

Other Kinds of Correlation in SPSS

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 SPSS, call for a partial correlation by choosing ANALYZE > CORRELATE > PARTIAL command.

If you want to follow along, I'm using the LarsonHallPartial.sav file. The dialogue box is almost the same as the one for regular correlations, except that it asks you to put factors you want to control for in the box labeled CONTROLLING FOR (see Figure 1). In order to get a confidence interval, open the "Bootstrap" button and tick the "Perform bootstrapping" box. Change the confidence intervals type to BCa from the default Percentile.

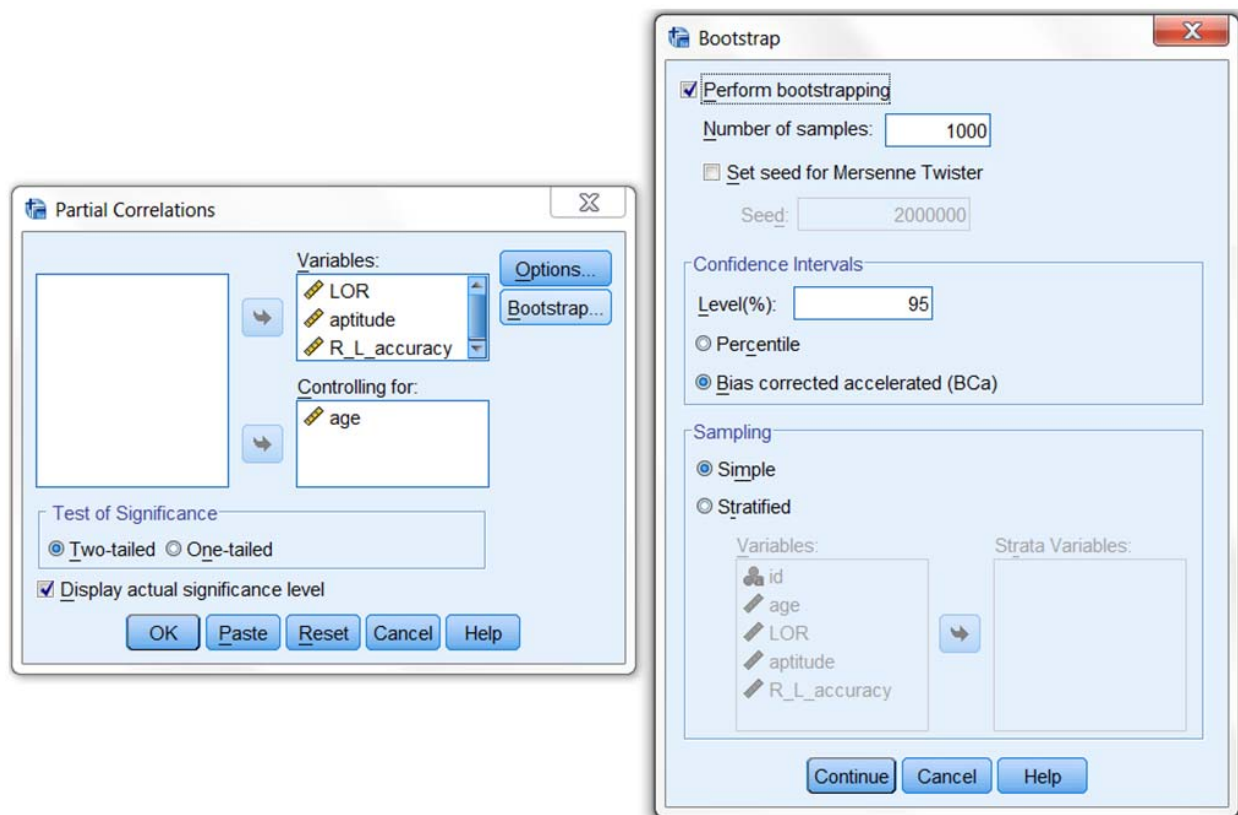


Figure 1 Calling for a partial correlation in SPSS.

The output shown in Table 1 is almost identical to the normal correlation matrix output except that degrees of freedom (df) are shown instead of N. The output shows that the correlations between length of residence (LOR) and production accuracy are now slightly smaller but still quite substantial, even given the lower limit of the confidence interval ($r = -.75$, 95% BCa CI $[-.89, -.55]$), while the correlation between the language aptitude score and LOR now has no effect, as the CI passes through zero and is quite wide ($r = .03$, $[-.61, .74]$). This seems to imply that age played a large role in explaining the relationship of LOR and the aptitude scores, but not as great a role in the correlation between LOR and production accuracy. There is still a strong negative correlation between length of residence and production accuracy even when the effects of age are statistically subtracted.

