chi-square command in R.

The data in beqDom is not in summary form, so I chose TWO-WAY TABLE. We have previously seen the dialogue box that comes up, which lets us choose the row and column variables, whether we want to see percentages of the data, and which hypothesis tests we can pick (this is shown again in Figure 21). For the statistical test, make sure the box under HYPOTHESIS TESTS called "Chi-square test of independence" is checked. The other choice you might want here is "Fisher's exact test" if you had a problem with small sample size.

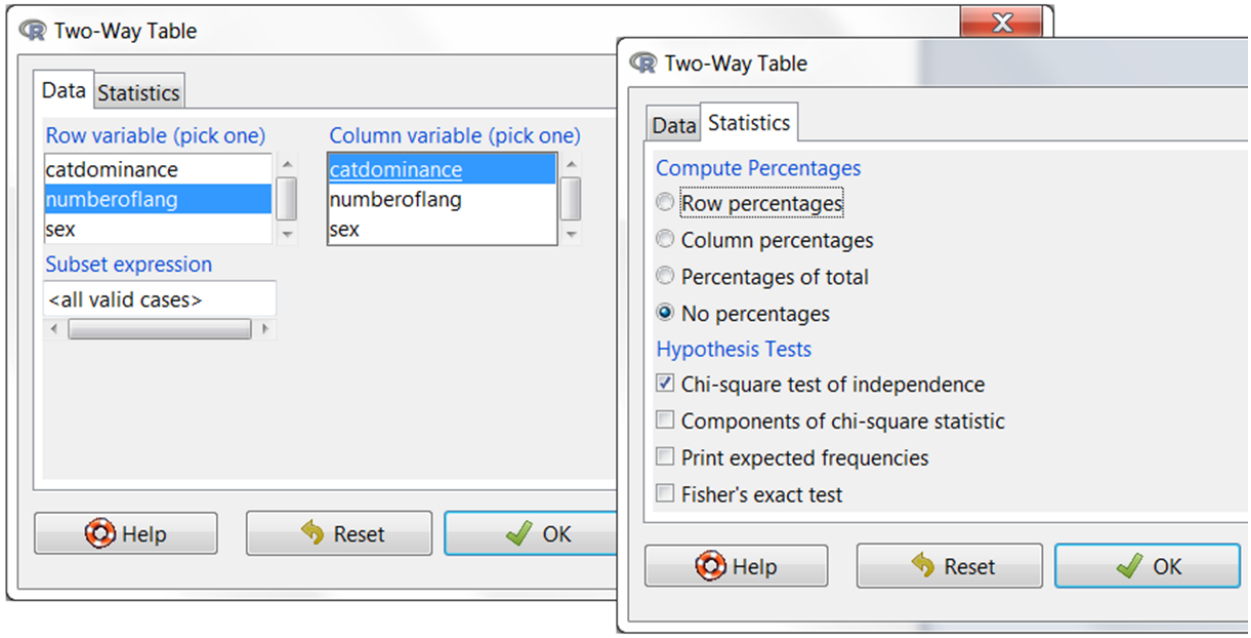

Figure 21 Dialogue boxes for a two-way group-independence chi-square in R.

I did not choose anything besides the chi-square test, so besides the crosstabs summary count (seen previously in the section called "Summary Tables for Goodness-of-Fit Data in R") my output looks like this:

```
        Pearson's Chi-squared test

data:  .Table
X-squared = 59.5807, df = 6, p-value = 5.476e-11
```

The result of the chi-square is that the χ^2 value is very large (59.6) on 6 degrees of freedom, and so the *p*-value is quite small (p=.000000000055). We reject the null hypothesis and say that there is a relationship between the number of languages someone speaks and which language they say is dominant. Note that the test itself does not tell us anything about the nature of this relationship with regard to specific details about the levels of the variables.

For an explanation of the relationship, we can look at the residuals. Actually, we have already seen a plot of the residuals in the Association plot from Figure 11. We can call for these residuals using R code and interpret them in light of the fact that our main chi-square test is statistical. I'll return to this momentarily after we have looked at the R code.

Along with looking at the output, don't forget to check R Commander's message box for a warning if any of the expected frequencies are too low. This was not a problem with the Dewaele and Pavlenko data (this is a huge dataset!), but Figure 22 is an example where the expected frequencies are too small.

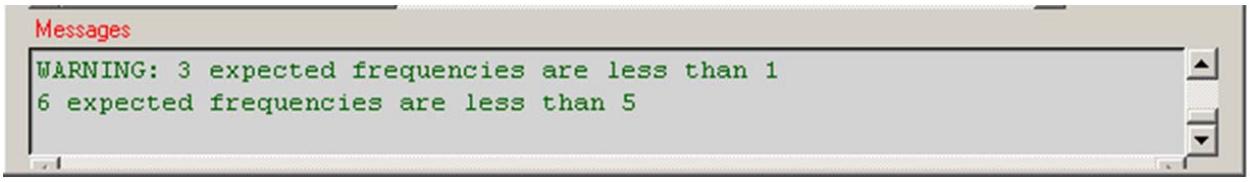

Figure 22 R Commander's warning about low expected frequencies.

