Other Kinds of Correlations in R

Partial Correlation

Do you think that how well second language learners can pronounce words in their second language gets worse as they get older? I certainly didn't suspect this might be the case when I performed an experiment designed to see how well 15 Japanese speakers living in the United States for 12 years or more pronounced words beginning in /r/ and /l/ (Larson-Hall, 2006).

In every experimental condition the researcher wants to manipulate some variables while holding all other variables constant. One way to do this involves controlling for the variable before experimental participants are chosen. If I had thought age was a concern for pronunciation accuracy, I would have set experimental parameters to exclude participants over, say, age 50. When I found, after the fact, that pronunciation accuracy as well as scores on a timed language aptitude test declined with age, the only way left to hold the age variable constant was to use partial correlation to subtract the effects of age from the correlations I was interested in.

I found a strong (as judged by effect size) and statistical negative correlation between length of residence (LOR) and production accuracy (as later judged by native speaker judges; $r = -.88$) as well as LOR and scores on a language aptitude test ($r = -.55$). This meant that, as the participants lived in the US longer, their scores went down on both measures. However, I also found that *age* correlated negatively with both production

accuracy and aptitude scores! Of course age also correlated positively with LOR (the longer a person had lived in the US, the older they were; r = .74). Thus, in order to determine the true relationship between length of residence and production accuracy, I needed to use a partial correlation. The partial correlation can tell me how LOR and accuracy vary together by subtracting out the effects of age.

Calling for a Partial Correlation

In R Commander a partial correlation is done by including only the pair of variables that you want to examine and the variable you want to control for. In other words, in order to get the correlation between LOR and aptitude while controlling for age, I will include only these three variables in the correlation. The partial correlation command in R will return a matrix of partial correlations for each pair of variables, always controlling for any other variables that are included.

If you want to follow what I am doing, import the SPSS file LarsonHallPartial.sav and name it partial. The steps to performing a partial correlation are exactly the same as to performing any other correlation in R Commander: STATISTICS > SUMMARIES > CORRELATION MATRIX. Choose the two variables you are interested in seeing the correlation between (I want aptitude and LOR), and then the variable(s) you want to control for (age, although it actually doesn't matter what order you pick these in), and click the Partial button (it doesn't matter if you check the box for pairwise *p*-values; you will not get *p*-values with the partial correlation command). The syntax for this command, in the case of the partial correlation between LOR and aptitude while controlling for age, is:

```
partial.cor(partial[,c("age","lor","aptitude")], use="complete")
```

This syntax should look familiar as it is almost exactly the same as the `rcorr.adjust( )` command; just the command `partial.cor( )` is new. The argument `use="complete"` is inserted in order to specify that missing values should be removed before R performs its calculations. As always, it's best if you can impute missing values instead of removing cases with missing data.

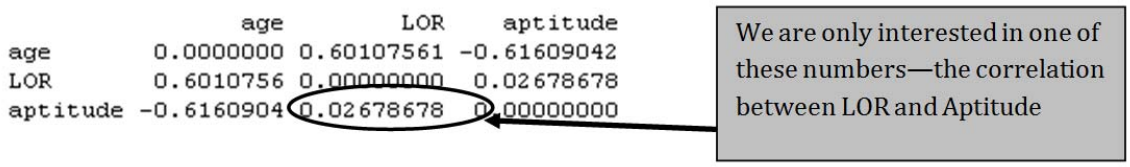

Figure 1 Pairwise correlation output.

The output shown in Figure 1 gives the Pearson r value for the pairwise correlation. The results show that the correlation between LOR and aptitude are not important when age is removed from the equation (compare that to the non-partial correlation coefficient, which was r = -.55!). Note that no *p*-value is given in this command. That doesn't bother us because the very low correlation means that the correlation is not important. However, the fact that we don't have a confidence interval does bother us.

What the fact that the effect size of the correlation is negligible with age partialled out means is that declines in scores on the aptitude test are almost all actually due to age. In

order to test the other correlation we are interested in, that between LOR and accuracy when age is controlled, we would need to perform another partial correlation with only those three variables included. Doing so shows that there is still a strong correlation between LOR and accuracy (r=-.75).

