Questions and answers about language testing statistics:

Calculating reliability of dictation tests: Does K-R21 work?

James Dean Brown brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

For many tests like multiple-choice, true-false, and fill-in, we have item statistics which we can use in calculating reliability statistics like K-R20 and alpha. But for dictations, we only count-up total scores. So, my question is this: (a) can we use K-R21 based on the mean, standard deviation, and number of items for the total scores to calculate the reliability of a dictation, and (b) if so, how long should a dictation be in order to be reliable?

Answer:

This is the first of two columns that I will use to answer your questions. In the next column, I will discuss the relationship between dictation length and reliability. In this one, I will explore some problems and solutions for calculating the reliability of dictations. To do so, I will address four central questions:

1. What data serve as the basis for the current column?
2. What are some options for calculating reliability for dictations and what are the relationships among them?
3. What else is important in interpreting these common reliability estimates?
4. What does all this mean for calculating the reliability of dictation scores?

What data serve as the basis for the current column?

The participants were 220 graduate and undergraduate students taking the English Language Institute Placement Test (ELIPT) at UHM in fall 2015 and spring 2016. Their scores on the Internet-based TOEFL ranged roughly from 61 to 100.

Three dictations were involved here [For fuller descriptions, see Brown & Trace (2018).]:

1. The traditional academic dictation test (DCT) was a 50-word passage on a general academic topic that was recorded by a male native speaker of English who read three times: once at regular speed, once with pauses, and once again at regular speed. The 50 words were scored one point each. Spelling was not counted if the word was morphologically correct.
2. The connected-speech narrative (CSN) dictation was based on a passage about international student life in the US (from Prator & Robinett, 1972). It was recorded in much the same way as the DCT. However, the speaker was told to speak use connected speech just as in natural speech. While the passage contained 187 words, only the first 50 words involved in connected speech were scored as correct or incorrect. Connected speech was defined as changes from the dictionary pronunciation of the words including instances of adding, dropping, transitioning, or changing sounds. Otherwise, this dictation was scored the same as the DCT.

3. The connected-speech conversation (CSC) dictation was a 10-turn informal dialogue between a male and female about travel plans (based on Brown & Kondo-Brown, 2006), both speakers purposely used connected speech as appropriate. The CSC contained 83 words, 50 of which were scored because they involved connected speech. Otherwise, this dictation was scored the same as the DCT.

Note that we took the unusual step for all three dictations of compiling item level data, where each word was scored right or wrong and represented one item.

What are some options for calculating reliability for dictations and what are the relationships among them?

Since responsible interpretation of reliability estimates depends on descriptive statistics, Table 1 presents the mean, standard deviation (*SD*), minimum score, maximum score, and range for each of the three sets of dictation scores being considered here: the DCT, CSN, and CSC.

Table 1
Descriptive Statistics for the DCT, CSN, and CSC Dictations

Statistic	DCT	CSN	CSC
Mean	31.40	30.35	38.09
SD	8.34	8.84	6.58
Minimum score	11	14	16
Maximum score	50	50	50
Range	40	37	35

Notice in Table 1 that the CSC dictation has the highest mean at 38.09, and that the DCT and CSN were considerably lower at 31.40 and 30.35, respectively. This probably means that the conversational English in the CSC was easier for L2 learners to understand. Notice also that the *SD* for the CSC is considerably lower at 6.58, than those for the DCT (8.34) and CSN (8.84). This means that the scores on the CSC were less widely dispersed than those on the DCT and CSN. The maximum values of 50 indicate that at least one of the examines scored the highest possible score of 50 on each of the three dictations. The minimum values indicate the lowest scores were 11, 14, and 16, respectively. The range is another indicator of the relative spread of scores with the DCT being the widest (40) and the CSC the narrowest (35) and the CSN in between (37).

Table 2 shows four reliability estimates each for the DCT, CSN, and CSC dictations: (a) Kuder-Richardson formula 20 (K-R20); (b) Cronbach alpha; (c) Kuder-Richardson formula 21 (K-R21); and (d) split-half adjusted based on odd and even scores (for more on these various reliability estimates, see Brown, 2005, pp. 169-198, or Brown, 2016, pp. 107-153). Notice that K-R20 (.89, .90, & .87) and alpha (.89, .89, & .87) estimates are very similar for each of the dictations, which is what theory would predict. Notice also that the K-R21 estimates are systematically the lowest at .85, .87, and .81, respectively, and that the split-half adjusted estimates are the highest at .93, .92, and .91, respectively. In the next section, I will consider each of these four internal consistency reliability statistics in more detail.