R code that can generate the chi-square test of group independence is:

```
(.Test<-chisq.test(xtabs(~catdominance+numberoflang,data=beqDom),
correct=FALSE))
```

( )	Putting parentheses around the command will print the results without any additional work
.Test<- chisq.test( )	Putting the results of the chi-square test into an object named .Test means we will be able to call for components of the test later (see

	below for residuals)
chisq.test	Performs a chi-square test for contingency tables as well as goodness-of-fit
xtabs	Creates a contingency table of the specified variables
(~catdominance + numberoflang)	The variables that are tested in this two-way chi-square
data=beqDom	Specifies the dataset to be used
correct=FALSE	Specifies that the continuity correction should not be applied. The correction is generally only applied when there are only two levels for each of the two variables, which is called a 2 × 2 table, but Howell (2002) argues that you should not use it even in that case. The default setting is TRUE, so be sure to include this argument

In understanding the chi-square test in more detail, we can look at the standardized residuals. If the value of the residual is greater than $\pm$ 1.96 then the result is statistical at the $p < .05$ level; if the residual is greater than $\pm$2.58 then the result is statistical at the $p < .01$ level, and if the residual is greater than $\pm$3.29 then the result is statistical at the $p < .001$ level (Field, Miles & Field, 2012). The Association plot (Figure 11) and Mosaic plot (Figure 12) of this data seen earlier in the paper show us a pattern of results however, and providing readers with a plot like this may be preferable to providing the specific residual numbers (I am reprinting the Association plot of Figure 11 here as Figure 23 so you can compare it to the residual numbers).

Nevertheless, we can draw out the specific residuals with this command, providing we have put the results of the chi-square test into an object named .Test:

```
> .Test$residuals
             catdominance
numberoflang        YES         NO    YESPLUS
       Two    2.3527617  3.2511987 -4.6026463
       Three  1.2195719 -0.2229896 -1.3733499
       Four  -0.2722572 -1.0268126  0.8777912
       Five  -2.2952267 -0.9189132  3.2948414
```

The size and the sign (positive or negative) of the residuals give us information about departures from expected values.

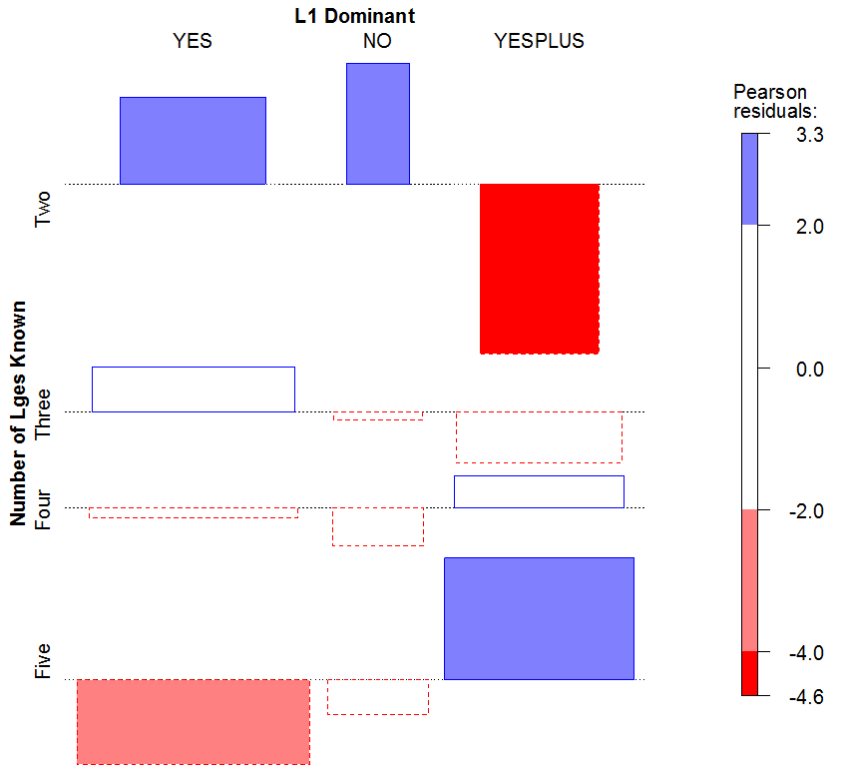

Figure 23 Association plot of Dewaele & Pavlenko data (reprised).

The first residual, those with two languages who are dominant in their L1, is 2.35, and is larger than would be expected if we thought each category should be equally likely. Therefore, it is found above the line in a shaded blue box in the Association plot. On the other hand, of those with two languages who say they are dominant in their L1 plus another language, the residual is -4.60, and is much smaller than would be expected if every choice were equally likely. The Association plot shows it below the line, shaded bright red because it is greater than 4.00. Meyer, Zelis and Hornik (2003) state that "the main purpose of the shading is not to visualize significance but the *pattern* of deviation from independence" (p. 4, emphasis original).

Overall, we can see a number of patterns. For example, for bilinguals, the pattern shows that they are much more likely to be dominant in only one language, although this may be their L1 (L1 dominant=YES) or L2 (L1 dominant=NO). Another pattern we can see is that there are fewer people than would be expected who say they are dominant in more than one language if they only know two languages, while there are more people than would be expected who say they are dominant in more than one language if they know five languages. The barplot (Figure 10) was also informative in showing that the number of people who were dominant in more than one language seemed to increase monotonically as the number of languages known increased. Suffice it here to say that plots should always be used to augment the information given in a chi-square test. Residuals can additionally be used to understand the pattern of what is happening in the data.