For confidence intervals, we can use the psych package (Revelle, 2015). We'll need to take just the partial correlation matrix that was calculated from the partial.cor( ) function and put it into an object. If you look at this object, you'll see that just the part $R is the matrix, so we will enter that into the corr.p( ) command, which returns *p*-values and confidence intervals, if we set short = F.

```
library(psych)
m <- partial.cor(partial[,c("age","lor","aptitude")], use="complete")
#use command from above
cp <- corr.p(m$R, n=13)
#set n=number of participants minus two (there were 15 in this dataset)
print(cp, short=F)
```

output omitted. . .

```
 Confidence intervals based upon normal theory.  To get bootstrapped values, try cor.ci
          lower     r upper    p
age-lor    0.07  0.60  0.87 0.03
age-apttd -0.87 -0.62 -0.10 0.02
lor-apttd -0.53  0.03  0.57 0.93
```

We want the correlation between LOR and aptitude so we would look at the third line in this output (which has age partialled out) and see that the confidence interval is [-.53, .57], clearly showing a lack of any correlation when age is partialled out.

Partial Correlation in R
In R Commander, choose STATISTICS > SUMMARIES > CORRELATION MATRIX.
Choose the radio button for "Partial."
Choose three variables—the two you want to see correlated with the third variable the one you want to partial out.
By the way, you can partial out more than one variable at a time; just add another to the mix to partial out two, for example.
R code (looking at correlation between two of the variables with the third partialled out):

```
partial.cor(partial[,c("age","LOR","aptitude")], use="complete.obs")
#(N.B. items in red should be replaced with your own data name)
#also note that dataset partial is attached for this command
```

To get confidence intervals, use the psych package and put the results from the partial correlation into an object m, whose $R component contains the partial correlation matrix:

```
m <- partial.cor(partial[,c("age","lor","aptitude")], use="complete")
cp <- corr.p(m$R, n=13)
#set n=number of participants minus two
print(cp, short=F)
```

Reporting Results of Partial Correlation

To report the results found for my data, I would say:

A partial correlation controlling for age found a strong correlation between length of residence and production accuracy of R/L words. The Pearson *r* correlation coefficient was negative ($r = -.75$), meaning scores on production accuracy decreased with increasing length of residence, and a 95% BCa CI of [-.89, -.55] showed that there was an effect for this partial correlation. The width of the interval means the correlation

coefficient is not precise, but even the lower limit of the CI shows that we can be confident that there is a strong relationship between accuracy and length of residence, and the effect size was large ($R^2 = .56$). Controlling for age, the correlation between LOR and scores on the language aptitude test was very small and we can say there was basically no effect ($r = .03$, 95% CI [-.53, .57]).

Point-Biserial Correlations

It is also permissible to enter a categorical variable in the Pearson's *r* correlation if it is a **dichotomous variable**, meaning there are only two choices (Howell, 2002). In the case of a dichotomous variable crossed with a continuous variable, the resulting correlation is known as the **point-biserial correlation** (r_{pb}). Often this type of correlation is used in the area of test evaluation, where answers are scored as either correct or incorrect.

For example, in order to test the morphosyntactic abilities of non-literate bilinguals I created an oral grammaticality judgment test in Japanese. The examinees had to rate each sentence as either "good" (grammatical) or "bad" (ungrammatical), resulting in dichotomous (right/wrong) answers. Since this was a test I created, I wanted to examine the validity of the test, and see how well individual items discriminated between test takers. One way to do this is by looking at a discrimination index, which measures "the extent to which the results of an individual item correlate with results from the whole test" (Alderson, Clapham, & Wall, 1995). Such a discrimination index investigates whether test takers who did well overall on the test did well on specific items, and whether those who did poorly overall did poorly on specific items. It therefore examines

the correlation between overall score and score on one specific item (a dichotomous variable). Scores are ideally close to +1.

One way to determine item discrimination in classical test theory is to conduct a corrected point-biserial correlation, which means that scores for the item are crossed with scores for the entire test, minus that particular item (that is the "corrected" part in the name).