Partial Corr

| Correlations | | | | | | |
|-------------------|-----|-------------------------|-----------------------------|-------|----------|--------------|
| Control Variables | | | | LOR | aptitude | R_L_accuracy |
| age | LOR | Correlation | | 1.000 | .027 | -.753 |
| | | Significance (2-tailed) | | . | .928 | .002 |
| | | df | | 0 | 12 | 12 |
| | | Bootstrap ^a | Bias | .000 | .024 | .008 |
| | | | Std. Error | .000 | .300 | .122 |
| | | | BCa 95% Confidence Interval | | Lower | |
| | | | | Upper | | |

Table 1 Output from a Partial Correlation in SPSS.

Summary Calculating Partial Correlations in SPSS

In the drop-down menu choose ANALYZE > CORRELATE > PARTIAL. Put the variable you want to control for in the CONTROLLING FOR box, and the other variables in the VARIABLES box. Open the BOOTSTRAP button and tick the “Perform bootstrapping” box. Change the type of confidence interval to “Bias corrected accelerated (BCa)”.

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).

Calling for Point-Biserial Correlations

In SPSS, this is easily done by choosing ANALYZE > SCALE > RELIABILITY ANALYSIS. Move the total test score and the dichotomous scores for each item to the ITEMS box on the right. Click the STATISTICS button, and be sure to check the box for “Scale if item deleted” under DESCRIPTIVES FOR. This will give you a box labeled Item-Total Statistics in the output, where you can see the Corrected Item-Total Correlation, which is the point-biserial correlation for each item. 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.

Calculating Point-Biserial Correlations

In the drop-down menu choose **ANALYZE > SCALE > RELIABILITY ANALYSIS**. Put the score for the total test and also the individual items in the “Items” box. Open the **STATISTICS** button and tick “Scale if item deleted.” If point-biserial correlations are low, you should probably think about eliminating these items from your test (but more modern methods of test analysis such as IRT are really better ways to consider your test data and I would urge you to find out how to use these).

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 4 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 10 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 calculate the intraclass correlation for a group of raters, go to `ANALYZE > SCALE > RELIABILITY ANALYSIS`. You will see the dialogue box for Reliability Analysis shown in Figure 2. Move the scores for your participants to the “Items” box. The columns you enter here should consist of the rating for each participant on a different row, with the column containing the ratings of each judge. Therefore, 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. Leave the “Model” menu set to ALPHA. Other choices here are SPLIT-HALF, GUTTMAN, PARALLEL, and STRICT PARALLEL, but what you want to call for is Cronbach’s coefficient alpha.