Going back to the choice of entering data directly into a contingency table (STATISTICS > CONTINGENCY TABLES > ENTER AND ANALYZE TWO-WAY TABLE in Figure 20), let's say that you

were looking at a contingency table in a published report but did not have the raw data. It would be easy to perform a chi-square test (but not draw up an association or mosaic plot!) if you only had summary data. An example of this is a study by Shi (2001) in which the author asked whether teachers whose L1 was English emphasized different aspects of essay grading from teachers whose L1 was Chinese. The teachers graded holistically, and then could list three comments, in order of their importance. The chi-square we will conduct will examine whether the teachers differed in the amount of importance they attached to different areas (where the importance of these areas was simply the frequency with which they made comments which fell into these categories). Actually, the author did not give a frequency table, but I am estimating the frequencies from the barplot in Figure 2 of the paper, where the author organized comments from the teachers into 12 different categories. My interpolated data is in Table 19.

Table 19 Grading importance in Shi (2001) study of writing teachers.

	General	Content	Ideas	Argument	Organization	Paragraph	organization	Transitions	Language	Intelligibility	Accuracy	Fluency	Length
Eng L1	11	19	35	57	23	9		5	3	23	22	14	4
Chin. L1	22	4	58	64	42	16		0	0	15	2	4	3

Right off the bat we can guess that this test will be statistical, just because there are so many different categories that it would be surprising to find that the native English speaking teachers performed in exactly the same way as the native Chinese speaking teacher. But I am simply

using this data to illustrate how to perform a chi-square if all you have is summary data (and, surprisingly, even summary data is hard to find in many published reports using chi-square).

In R Commander if you go to STATISTICS > CONTINGENCY > ENTER AND ANALYZE TWO-WAY TABLE (as shown in Figure 20), then you will see the dialogue box in Figure 24.

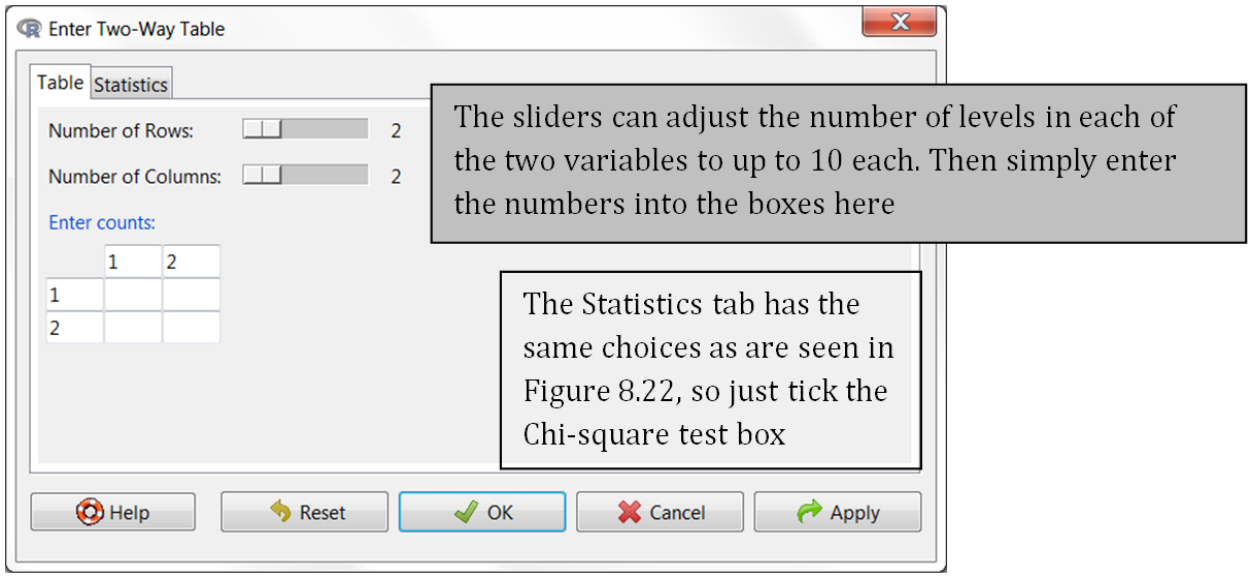

Figure 24 Enter a two-way table for a chi-square test in R Commander.

Since I couldn't enter all of the variables from Shi's table here (she had 12 variables and R Commander's dialogue box only lets me put in 10), I instead used R syntax to get the entire table analyzed:

```
.Table <- matrix(c(11,19,35,57,23,9,5,3,23,22,14,4,22,4,58,64,42,16,0,0,15,2,4,3), 2, 12, byrow=TRUE) #put in entire first row numbers, then second row
rownames(.Table) <- c('EngL1', 'ChinL1')
```

```
colnames(.Table) <- c('General', 'Content', 'Ideas', 'Argument', 'Organization',
'Paragraph Org.', 'Transitions', 'Language', 'Intelligibility', 'Accuracy', 'Fluency', 'Length')
.Table #Use this to check that your table is structured correctly
chisq.test(.Table, correct=FALSE)
```

Here are the results of the chi-square test:

```
Warning in chisq.test(.Table, correct = FALSE) :
  Chi-squared approximation may be incorrect

        Pearson's Chi-squared test

data:  .Table
X-squared = 59.0577, df = 11, p-value = 1.387e-08
```