Table 2
Four Types of Reliability Estimates for the DCT, CSN, and CSC Dictations

Reliability Estimate	DCT	CSN	CSC
K-R20	.89	.90	.87
Cronbach alpha	.89	.89	.87
K-R21	.85	.87	.81
Split-half adjusted (odd-even)	.93	.92	.91

What else is important in interpreting these common reliability estimates?

The *K-R20* formula (originally from Kuder & Richardson, 1937) is based on item- and test-level statistics, and it assumes "that the matrix of inter-item correlations has a rank of one and that all intercorrelations are equal" (p. 156). These assumptions are satisfied on a test where all items are measuring the same factor. Thus, unidimensionality is assumed. An additional design condition for K-R20 is that the items must be scored dichotomously (e.g., right or wrong).

Cronbach alpha is similar to K-R20 in that both are based on item- and test-level statistics, and alpha also assumes unidimensionality, but alpha has the advantage over K-R20 of being applicable to scales that are not dichotomous like weighted items (e.g., 0 = wrong, 1= partial credit, and 2 = completely correct), Likert items (e.g., 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree), etc.

The *K-R21* formula is based here on just three test-level statistics (the mean, standard deviation, and number of items), and it assumes unidimensionality, but also "that all items have the same difficulty" (Kuder & Richardson, 1937, p. 158). That last assumption is met if the item facility values on a test are approximately equal as on a test developed by selecting items from a pilot version that have item facility values ranging from .30 to .70 in item facility (i.e., the item difficulties are approximately equal) for a final version of the test (see Brown, 2005, pp. 66-68; Brown, 2016, pp. 63-67).

The *split-half adjusted* approach is based on scoring the odd-numbered and even-numbered items separately—producing two separate scores for each examinee and then calculating the correlation between the two sets of scores to find the half-test reliability and adjusting for full-test reliability using the Spearman-Brown prophecy formula (see Brown, 2005, pp. 177-179; Brown, 2016, pp. 125-130).

One way or another, all the internal consistency statistics discussed here estimate the degree to which items on a test are interrelated. And, one assumption statistically and logically of these estimates is unidimensionality, which means that the items on the test should all be measuring the same construct. While items on a test will most often be interrelated, that does not necessarily mean they are unidimensional as a set (as we will see below). So, in a sense, interrelatedness is a precondition, but is not sufficient in itself to assure unidimensionality. It would therefore be a mistake to be satisfied with a high degree of item interrelatedness (i.e., reliability) without also examining unidimensionality.

One way to examine the unidimensionality of a set of items is to run a factor analysis to see how many factors underlie what they are testing. A factor analysis—actually, a principle components analysis (PCA)—was performed for each of the three dictations with the 50 items in each case as the variables (for much more on factor analysis, see Brown, 2016, pp. 237-276). As shown in Table 3, the PCAs for DCT, CSN, and CSC produced 17, 16, and 18 Eigen values above 1.00, respectively, which accounted for 67.8, 67.8, and 72.4 percent of the variance, respectively. This is a clear indication that a number of components (or dimensions) underlie whatever these dictations are testing. Hence, all three seem to violate the assumption of unidimensionality, and as a result the reliability statistics found here probably provide

underestimates of the reliability of the scores. Such estimates are often referred to as lower-bound estimates of reliability, that is, the reliability of the scores will be at least as high as the estimate but may be higher.

Table 3
Item Facility, Eigen Values, and Percentage of Variance for the DCT, CSN, and CSC Dictations

Statistic	DCT	CSN	CSC
Number of Eigen values over 1.00	17	16	18
Percentage of variance accounted for	67.80	67.80	72.40
IF minimum	.10	.09	.05
IF maximum	1.00	1.00	1.00

Also recall that the K-R21 assumes that the item facility values on a test are approximately equal—say between .30 and .70. Table 3 indicates that the IF values ranged much more widely than that: from .10 to 1.00, .09 to 1.00, and .05 to 1.00, respectively. These violations of the assumption of equal difficulty may explain why the K-R21 estimates shown in Table 2 are consistently lower than all the other estimates. Brown (1983) reported similar but much bigger underestimates for K-R21 for cloze tests, where the assumption of equal difficulty is also violated.