Point-Biserial correlations using R

Import the SPSS file LarsonHallGJT.sav as LHtest. To conduct the reliability assessment in R Commander choose STATISTICS > DIMENSIONAL ANALYSIS > SCALE RELIABILITY. Pick the total test score (totalscore) and the dichotomous scores for each item (for demonstration purposes I will show you the output just for the last three items of the test, Q43, Q44 and Q45). Below is the R code for this same procedure.

```
reliability(cov(LHtest[,c("Q43","Q44","Q45","totalscore")], use="complete.obs"))
```

```
Alpha reliability =  0.1911
Standardized alpha =  0.5545

Reliability deleting each item in turn:
            Alpha Std.Alpha r(item, total)
Q43        0.1729    0.5803          0.2464
Q44        0.1491    0.4920          0.2780
Q45        0.1198    0.4355          0.3939
TotalScore 0.4198    0.4056          0.4282
```

The output first shows the overall reliability for these three items (it is low here but would be higher with more items). The point-biserial correlation for each item is the third column of data titled "r(item, total)" and the line above the columns informs us that,

appropriately, this has been calculated by deleting that particular item (say, Question43) from the total and then conducting a correlation between Q43 and the total of the test (also called the Corrected Item-Total Correlation). Oller (1979) states that, for item discrimination, correlations of less than .35 or .25 are often discarded by professional test makers as not being useful for discriminating between participants.

More modern methods of test item analysis have become more popular, however, now that computing power has increased. In particular, item response theory (IRT) provides a way to analyze test items by positing a latent or unmeasured trait that is linked to the dichotomous scores. McNamara and Knoch (2012) state that IRT as a tool for analyzing language tests "appears to have become uncontroversial and routine" (p. 569). Although there is not space in this book to detail how IRT works, interested readers are directed to edited collections by Baker and Kim (2004) and van der Linden and Hambleton (1997), and more recent articles by Ellis and Ross (2013).

In other cases where you may have a dichotomous variable such as gender (male versus female) or group membership with only two categories (student versus employed, for example) that you want to correlate with a continuous variable such as TOEFL scores, it generally does not make sense to conduct a correlation (whether Pearson or Spearman) because you have so little variation in the dichotomous variable (there are some exceptions; see Hatch & Lazaraton, 1991, p. 450, for additional information). It would be better in this case to compare means for the two groups using a *t*-test or one-way ANOVA.

Summary Conducting a Point-biserial Correlation with R
In R Commander choose STATISTICS > DIMENSIONAL ANALYSIS > SCALE RELIABILITY. If doing test item analysis, pick the total test score and the dichotomous scores for each item.

The R code is:

```
reliability(cov(LHtest[,c("Q43","Q44","Q45","TotalScore")],
use="complete.obs"))
```

(N.B. items in red should be replaced with your own data name):

Inter-rater Reliability

It often happens in second language research that you will have a set of judges who will rate participants. The judges may rate the participants' pronunciation accuracy or writing ability or judge the number of errors they made in past tense, for example. In this case you will have multiple scores for each participant that you will average to conduct a statistical test on the data. However, you should also report some statistics that explore to what extent your raters have agreed on their ratings.

If you think about what is going on with judges' ratings, you will realize that you want the judges' ratings to differ based on the participants that they rated. For example, Judge A may give Participant 1 an 8 and Participant 2 a 3 on a 10-point scale. You would then hope that Judge B will also give Participant 1 a high score and Participant 2 a low score, although they may not be exactly the same numbers. What you don't want is for judges' scores to vary based on the judge. If this happened, Participant 1 might get an 8 from

Judge A but a 2 from Judge B and a 10 from Judge C. In other words, you want to see that the variability in scores is due to variation in the sample and not variation in the judges. Any variation that is seen in the judges' scores will be considered error, and will make the rating less reliable. DeVellis (2005) defines **reliability** as "The proportion of variance in a measure that can be ascribed to a true score" (p. 317). Mackey and Gass (2005) define reliability as consistency of a score or a test. They say a test is reliable if the same person taking it again would get the same score. You can see that these two definitions of reliability are similar, for they both address the idea that a test result can be confidently replicated for the same person. Therefore, the more reliable a measurement is, the more it will measure the right thing (the true score) and the less error it will have.