As predicted, this chi-square has a very small *p*-value (the author gives the following chi-square results: $\chi^2 = 51.14, 19.99, 58.42$; df. = 11, $p < 0.001$; it is not clear why the author gives three chi-square numbers, but the third one looks very close to my result). Also, note the warning in the output. This is the warning that is generated when expected frequencies are smaller than 5. The easiest way to tell how many cells have expected frequencies less than 5 is to use R Commander and look for the warning. However, there is also a way to draw this information out from R. If you put the results of the chi-square test into an object, here named .Test, you can look at the expected values and see how many are less than 5.

```
.Test=chisq.test(.Table,correct=FALSE) #Remember, the FALSE is for continuity correction
.Test$expected
```

```
          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
[1,] 16.31868 11.37363 45.98901 59.83516 32.14286 12.36264 2.472527 1.483516
[2,] 16.68132 11.62637 47.01099 61.16484 32.85714 12.63736 2.527473 1.516484
          [,9]    [,10]    [,11]    [,12]
[1,] 18.79121 11.86813 8.901099 3.461538
[2,] 19.20879 12.13187 9.098901 3.538462
```

Here, expected cell counts for columns 7, 8, and 12 are less than 5, so there are 6 cells with counts less than 5. Remember that in the section on assumptions for chi-square I noted that violating this assumption of a "normal" distribution was problematic only insofar as it would result in low power. Since we have found a statistical result we really don't have to worry about the fact that expected values are low.

I would like to note that the output from R Commander's chi-square test gives the chi-square value, the degrees of freedom, and a *p*-value. However, it does not give any measure of effect size. Howell (2002) recommends the phi-coefficient (φ) and Cramér's V as appropriate effect sizes for a chi-square test. Phi is used when there are only 2 levels of each variable. If you want these numbers you can use the command assocstats( ) for the two-way chi-square (this command comes from the vcd package, so download that (install.packages("vcd")) and open the package (library(vcd))before you use the next command). I will return to the Dewaele and Pavlenko data now.

```
summary(assocstats(xtabs(~catdominance+numberoflang, data=beqDom)))
```

Here is the output generated by the assocstats( ) command:

```
Call: xtabs(formula = ~catdominance + numberoflang, data = beqDom)
Number of cases in table: 1036
Number of factors: 2
Test for independence of all factors:
        Chisq = 59.58, df = 6, p-value = 5.476e-11
                     X^2 df   P(> X^2)
Likelihood Ratio 63.742  6 7.7904e-12
Pearson          59.581  6 5.4760e-11

Phi-Coefficient   : 0.24
Contingency Coeff.: 0.233
Cramer's V        : 0.17
```

Now we know that we had 1036 valid cases tested here. We also have results for not just the Pearson chi-square, but also the Likelihood ratio test, which is a popular alternative to the Pearson test that also uses the χ^2 distribution. Howell (2002) says that the likelihood ratio will be equivalent to the Pearson when the samples sizes are large (as we see is the case here). The printout also shows effect sizes phi (φ=.24) and Cramer's V (V=.17). Since this is a 3 × 4 table (3 levels in CatDominance and 4 in NumberOfLangs), it is appropriate to use the Cramer's V effect size. Phi and Cramer's V are percentage variance effect sizes so this means that the number of languages that an individual knows explains 17% of the variance in the variable of L1 dominance. According to Cohen's guidelines (found in Table 4.5 of the book) for effect size strength, w = .17(sqrt(3-1)) = .24, this is a small-to-medium effect size.

If you are dealing with ordinal data you could report the results of the Linear-by-Linear association test, which assumes that both variables are ordinal. This test can be obtained by using the coin package (Hothorn, Hornik, van de Wiel & Zeileis, 2006), with the independence_test command, like this:

```
independence_test(catdominance~numberoflang,data=beqDom,teststat="quad")
```

Performing a Two-way Group Independence Chi-square in R

1. In R Commander, Choose STATISTICS > CONTINGENCY > TWO-WAY TABLE (if you have raw data) OR ENTER AND ANALYZE TWO-WAY TABLE (if you have summary data).
2. If you have raw data, choose two variables. Go to the "Statistics" tab and tick the Chi-square test box. If you have summary data, adjust the sliders to the correct number of levels for each of the two variables and enter the summary data.
3. The basic syntax in R is (replace the red parts with your own data):
   ```
   chisq.test(xtabs(~catdominance+numberoflang, data=beqDom), correct=FALSE)
   ```
4. If you get a warning that the approximation may be incorrect, this means the expected count in at least one of your cells is less than 1. Check the warning message at the bottom of R Commander, or put the chi-square test into an object and pull up the expected counts this way:
   ```
   .Test= chisq.test(xtabs(~catdominance+numberoflang,data=beqDom), correct=FALSE)
   .Test$expected
   ```
5. In order to get effect sizes and the likelihood ratio test, use the assocstats command in the vcd library:
   ```
   install.packages("vcd")
   library(vcd)
   summary(assocstats(xtabs(~CatDominance+NumberOfLang, data=beqDom)))
   ```
6. If you have ordinal data and want to get the linear-by-linear association, use the coin library:
   ```
   install.packages("coin")
   library(coin)
   independence_test(CatDominance~NumberOfLang,data=beqDom, teststat="quad")
   ```

Application Activities with Chi-Square in SPSS

1. Using the Geeslin and Guijarro-Fuentes (2006) data (GeeslinGF3_5), analyze Item 15 to see whether the native speakers of Spanish chose each possibility with equal probability. Additionally, generate some kind of visual to help you better understand the data, and describe what you have found.
2. Using the same data as in 1 above, test the hypothesis that there is only a 10% probability

that speakers will choose answer 3 ("both verbs equally"), while the probability that they will choose the other two choices is equal.