Local item independence is seldom written about in language testing, but important for thinking about the reliability estimates for dictations. Essentially, many item and test statistics require that the items be independent, which is to say that they should not be correlated for reasons other than the fact that they are testing the same construct (for more on this, see Yen, 1993). This may be a problem when five items are based on the same reading or listening passage because being based on the same passage may cause items to be related for reasons other than the fact that they are testing the same construct. Local independence may also be a problem in cloze tests because answering one item correctly may help (for reasons beyond the construct being tested) answer the next blank because more context is available. The second study in Brown (1983, pp. 243-250) was an experiment that showed that, if lack of local independence is a problem for cloze procedures, it does not affect reliability.

Dictations may have the same problem of lack of local independence because writing down one word correctly may help (for reasons beyond the construct being tested) guess/know the next or other words because more context is provided. The net effect of dictations lacking local item independence might be that such reliability estimates would provide inflated estimates of the true state of affairs. This is troubling, and unfortunately, there is no research on this issue for dictations that I am aware of.

Interestingly, since the odd- and even-item scores used to calculate split-half adjusted reliability are more independent (i.e., items are not right next to each other) than the individual items (right next to each other) used in calculating the K-R20, alpha, and K-R21, I would expect any such inflation of the reliability estimates to affect the split-half adjusted estimates to a lesser degree than the others. Yet, it turned out that the split-half adjusted estimates were higher than the others. Thus, it appears that the magnitude of any inflation due to lack of independence, may be less than the magnitude of any underestimation due to lack of unidimensionality. This conclusion is based on relatively small data sets and a small number of dictations, so further research is clearly warranted before accepting any such conclusion.

For testers who are worried about this issue, a test-retest or parallel-forms reliability estimates (see Brown, 2005, pp. 175-176) would get around this problem—at least for statistically estimating reliability.

What does all this mean for calculating the reliability of dictation scores?

So, what are language testers to do if they want to know how reliable the scores on their dictations are? If they are concerned about lack of unidimensionality, the internal consistency estimates used here will work fine, but must be interpreted as underestimates. In any case, interpretation should be done cautiously (as in Brown, Phung, Hsu, Trace, Harsch, & Faucette, 2018, p. 24, where we put a footnote in the table containing our DCT reliability estimate as follows: "**K-R21 = very rough estimate"). In addition, the fact that dictations are clearly measuring multiple dimensions must be considered.

If language testers are concerned about lack of local independence, they could use either a test-retest or parallel-forms approach to calculate reliability because those approaches are based on independent total scores. These approaches involve considerably more work (for both the tester and examinees) than any of the internal consistency reliability estimates reported here, but they do avoid the local independence problem.

Conclusion

In this column, I (a) described the data that serve as the basis for the discussion in this column; (b) provided some options for calculating the reliability of dictations and looked at the relationships among them; (c) considered other important issues involved in interpreting these reliability estimates; and (d) suggested what all this means for calculating the reliability of dictation scores. In the process, for the internal consistency estimates presented here, it became clear that (a) they likely violate the assumption of unidimensionality and (b) that they also lack local independence across items. I ended by suggesting strategies for dealing with both problems. In direct answer to our question, yes, you can use K-R21, but only very cautiously while taking into account the issues discussed in this column.

I hope this column addressed the first part of your question adequately and provided you with the information you need for calculating and describing the reliability of your dictation tests in the future. [For much more on calculating, describing, and using reliability estimates see Brown (2005, pp. 169-198).] In the next column, I will explore the relationship between reliability and dictation length.

References

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J.W. Oller, Jr. (Ed.) *Issues in Language Testing Research* (pp. 237-250). Rowley, MA: Newbury House (also available from ERIC: ED 227 695).

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New edition). New York: McGraw-Hill.

Brown, J. D. (2016). *Statistics corner: Questions and answers about language testing statistics.* Tokyo: JALT.

Brown, J. D., & Kondo-Brown, K. (2006). Testing reduced forms. In J. D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 247-264). Honolulu, HI: University of Hawaii, National Foreign Language Resource Center.

Brown, J. D., & Trace, J. (2018). Connected-speech dictations for testing listening. In G. Ockey & E. Wagner (Eds.). *Assessing L2 listening: Moving towards authenticity* (pp. 45-63). Philadelphia, PA: John Benjamins

Brown, J. D., Phung, H., Hsu, W-L, Trace, J., Harsch, K., & Faucette, M. P. (2018). 2016-2017 English Language Placement Test (ELIPT) revision project. *Second Language Studies, 36*(2), 1-25.

Kuder, G. F. & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*, 151-160.

Prator, C. H., & Robinett, B. W. (1972). *Manual of American English pronunciation*. New York: Holt, Rinehart, & Winston.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Where to submit questions:

Your question can remain anonymous if you so desire. Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.
JD Brown
Department of Second Language Studies, University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822 USA