Howell (2002) says the best way to calculate **inter-rater reliability** for cases of judges rating persons is to look at the intraclass correlation. This will not only take into account the correlation between judges, but also look at whether the actual scores they gave participants differed. We will look at **Cronbach's alpha** as a measurement of intraclass correction. Cortina (1994) says that coefficient alpha is an internal consistency estimate, "which takes into account variance attributable to subjects and variance attributable to the interaction between subjects and items [on a test, or for our purposes here, judges]" (p. 98).

In general, we might like a rule of thumb for determining what an acceptable level of overall Cronbach's alpha is, and some authors do put forth a level of 0.70–0.80. Cortina (1994) says determining a general rule is impossible unless we consider the factors that

affect the size of Cronbach's alpha, which include the number of items (judges in our case) and the number of dimensions in the data. In general, the higher the number of items, the higher alpha can be even if the average correlations between items are not very large and there is more than one dimension in the data. Cortina says that, "if a scale has enough items (i.e. more than 20), then it can have an alpha of greater than .70 even when the correlation among items is very small" (p. 102).

In this section I will use data from a study by Munro, Derwing, and Morton (2006). These authors investigated to what extent the L1 background of the judges would affect how they rated ESL learners from four different L1 backgrounds—Cantonese, Japanese, Spanish, and Polish. The judges themselves were native speakers also of four different backgrounds—English, Cantonese, Japanese, and Mandarin, but I will examine the data only from the ten Mandarin judges here. The judges rated the samples on three dimensions—their comprehensibility, intelligibility, and accentedness. I will examine only scores for accentedness here using the file MunroDerwingMorton.sav.

Calling for Inter-rater Reliability

To follow along, import the SPSS file MunroDerwingMorton.sav into R as "MDM." To calculate the intraclass correlation for a group of raters, in R Commander choose STATISTICS > DIMENSIONAL ANALYSIS > SCALE RELIABILITY. Choose all of the variables except for "Speaker." The columns you choose should consist of the rating for each participant on a different row, with the column containing the ratings of each judge. Therefore, in the MDM dataset, variable M001 contains the ratings of Mandarin Judge 1 on the accent of 48 speakers, M002 contains the ratings of Mandarin Judge 2 on the

accent of the 48 speakers, and so on. The reliability( ) command will calculate Cronbach's alpha for a composite scale.

```
reliability(cov(MDM[,c("m001","m002","m003","m004","m005","m006","m007",
"m008","m009","m010")], use="complete.obs"))
```

```
Alpha reliability =  0.8848
Standardized alpha =  0.8924

Reliability deleting each item in turn:
      Alpha Std.Alpha r(item, total)
m001 0.8893    0.8936         0.4486
m002 0.8699    0.8796         0.6681
m003 0.8770    0.8827         0.6129
m004 0.8645    0.8749         0.7320
m005 0.8830    0.8916         0.4841
m006 0.8721    0.8806         0.6602
m007 0.8719    0.8816         0.6552
m008 0.8679    0.8786         0.6905
m009 0.8761    0.8828         0.6080
m010 0.8608    0.8713         0.7757
```

Whereas for test analysis we were most interested in the third column, the corrected item-total correlation, here we will be interested in the second column, which contains the standardized Cronbach's alpha. For the Mandarin judges overall, Cronbach's alpha is 0.89 (the "standardized alpha" at the top of the printout). This is a high correlation considering that there are ten items (judges).

Remember that we do not have any general rule of thumb for determining what level of Cronbach's alpha is acceptable, and it is important to look at the correlations between pairs of variables. This means we should look at a correlation matrix between all of the

variables. Call for this in R code with the rcorr.adjust( ) command:

```
rcorr.adjust(MDM[c("m001","m002","m003","m004","m005","m006","m007","m008","m009","m010")],type="pearson")
```

```
      m001 m002 m003 m004 m005 m006 m007 m008 m009 m010
m001 1.00 0.35 0.34 0.41 0.23 0.24 0.34 0.38 0.41 0.32
m002 0.35 1.00 0.33 0.53 0.41 0.48 0.52 0.61 0.51 0.54
m003 0.34 0.33 1.00 0.60 0.43 0.41 0.33 0.39 0.58 0.61
m004 0.41 0.53 0.60 1.00 0.43 0.62 0.51 0.50 0.44 0.62
m005 0.23 0.41 0.43 0.43 1.00 0.37 0.31 0.25 0.31 0.48
m006 0.24 0.48 0.41 0.62 0.37 1.00 0.55 0.55 0.31 0.67
m007 0.34 0.52 0.33 0.51 0.31 0.55 1.00 0.58 0.39 0.59
m008 0.38 0.61 0.39 0.50 0.25 0.55 0.58 1.00 0.51 0.58
m009 0.41 0.51 0.58 0.44 0.31 0.31 0.39 0.51 1.00 0.55
m010 0.32 0.54 0.61 0.62 0.48 0.67 0.59 0.58 0.55 1.00
```