3 Using the Mackey and Silver (2005) data, investigate the question of whether there was any relationship between question development and experimental group for the immediate posttest (use the DevelopPost variable). We saw in the sections called "Two-Way Group-Independence Chi-Square" for either SPSS or R that there was a relationship between experimental group and delayed posttest. Does this hold true for the immediate posttest? Be sure to report on effect size.

4 Examine different data from Dewaele and Pavlenko (2001–2003) to determine if there is a statistical relationship between language dominance (L1dominance) and educational level (degree). Educational level has 4 levels: PhD, MA, BA and A level. Use the BEQ.Swear.sav file (if importing into R, call it beqSwear) to investigate this question using the chi-square test. Additionally, generate some kind of visual to help you better understand the data, and describe what you have found. If you choose a mosaic plot, don't worry about changing the default labels on the graph, and instead add this argument to your plot: labeling_args=list(rep=c(degree=F) (it cuts out repetition of the educational level labels).

5 Bryan Smith (2004) tested the question of whether preemptive input or negotiation in a computer-mediated conversation was more effective for helping students to learn vocabulary. Smith considered the question of whether these strategies resulted in successful uptake. He found that 21 participants who heard preemptive input had no uptake, while 2 had successful uptake. He further found that 37 participants who heard negotiated input had no uptake and 6 had successful uptake. Assuming that this data is

independent, is there any relationship between the way the participants heard the vocabulary and whether they had successful uptake (use the variables INPUT and UPTAKE)? Since we have only summary data, you'll have to do something a little different here. Open a new data file and enter the data by specifying rows and tables like this:

Input	Uptake	Result
Preemtive input	yes	2
Negotiated input	yes	6
Preemptive input	no	21
Negotiated input	no	37

Answers to Application Activities with Chi-Square in SPSS (No Answers Given for R)

1 Geeslin and Guijarro-Fuentes (GeeslinGF3_5.sav)

This is a one-way goodness-of-fit chi-square, so choose ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > CHI-SQUARE. Put Item 15 into the Test Variable List and click OK. Results show that we should reject the null hypothesis that all choices are equally likely (χ^2 = 6.421, df = 2, p = .04).

To get a visual, we will look at a barplot of the data. Choose GRAPHS > LEGACY DIALOGS > BAR, then SIMPLE, SUMMARIES FOR GROUPS OF CASES. Press Define. Put Item 15 in the Category Axis box and press OK. The barplot shows that #1 (estar) was the most frequent choice, #2 (ser) was the second most frequent, and #3 was the least frequent.

2 Geeslin and Guijarro-Fuentes (GeeslinGF3_5.sav)

Choose ANALYZE > NONPARAMETRIC TESTS > LEGACY DIALOGS > CHI-SQUARE. Put Item 15 into the Test Variable List and then under "Expected Values" click on the "Values" button. Enter: 45 (click ADD), 45 (Add), 10 (Add), then press OK.

Results show that we cannot now reject the null hypothesis (χ^2 = 1.468, df = 2, p = .48).

3 Mackey and Silver (2005)

This is a two-way group independence chi-square test, so choose ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS. Put EXPGROUP into "Row(s)" and DEVELOPPOST into "Column(s)." Tick "Display clustered bar charts." Open STATISTICS button and tick "Chi-square" and "Phi & Cramer's V" boxes; click CONTINUE. Open CELLS button and tick "Expected frequencies" plus any percentages you'd like; CONTINUE, then click OK.

There should be 26 cases and results show that we cannot reject the null hypothesis (Pearson χ^2 = .097, df = 1, p = .756). 1 cell has less than expected count. The effect size is very small (φ = .06) and not statistical.

4 Dewaele and Pavlenko (2003) BEQ.Dominance file

This is a two-way group independence chi-square test, so choose ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS. Put CATDOMINANCE into "Row(s)" and NUMBEROFLANG into "Column(s)" (or vice versa). Tick "Display clustered bar charts." Open the STATISTICS button

and tick "Chi-square" and "Phi & Cramer's V" boxes; click CONTINUE. Open CELLS button and tick "Expected frequencies" plus any percentages you'd like; CONTINUE, then press OK.

There should be 1036 valid cases, and results show that we can reject the null hypothesis that there is no relationship between the variables (Pearson χ^2 = 59.58, df = 6, p = .000). No cells have less than the expected count. The effect size is fairly small (Cramer's V = .17).

5. Smith (2004)

To weight the data in SPSS, go to DATA > WEIGHT CASES and put the variable RESULT into the line that says "Weight cases by." This will tell SPSS the data are summary data.

This is a two-way group independence chi-square test, so choose ANALYZE > DESCRIPTIVE STATISTICS > CROSSTABS. Put INPUT into "Row(s)" and UPTAKE into "Column(s)" (or vice versa). Tick "Display clustered bar charts." Open the STATISTICS button and tick "Chi-square" and "Phi & Cramer's V" boxes; click CONTINUE. Open CELLS button and tick "Expected frequencies" plus any percentages you'd like; CONTINUE, then press OK. We cannot reject the null hypothesis that there is no relationship between the variables (Pearson χ^2 = .39, df = 1, p = .53), so we assume that type of input makes no difference to whether students have uptake or not. The effect size is small (φ = .08).