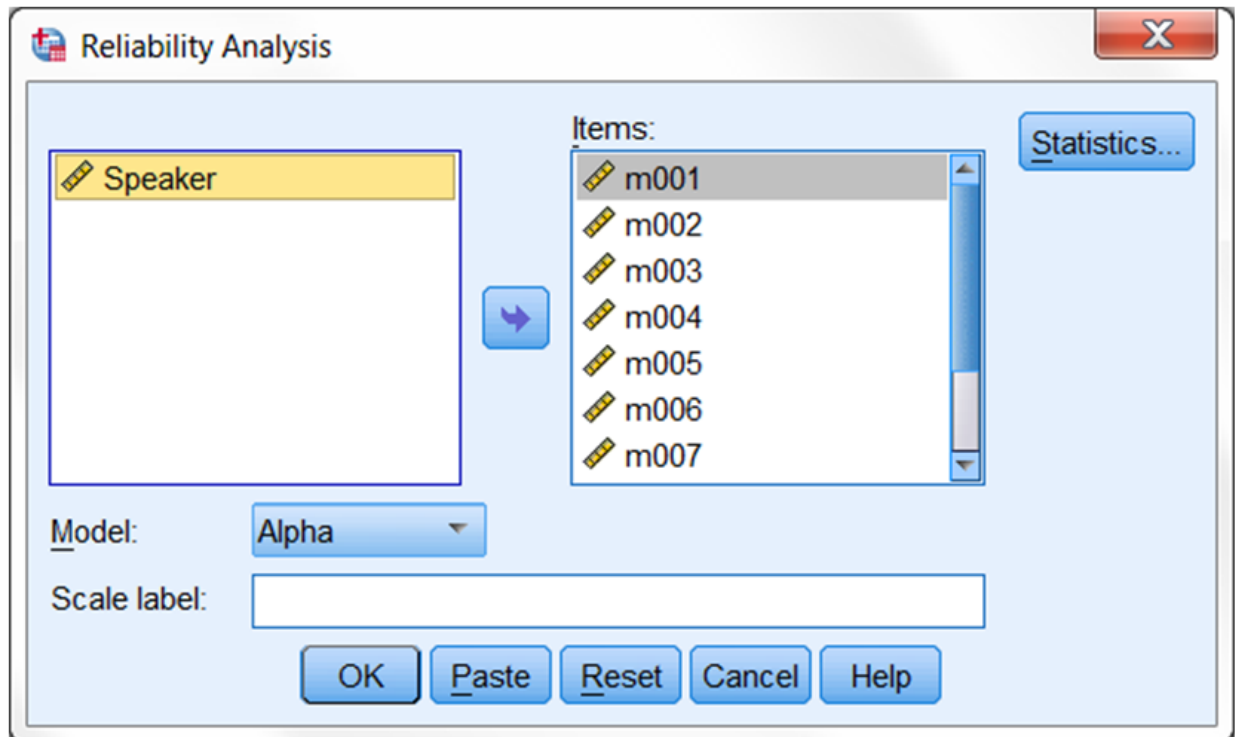
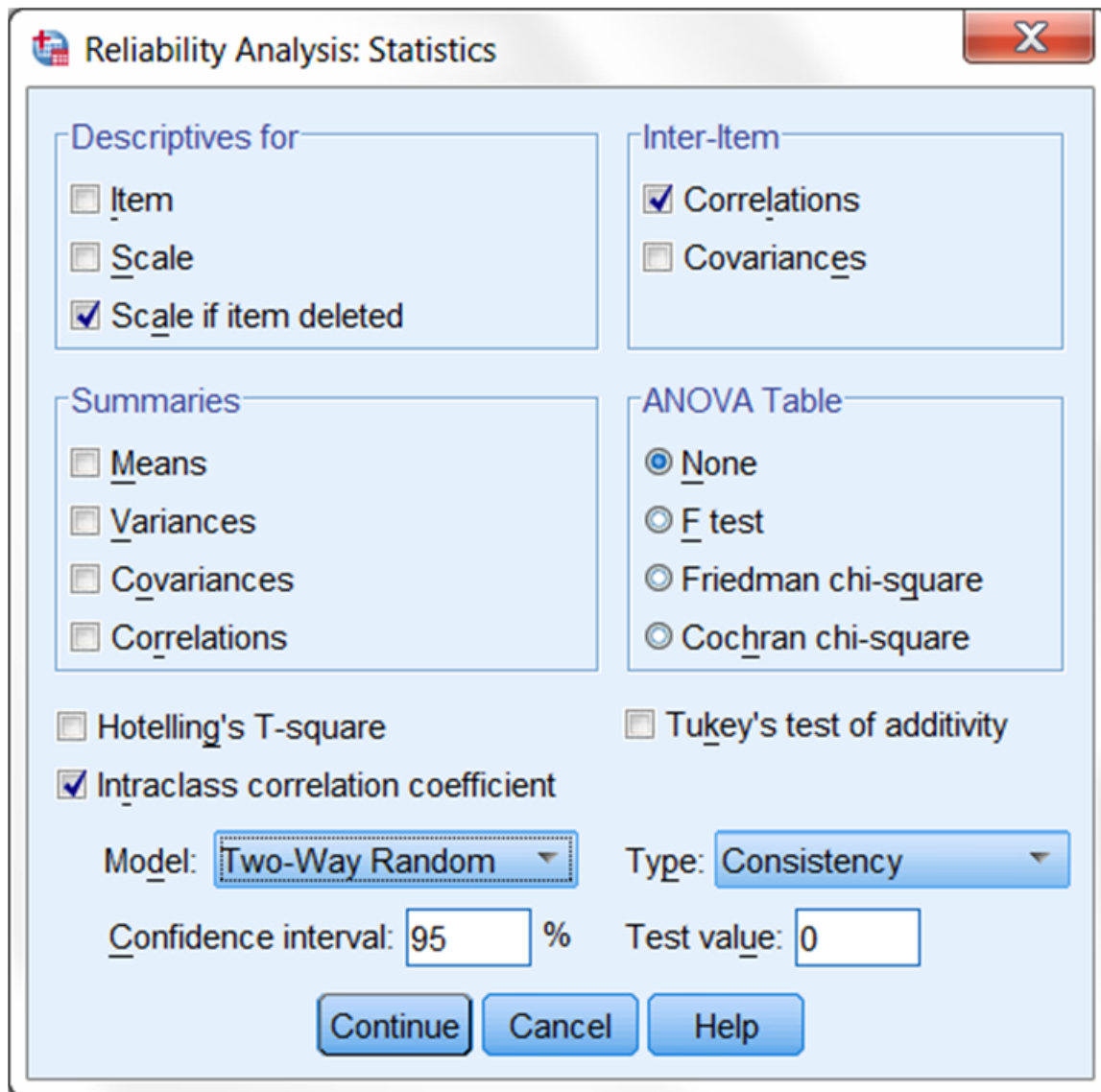


Figure 2 Dialogue box for Reliability Analysis in SPSS.