Because the matrix is repeated above and below the diagonal line, you only need to look at one side or the other. By and large the paired correlations between judges are in the range of 0.30–0.60, which are medium to large effect sizes, and this Cronbach's alpha can be said to be fairly reliable. However, if the number of judges were quite small, say three, then Cronbach's alpha would be quite a bit lower than what is obtained with 10 or 20 items even if the average inter-item correlation is the same. Try it yourself with the data— randomly pick three judges and see what your Cronbach's alpha is (I got .65 with the three I picked).

Why don't we just use the average inter-item correlation as a measure of reliability between judges? Howell (2002) says that the problem with this approach is that it cannot tell you whether the judges rated the same people the same way, or just if the trend of higher and lower scores for the same participant was followed.

Another piece of output I want to look at is the reliability if each item (judge) individually were removed. If judges are consistent then there shouldn't be too much variation in these numbers. This information is found in the first column of data next to each of the judges (m001, m002, etc.) in the output for the reliability( ) command above. Looking at this column I see that there is not much difference in overall Cronbach's alpha if any of the judges is dropped for the Munro, Derwing, and Morton (2006) data (nothing drops lower than 86%), and that is a good result. However, if there were a certain judge whose data changed Cronbach's drastically you might consider throwing out that judge's scores.

Overall test reliability is often also reported using this same method. For example, DeKeyser (2000) reports, for his 200-item grammaticality judgment test, that "The reliability coefficient (KR-20) obtained was .91 for grammatical items [100 items] and .97 for ungrammatical items" (p. 509) (note that, for dichotomous test items, the Kuder–Richardson (KR-20) measure of test reliability is equal to Cronbach's alpha). DeKeyser gives raw data in his article, but this raw data does not include individual dichotomous results on each of the 200 items of the test. These would be necessary to calculate the overall test reliability. Using the file LarsonHallGJT.sav file (imported as LHtest) I will show how to obtain an overall test reliability score if you have the raw scores (coded as 1s for correct answers and 0s for incorrect answers).

I could use the previous reliability( ) command but I'm going to introduce another possibility here. That is the alpha( ) command from the psych package. I like it because

I don't have to list individual items, but can instead put the whole data frame into the command. For the LHtest data though, I just have to make sure I delete any variable (like ID or TotalScore) that are not individual items I want analyzed. Do this easily in R Commander by going to DATA > ACTIVE DATASET > SUBSET DATASET. In Figure 2 you can see I ticked off "Include all variables" and instead chose only the individual items. I then gave the new dataset a different name, LHtest.short.