Testing for Independence in Chi-Square when There Are More than Two Levels in Each Factor

Up to this point I have assumed that, if you have a contingency table where there are more than two levels in each of the two variables (a 2 × 2 table), you will simply get an overall chi-square result and look at the contingency table to make further inferences. In fact, if you are working

with data that is larger than 2 × 2 (but still only has two variables), you can do further testing to find out which groups are different from each other. You can partition the contingency table and then test the smaller 2 × 2 tables (Agresti, 2002). In order to demonstrate this idea, let us suppose Mackey and Silver had had three different experimental groups (this data is now made up) as shown in Table 20.

Group	*Developed*	*Did not develop*
Experimental 1	10	4
Experimental 2	9	2
Control	4	8

Table 20 An example of a 2×3 Contingency Table.

An overall chi-square test for group independence (results: $\chi^2 = 6.6$, df = 2, $p = .04$) would not tell you which groups were different from each other. If we have a 2 × J component table (where J = 3 in this example), then one can partition the table into (J– 1) tables (two in this case). Partitioning is done by computing first a 2 × 2 table using the first two rows and first two columns of the original table. The next partition will combine the first two rows and compare them to the third row. Thus we will have two partition tables to test with the group-independence chi-square test. Table 2 shows this partitioning of the made-up data in Table 21.

Group	*Developed*	*Did not develop*	*Group*	*Developed*	*Did not develop*
Experimental 1	10	4	Experimental 1 + Experimental 2	19	6
Experimental 2	9	2	Control	4	8

Table 21 Partitioning the 2 × 3 Table into 2 × 2 Tables.

In order to keep the tests independent, the rules for doing this partitioning are (Agresti, 2002, p. 84):

1. "The df for the subtables must sum to df for the full table."
2. "Each cell count in the full table must be a cell count in one and only one subtable" (the other times the original cell of the full table will be combined with another cell).
3. "Each marginal total [the row or column total] of the full table must be a marginal total for one and only one subtable" (but this rule will be upheld if 2 above is upheld).

You can see that the second rule is satisfied in the partitioning shown in Table 21. If we ran a chi-square for these two partitions, the results are that, for the leftmost table, χ^2=.36, df=1, p = .55 and, for the rightmost table, χ^2=6.3, df=1, p = .01. Thus the df for each of the partitions is 1 + 1, which sums to the 2 df for the full table. We also see that the difference between groups lies in the difference between the control group and the experimental groups, and not between the experimental groups themselves.

At this point, you might wonder how to test these smaller 2 × 2 contingency tables. Will you have to rearrange your data? One way to quickly do it would just be to enter summary data. In

SPSS you would open a new dataset, enter the numbers, and then weight the data (DATA > WEIGHT CASES) and put the variable that contains your counts into the "Weight Cases By" box. In R Commander, follow the menu command for STATISTICS > CONTINGENCY TABLES > ENTER AND ANALYZE TWO-WAY TABLE.

Returning to the question of analyzing even larger contingency tables, we would follow the same rules. Let's say our larger contingency table has I number of columns and J number of rows. As an example, take the Dewaele and Pavlenko data we have looked at in this paper, which is a 3 × 4 contingency table, so I = 3 and J = 4. There will be (I-1)(J-1) partitions, which in this case means (3–1)(4–1) = (2)(3) = 6 (see Table 22).

No. of languages	*L1 dominant (YES)*	*Other dominant (NO)*	*L1 + other(s) dominant (YESPLUS)*
Two	94	26	17
Three	159	26	83
Four	148	23	110
Five	157	30	163

Table 22 An example of a 3 × 4 Contingency Table.

To further examine differences, we would look at the first two columns and first two rows. Next we add together either the first two rows or the first two columns and compare them to the next row or column. This may become clearer as you look at what I do in Table 23. The second 2 × 2 table adds together the first two columns, YES + NO, and compares them to the third column, still for the first two rows. The third 2 × 2 table then adds the first two rows together and compares them to the fourth row, and so on. The bolded numbers indicate original numbers from

Table 23.

	YES	*NO*		*YES+NO*	*YESPLUS*		*YES+NO*	*YESPLUS*
Two	**94**	**26**	Two	120	17	Two+ Three	305	100
Three	**159**	**26**	Three	185	**83**	Four	171	**110**
	YES+NO	*YESPLUS*		*YES*	*NO*		*YES*	*NO*
Two+ Three+ Four	476	210	Two+ Three	253	52	Two+ Three+ Four	401	75
Five	187	**163**	Four	**148**	**23**	Five	**157**	**30**

Table 23 Partitioning the 3 × 4 Table into 2 × 2 Tables.

This was a little complicated to set up, but actually it's very helpful to check Agresti's Rule 2 above that every cell of the original table is found once in the partitioned tables (each original cell *must* be found once in a partitioned table). I've bolded the original cells in the partitioned table. The original table had a df = 6, and each of these six 2 × 2 tables will have a df = 1, which will sum to 6.

Effect Sizes for Chi-Square

Because one-way goodness-of-fit chi-squares are just looking at fit, there are no measures of effect size for this test.

For the chi-square group-independence test, one type of effect size you can use is phi (φ) or Cramer's V. These two numbers are the same when the test has only two levels for each factor (2 × 2), but Cramer's V should be used when there are more than two levels. This number is a

correlation (percentage variance, *r*-family effect size) and its value can range from 0 to ±1.