Next, open the STATISTICS button and you'll see the box in Figure 3. The most important thing to do here is to tick the "Intraclass correlation coefficient" box. When you do this, two drop-down menus will become visible. In the first one choose TWOWAY RANDOM. This choice specifies both the item effects (the judges/the columns) as random variable and the subject effects (the participants/the rows) as random as well. Since both the rows and the columns contain subjects, they are both random effects (we want to generalize to more than just the actual judges and more than just the actual participants; I discussed the difference between fixed and random effects in Section 2.1.6 of the book). You should also tick on the boxes "Descriptives for . . . Scale if item deleted" and "Inter-item correlations", as shown in Figure 3.

In the second drop-down menu you can choose whether you'd like a measure of CONSISTENCY or

ABSOLUTE AGREEMENT, but in truth this doesn't matter for the Cronbach's alpha result so just leave the default of CONSISTENCY chosen. Also tick the boxes that say "Scale if item deleted" and "Correlations."



The image shows the 'Reliability Analysis: Statistics' dialog box in SPSS. It is a standard Windows-style dialog box with a title bar, a close button (X), and several groups of options. The 'Descriptives for' group has checkboxes for 'Item', 'Scale', and 'Scale if item deleted' (checked). The 'Inter-Item' group has checkboxes for 'Correlations' (checked) and 'Covariances'. The 'Summaries' group has checkboxes for 'Means', 'Variances', 'Covariances', and 'Correlations'. The 'ANOVA Table' group has radio buttons for 'None' (selected), 'F test', 'Friedman chi-square', and 'Cochran chi-square'. Below these are checkboxes for 'Hotelling's T-square' and 'Tukey's test of additivity'. The 'Intraclass correlation coefficient' checkbox is checked. At the bottom, there are dropdown menus for 'Model' (set to 'Two-Way Random') and 'Type' (set to 'Consistency'). Below these are input fields for 'Confidence interval' (95 %) and 'Test value' (0). At the very bottom are 'Continue', 'Cancel', and 'Help' buttons.

| Group | Option | Status |
|------------------------|------------------------------------|----------------|
| Descriptives for | Item | Unchecked |
| | Scale | Unchecked |
| | Scale if item deleted | Checked |
| Inter-Item | Correlations | Checked |
| | Covariances | Unchecked |
| Summaries | Means | Unchecked |
| | Variances | Unchecked |
| | Covariances | Unchecked |
| | Correlations | Unchecked |
| ANOVA Table | None | Selected |
| | F test | Unselected |
| | Friedman chi-square | Unselected |
| | Cochran chi-square | Unselected |
| Other Tests | Hotelling's T-square | Unchecked |
| | Tukey's test of additivity | Unchecked |
| Intraclass Correlation | Intraclass correlation coefficient | Checked |
| | Model | Two-Way Random |
| Type | Type | Consistency |
| | Confidence interval | 95 % |
| Test Value | Test value | 0 |

Figure 3 Statistics for the reliability analysis in SPSS.

The first box you will see in the output will just be a summary of how many cases were

analyzed. Of course you should check this to make sure that all the cases you thought were going to be analyzed actually were (there were 48 in the Munro, Derwing, & Morton data). The last box in the input contains Cronbach's alpha, which is the major item you are interested in, although it is not labeled as such, but just as "Intraclass Correlation Coefficient" (see Table 2). Using the line that says "Average Measures", we see that Cronbach's alpha is 0.89, 95% CI [.83, .93]. This is a high correlation considering that there are ten items (judges).

| Intraclass Correlation Coefficient | | | | | | | |
|------------------------------------|-------------------------------------|-------------------------|-------------|--------------------------|-----|-----|------|
| | Intraclass Correlation ^b | 95% Confidence Interval | | F Test with True Value 0 | | | |
| | | Lower Bound | Upper Bound | Value | df1 | df2 | Sig |
| Single Measures | .434 ^a | .327 | .562 | 8.681 | 47 | 423 | .000 |
| Average Measures | .885 | .829 | .928 | 8.681 | 47 | 423 | .000 |

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

Table 2 Cronbach's alpha output from the reliability analysis in SPSS.