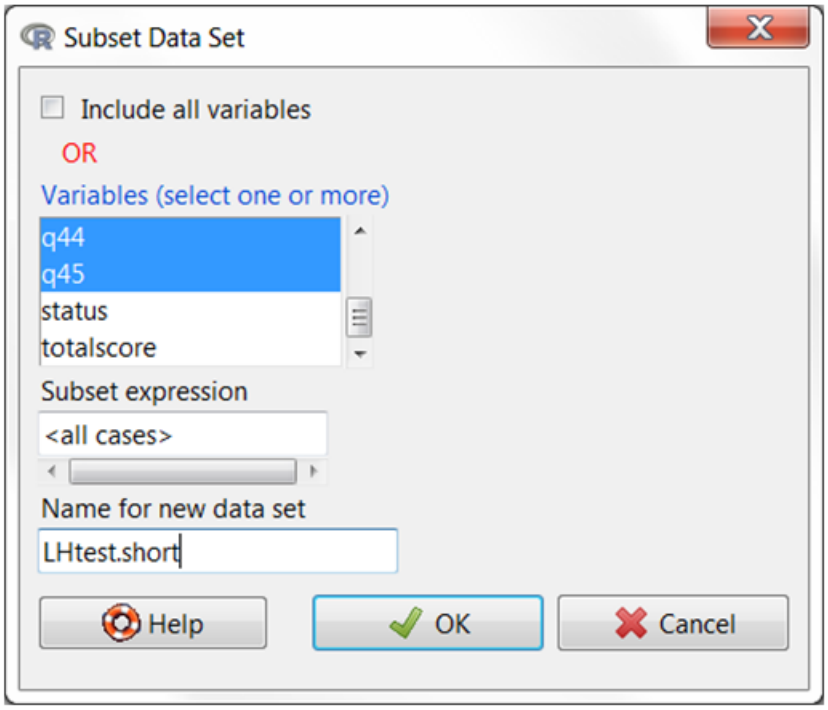

Figure 2 Getting rid of unwanted columns in your dataset using R Commander.

Now open the psych package and use the alpha command on the subsetted data (this must be done in R). If you don't have the psych package you can easily install it by typing:

```
install.packages("psych")
```

which will work if you have root access in your computer to a directory that the package can be stored in.

```
library(psych)
alpha(LHtest.short)
```

```
Warning in alpha(LHtest.short) :
  Some items were negatively correlated with total scale and were automatically reversed.

Reliability analysis
Call: alpha(x = LHtest.short)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd
      0.69      0.68    0.89     0.051 2.2 0.059 0.57 0.13

 lower alpha upper     95% confidence boundaries
0.57 0.69 0.8

 Reliability if an item is dropped:
     raw_alpha std.alpha G6(smc) average_r S/N alpha se
q4         0.69      0.68    0.88     0.053 2.2    0.059
```

The beginning of the output shows me that with all 40 items I have a Cronbach's alpha (under the "raw_alpha" column) of 0.69, which can also be reported as a KR-20 score of .69. This is not very high considering how many items I have, so it would be hard to call this a highly reliable test (I made it up myself and it clearly needs more work! I actually presented a conference paper at AAAL 2008 where I used R to analyze the data with IRT methods, and I would be happy to send you this presentation if you are interested).

Summary: Calculating Inter-rater Reliability
In R Commander, get a measure of the intra-class correlation (measured as Cronbach's alpha) and the reliability of the ratings if each item (judge) were removed by choosing STATISTICS > DIMENSIONAL ANALYSIS > SCALE RELIABILITY and choosing all of the items in your test (often the columns (the variables) are different judges who judged the data)

The R code for this is:
```
reliability(cov(MDM[,c("m001","m002","m003","totalScore")], use="complete.obs"))
```
(N.B. items in red should be replaced with your own data names)

Alternatively, you don't have to type in individual names of variables if you use:

```
library(psych)
alpha(MDM)
```

Just make sure to delete any variables first that aren't test items!

One more piece of information you should examine is a correlation matrix among all your items, and make sure none of them have extremely low correlations. Call for the correlation matrix with the R code:

```
rcorr.adjust(MDM[c("m001"," m002"," m003"," m004"," m005")],type="pearson")
```

Bibliography

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. New York: Cambridge University Press.

Baker, F. B., & Kim, S.-H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Cortina, J. M. (1994). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*(4), 499–533.

DeVellis, R. F. (2005). Inter-rater reliability. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*(pp. 317–322). San Diego, CA: Academic.

Ellis, D. P. & Ross, S. J. (2013). Item response theory in language testing. In Kunnan, A. J. (Ed.), *The Companion to Language Assessment*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118411360.wbcla016/abstract

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

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

Larson-Hall, J. (2006). What does more time buy you? Another look at the effects of long-term residence on production accuracy of English /r/ and /l/ by Japanese speakers. *Language and Speech*, *49*(4), 521–548.

Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Erlbaum.

McNamara, T. & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, 29(4), 555–576.

Munro, M., Derwing, T., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*, 111–131.

Oller, J. W. (1979). *Language tests at school*. London: Longman.

Revelle, W. (2015). Package 'psych' [Software]. Retrieved from http:/cran.r-project.org/web/packages/psych/psych.pdf

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.