The effect size that Cohen (1988) uses for chi-square is called *w*. Volker (2006) says that *w* is equal to φ with 2 × 2 tables, and that $w = V\sqrt{r-1}$ when there are more than two levels to a variable, where V=Cramer's V and r=the number of rows or columns, whichever is smaller. For the Mackey and Silver (2005) data, we can obtain phi from the SPSS output ($\varphi = -.38$), so $w = \varphi = .38$, which is a medium effect size (we don't need to consider the negative sign if we think about this as an effect size). Calculating the effect size for the Dewaele and Pavlenko group-independence chi-square, we can obtain V from the print-out (V = .17) and the smallest number of levels is 3, in the "Categorical Dominance" variable, so $w = .17\sqrt{(3-1)} = .17\sqrt{2} = .24$, which could be considered a small to medium effect size.

Howell (2002) also notes that an **odds ratio** can be an intuitively understandable way to present results and understand the strength of the effect. Heiberger and Holland (2004) further note that odds ratios are not much affected by group size, unlike the chi-square statistic. I'll show how an odds ratio could be calculated. Table 24 shows the crosstab output for the Mackey and Silver (2005) data in the delayed posttest condition.

Group ' DevelopDelPost Crosstabulation

Count

		DevelopDelPost		
		Developed	Not developed	Total
Group	Control	4	8	12
	Experimental	10	4	14
Total		14	12	26

Table 24 Crosstabs for Mackey and Silver (2005).

For the Mackey and Silver (2005) data, which was a 2 × 2 table, if we look at the odds of developing on question formation given that a person was in the experimental group, this is equal to the number of people in that group who developed divided by the number of people in that group who did not develop, so the odds equal 10/4 = 2.5. The odds of developing in the control group are 4/8 = .5. Now we make an odds ratio with both of these odds: 2.5/.5= 5. You are five times more likely to develop your question-making ability if you were in the experimental group than if you were in the control group. An even easier way to calculate this odds ratio is to calculate the cross product of the contingency table ($n_{11}n_{22}/n_{12}n_{21}$), where the subscript on the n refers to position in the table, as shown in Table 25.

N_{11}	N_{12}
N_{21}	N_{22}

Table 25 Table Subscripts.

You just need to make sure that the ratio you are testing for is the one in the N11 position. With the Mackey and Silver data we need to flip Table 23 on its head so that the experimental group that developed is in the N11 position, and then we have $\frac{10*8}{4*4} = \frac{80}{16} = 5.$

To calculate odds ratios for contingency tables larger than 2 × 2 you need to collapse some categories to get to a 2 × 2 contingency table. Let's say that, with the Dewaele and Pavlenko data, we want to know the odds of being dominant in more than one language if you know five languages. We will then collapse the two categories of "Dominant in L1" and "Dominant in LX"

into one category. For number of languages that a person knows we will collapse this into "Knows five languages" and "Knows fewer than five languages." This collapsing results in the contingency table in Table 26.

	Five languages	*Fewer than five languages*
Dominant One	187	476
Dominant Two	163	210

Table 26 Dewaele and Pavlenko 2 × 2 Contingency Table.

The odds of being dominant in two (or more) languages given that you speak five languages is equal to $\frac{163*476}{187*210} = \frac{77588}{39270} = 1.97$ (again, we will consider the data in the intersection of two (dominant languages) and five (known languages) as our N_{11} position). You are about twice as likely (as bilinguals, trilinguals, and quadrilinguals) to be dominant in two languages than just one if you know five languages.

Reporting Chi-Square Test Results

For chi-square tests, it is imperative that you provide a contingency table with a summary of your data as well as the statistical results. I also recommend providing an informative graphic, which is usually not a barplot, but instead an Association plot, Mosaic plot, or Doubledecker plot. Since chi-square is not a directional test, it is impossible to understand the statistical results without seeing a contingency table. Once the reader understands how to interpret the newer types of graphics, these are also excellent for helping readers understand the statistical inferences in the data at a glance.

For the chi-square test itself, you should report the type of test that was performed, the chi-square value, the degrees of freedom, and the *p*-value of the test. If you have done a test for group independence, you should report effect sizes as phi (or Cramer's V if the table is larger than 2 × 2) or *w* and include confidence intervals for this effect size if possible. You might want to calculate odds ratios as well, as these are quite intuitive to understand.

For a one-way goodness-of-fit chi-square, let's consider a report about the Geeslin and Guijarro-Fuentes (2006) data:

> A one-way goodness-of-fit chi-square was conducted to see whether native speakers of Spanish were choosing each of the three choices for verbs equally in Items 3, 4, and 5. The frequency counts for each choice for each item are given in the following table:

Item	*Estar*	*Ser*	*Both*	*Total*
3	13	4	2	19
4	19	0	0	19
5	13	5	1	19

> A separate chi-square for each item revealed that native speakers of Spanish chose *estar* more frequently than would be predicted if all speakers were randomly picking one of the three choices (Item 3: $\chi^2 = 10.8$, df = 2, $p = .004$; Item 4: could not be tested because only *estar* was chosen; Item 5: $\chi^2 = 11.8$, df = 2, $p = .003$).