Remember that we do not have an absolute rule of thumb for determining what an acceptable level of Cronbach's alpha is, and we should look at the correlations between pairs of variables, and this is shown in the part of the output labeled "Inter-item Correlation Matrix", shown in Table 3.

| Inter-Item Correlation Matrix | | | | | | | | | | |
|-------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | m001 | m002 | m003 | m004 | m005 | m006 | m007 | m008 | m009 | m010 |
| m001 | 1.000 | .352 | .341 | .413 | .231 | .236 | .337 | .385 | .407 | .323 |
| m002 | .352 | 1.000 | .333 | .530 | .411 | .478 | .518 | .614 | .507 | .538 |
| m003 | .341 | .333 | 1.000 | .599 | .427 | .411 | .325 | .393 | .578 | .606 |
| m004 | .413 | .530 | .599 | 1.000 | .434 | .615 | .514 | .497 | .443 | .619 |
| m005 | .231 | .411 | .427 | .434 | 1.000 | .366 | .309 | .250 | .309 | .481 |
| m006 | .236 | .478 | .411 | .615 | .366 | 1.000 | .553 | .551 | .309 | .672 |
| m007 | .337 | .518 | .325 | .514 | .309 | .553 | 1.000 | .577 | .387 | .586 |
| m008 | .385 | .614 | .393 | .497 | .250 | .551 | .577 | 1.000 | .513 | .580 |
| m009 | .407 | .507 | .578 | .443 | .309 | .309 | .387 | .513 | 1.000 | .547 |
| m010 | .323 | .538 | .606 | .619 | .481 | .672 | .586 | .580 | .547 | 1.000 |

Table 3 Inter-Item Correlation Matrix from a Reliability Analysis.

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 thus 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.

The last piece of output I want to look at is shown in Table 4. This is the part of the output that shows what Cronbach’s alpha would be if each item (judge) individually were removed. If judges are consistent then there shouldn’t be too much variation in these numbers, and this is true

for the Munro, Derwing, and Morton (2006) data. However, if there were a certain judge whose data changed Cronbach's drastically you might consider throwing out that judge's scores.

| Item-Total Statistics | | | | | |
|-----------------------|----------------------------|--------------------------------|----------------------------------|------------------------------|----------------------------------|
| | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Squared Multiple Correlation | Cronbach's Alpha if Item Deleted |
| m001 | 51.48 | 172.893 | .449 | .265 | .889 |
| m002 | 50.04 | 169.530 | .668 | .529 | .870 |
| m003 | 50.46 | 185.020 | .613 | .559 | .877 |
| m004 | 52.63 | 158.495 | .732 | .594 | .864 |
| m005 | 50.44 | 190.294 | .484 | .321 | .883 |
| m006 | 51.21 | 157.317 | .660 | .576 | .872 |
| m007 | 53.13 | 175.601 | .655 | .479 | .872 |
| m008 | 52.42 | 163.355 | .690 | .558 | .868 |
| m009 | 51.40 | 182.457 | .608 | .512 | .876 |
| m010 | 52.44 | 158.719 | .776 | .671 | .861 |

Table 4 Item-Total Statistics Output from a Reliability Analysis.

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 LarsonHall2008 described in Section 6.5.4 of the book 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 have deleted the scores of native speakers of Japanese on this test, as I think native speakers may score quite differently from learners of Japanese.

Use the same reliability analysis as for the inter-rater reliability (ANALYZE > SCALE > RELIABILITY ANALYSIS). Here I will enter all 40 of my items into the “Items” box as shown in Figure 3. If all I want is to get Cronbach’s alpha, there is no need to open the STATISTICS button (the boxes you might tick in the STATISTICS button to look at item-total statistics and inter-item correlation would be a way of doing test analysis, although a mostly outdated one now). The output gives a Cronbach’s alpha of 0.67, which can also be reported as a KR-20 score of .67. 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 the R statistical program 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 the drop-down menu choose ANALYZE > SCALE > RELIABILITY ANALYSIS. Put all the items that contain judge’s ratings of the participants in the “Items” box. Open the STATISTICS button and tick the “Intraclass correlation coefficient” box. In the first drop-down menu choose TWO-WAY RANDOM, but leave the other drop-down menu alone. Also tick “Scale if item deleted” and “correlations”. Look for Cronbach’s alpha in the output.

For overall test reliability simply put all of your test items (coded as 0s and 1s) into the “Items” box in the Reliability analysis and obtain Cronbach’s alpha, which you can also call the KR-20 measure of reliability.

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