For a two-way group-independence chi-square, we will report the result of the Mackey and

Silver (2005) data:

A two-way group-independence chi-square was performed to assess the relationship between group membership and development in question formation. A contingency table for these data is shown below:

Group	Developed	Did not develop
Experimental	10	4
Control	4	8

The effect size of the chi-square test was phi=-.381, with a 95% CI of [.01, .70]. The wide confidence interval for this effect size means that membership in the group could have an effect as low as .01, explaining only 1% of the results, or as high as .70, explaining 70% of the results. I interpret this result as indicating that we cannot be confident yet in the relationship between group membership and development of question formation. The formal statistical results of the chi-square test are the following: likelihood ratio $\chi^2 = 3.86$, df = 1, p = .049). The odds of developing in the question-formation hierarchy for participants in the experimental group were five times greater than for participants in the control group, but the variance seen in the phi effect size tells us we cannot trust this odds ratio as a point estimate (odds ratio= [$\frac{\frac{10}{4}}{\frac{4}{8}} = \frac{2.5}{.5} = 5$]).

Summary of Chi-Square

In this section we have seen that chi-square is a test used when all of your variables are categorical. There are several situations when it is inappropriate to use chi-square, however, even

if all of your variables are categorical. All of the situations involve a violation of the assumption that each person will contribute only once to each cell of a contingency table summarizing the data.

When looking at a situation that satisfies this assumption, there are two types of chi-square tests that can be used. These tests differ in the number of variables that they require, and in hypotheses they make. The one-way goodness-of-fit chi-square test requires only one variable. It tests the hypothesis that every choice in the variable is equally likely (as a default, although other percentages can be specified).

The two-way group-independence chi-square test requires two variables. It tests the hypothesis that there is no relationship between the two categorical variables. This is probably the more common chi-square test in the field of second language research. Further sections showed how to analyze differences between levels of variables statistically (by partitioning them into 2×2 tables and testing them) and also how to calculate odds ratios as well as report effect sizes for these tests.

If you have more than two categorical variables you would like to test then you cannot use a chi-square test. Field (2005, 2012) recommends loglinear analysis, and I refer you to his books for more information.

Bibliography

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Aragon, T. J. (2012). epitools: Epidemiology Tools. R package version 0.5-7 [Software]. Available from http://CRAN.R-project.org/package=epitools

Beattie, G., Webster, K. & Ross, J. (2010). The fixation and processing of the iconic gestures that accompany talk. *Journal of Language and Social Psychology*, 29(2), 194–213.

Belia, S., Fidler, F.,Williams, J. & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10, 389–396.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Newbury Park, CA: Sage.

Crawley, M. J. (2002). *Statistical computing: An introduction to data analysis using S-PLUS*. New York: Wiley.

Crawley, M. J. (2007). *The R book*. Hoboken, NJ Wiley.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research*. New York: Routledge.

Dewaele, J.-M., & Pavlenko, A. (2001–2003). *Webquestionnaire: Bilingualism and Emotions*. London: University of London.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: Sage publications.

Flynn, K. S. (2012). *Stability of special education eligibility from kindergarten to third grade: Are there variables from fall of kindergarten that predict later classification status*? (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.

Friendly, M. (2000). *Visualizing categorical data*. Cary, NC: SAS.

Geeslin, K. L., & Guijarro-Fuentes, P. (2006). Second language acquisition of variable structure in Spanish by Portuguese speakers. *Language Learning*, *56*(1), 53–107.

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

Heiberger, R. M., & Holland, B. (2004). *Statistical analysis and data display: An intermediate course with examples in S-PLUS, R, and SAS*. New York: Springer.

Hothorn, T., Hornik, K., van de Wiel, M.A. & Zeileis, A. (2006). A Lego System for Conditional Inference. *The American Statistician,* 60(3), 257–263.

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

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31*(3), 368–390.

Mackey, A., & Silver, R. E. (2005). Interactional tasks and English L2 learning by immigrant children in Singapore. *System*, *33*(2), 239–260.

McDuffie, A., Kover, S. T., Hagerman, R, and L. Abbeduto (2013). Investigating word learning in Fragile X Syndrome: A fast-mapping study. *Journal of Autism Development Disorders*, 43, pp. 1676-1691.

Meyer, D., Zeileis, A. & Hornik, K. (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. Journal of Statistical Software, 17(3), 1-48.

Meyer, D., Zeileis, A., & Hornik, K. (2007, 2007–06–11). The vcd package. Retrieved October 3, 2007.

Meyer, D., Zeileis, A., & Hornik, K. (2014). vcd: Visualizing Categorical Data. R package version 1.3-2 [Software]. Available from http://cran.r-project.org/web/packages/vcd/

Oldham, D. S. (2012). Analysis of an early intervention reading program for fifth grade students in a metropolitan elementary school in Georgia (Unpublished doctoral dissertation). Walden University: Minneapolis, Minnesota.

Saito, H. (1999). Dependence and interaction in frequency data analysis in SLA research. *Studies in Second Language Acquisition*, *21*, 453–475.

Shi, L. (2001). Native- and nonnative-speaking teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303–325.

Smith, B. (2004). Computer-mediated negotiated interaction and lexical acquisition. *Studies in Second Language Acquisition*, *26*(3), 365–398.

Volker, M. A. (2006). Reporting effect size estimates in school psychology research. *Psychology in the Schools*, *43*(6), 653–672.

Wainer, H. (1996). Depicting error. *American Statistician*, *50*(2), 101–111.