Shiken: JALT Testing & Evaluation SIG Newsletter, 2 (1) October 1998 (p. 18-21)

Statistics Corner

Questions and answers about language testing statistics:

Cloze Tests and Optimum Test Length

James Dean Brown (University of Hawai'i at Manoa)

QUESTION: I am interested in including a cloze section in the entrance exam for our university. I know that the longer the test (that is, the more items) the more reliable, but I also need to keep the test as short as possible. Is there a point at which the increasing length of the test has diminishing returns? If so, how can I figure that out?

ANSWER: You are absolutely right that, in general, all other factors held constant, longer tests tend to be more reliable than shorter ones (as explained in Brown, 1996, pp. 194-195). Naturally, there is a point at which adding more items can have the opposite effect, as fatigue sets in and the students begin to get discouraged or stop taking the test seriously. However, unfortunately, reliability is not just a function of test length with cloze tests. At least three factors substantially affect the reliability of cloze tests: (a) variations in student ability levels and score ranges, (b) differences in passage difficulties, and (c) changes in numbers of items.

Variations in Student Ability Levels and Score Ranges

Reliability results for any test can vary markedly across groups of students who have different proficiency levels or ranges of language abilities. In Brown (1984), I showed how cloze test results could vary considerably for different groups. The results of that study are summarized in Table 1. Notice in Table 1 that the reliabilities (in this case, Cronbach alpha estimates) range from .31 to .95 in the fourth row. Thus, this table reports one of the highest reliability estimates ever published for a cloze test (at .95) and one of the lowest (.31). The startling thing is that all of these results are based on exactly the same cloze passage, but it was administered to different groups of students (at UCLA and Zhongshan University in Cuangzhou, China). So you can see that reliability results can vary considerably for different groups of students. More interesting, these results do not vary solely because they are produced by different groups of students. Notice how the standard deviations and ranges in the second and third columns of Table 1 systematically decrease as you look down the table in much the same way as the reliability estimates in the fourth row.

Table 1. *Ranges of talent in relationship to reliability and validity of cloze (adapted from Brown, 1983).*

Scoring Method/Sample	AC/1978	EX/1978	AC/1981	EX/1981	EX/1982a	AC/1982b	AC/1982c	EX/1982d
standard deviation	12.45	8.56	6.71	5.59	4.84	4.48	4.07	3.38
range	46	33	29	22	22	20	21	14
r_{xx}	.95	.90	.83	.73	.68	.66	.53	.31
r_{xy}	.90	.88	.79	.74	.59	.51	.40	.43

Clearly, the magnitude of the reliability estimates is directly related to the amount of variation in abilities in the various groups as that variation is reflected in the standard deviation and range of scores.

Differences in Passage Difficulties

Reliability estimates can also vary markedly for cloze tests developed from passages of varying difficulty. In Brown (1993), I showed how passage difficulty is related to reliability by administering 50 cloze tests to equivalent groups (2298 Japanese university students were randomly assigned to the 50 passages). The results from that study are summarized in Table 2.

Notice that the reliability estimates in the last row of Table 2 range from .283 to .816. Notice also that the standard deviations and ranges in the third and fourth rows systematically decrease as you look down those rows in much the same way that the reliability coefficients decrease.

Table 2. *Passage Difficulty in Relationship to Reliability of Cloze (adapted from Brown, 1993; arranged from high to low reliability).*

test	21	12	37	50	27	45	20	2	31	4	22
	11	35	15	9	6	7	1	41	42	49	19
	10	5	32	14	44	30	8	3	47	34	23
	24	48	25	28	33	39	29	38	16	46	43
	36	40	13	18	17	26					
mean	9.92	8.98	5.46	2.49	2.34	6.55	4.38	4.21	3.78	7.54	3.70
	5.94	6.63	9.18	2.85	5.11	6.14	5.23	2.87	4.41	4.56	4.76
	2.54	3.98	3.83	3.23	3.24	9.56	3.16	2.02	3.79	5.87	3.64
	2.96	2.69	5.36	2.58	2.14	2.51	2.32	1.71	1.36	2.16	1.43
	5.00	3.49	2.87	1.02	1.38	2.68					
SD	4.44	3.97	3.66	2.70	2.72	3.87	3.24	3.42	3.08	3.87	2.86
	3.36	3.66	3.42	2.46	3.23	3.41	3.16	2.51	3.10	2.81	2.88
	2.31	2.79	2.53	2.50	2.52	3.28	2.27	2.13	2.33	2.91	2.40
	2.26	2.12	2.74	2.17	1.87	1.98	1.77	1.57	1.41	1.82	1.45
	2.05	1.90	1.71	1.09	1.25	1.56					
range	19	21	13	12	13	16	15	13	15	14	11
	16	17	14	11	14	16	15	10	18	11	10
	8	13	9	9	10	13	8	10	11	13	11
	9	11	12	8	6	9	7	8	6	7	7
	9	9	8	3	5	5					
alpha	0.816	0.798	0.788	0.775	0.774	0.766	0.763	0.762	0.750	0.748	
	0.742	0.734	0.733	0.727	0.724	0.723	0.716	0.711	0.711	0.710	
	0.696	0.682	0.675	0.663	0.663	0.656	0.653	0.645	0.644	0.643	
	0.638	0.637	0.622	0.622	0.612	0.607	0.607	0.604	0.547	0.533	
	0.532	0.500	0.482	0.465	0.452	0.434	0.347	0.317	0.313	0.283	

Similarly, but to a somewhat lesser degree, the means shown in the second row of Table 2 also decrease similarly to the reliability estimates. A relatively high mean on a cloze test would indicate that it is based on a relatively easy passage because the students are scoring higher on average, and a low mean would indicate a relatively difficult passage because the students are scoring lower on average. Thus there appears to be some relationship between the difficulty of the cloze passages and the reliabilities that result when students of similar ability take them as cloze tests. In Table 2, the most difficult passages appear to be producing skewed distributions, a sort of "floor" effect, and are also producing commensurately low reliability estimates (as low as .283 in this case), while the easiest passages appear to be better centered and are also producing higher reliability estimates (as

high as .816 in this case). Thus, in addition to varying across groups of different ability levels and ranges, cloze test reliability estimates can vary according to how well the difficulty of the particular passage is related to the ability levels of the group of students being tested.

Changes in Numbers of Items

One study by Rand (1978) at UCLA directly addressed the issue of the effect of test length on the reliability of cloze tests. Based on a cloze passage that I developed for my MA thesis (Brown, 1978, also reported in Brown, 1980), Rand calculated the reliabilities for various lengths of that cloze test for three scoring methods: exact-answer (wherein only the original word found in the blank is counted as correct), acceptable-answer (wherein any answer judged as correct by native speakers for a give blank is counted as correct), and multiple-choice (wherein students are give three or four choices to select from for each blank). He found that the reliability for all three scoring methods began to "level off" at 20 items and the maximum reliability was nearly achieved at 25 items. He rather rashly concluded that: "This study has shown that the examiner using close tests can most efficiently use his own resources and his examinees' time by giving a cloze test of only twenty-five deletions and employing the acceptable-word method of scoring" (pp. 65-66). Unfortunately, Rand based his results on a single cloze test and single population; he therefore overlooked the importance of variations in student ability levels and score ranges as well as differences in passage difficulties.

Determining the Best Length of Cloze Test for Your Students

Clearly then, cloze tests can vary dramatically depending on which group of students is taking the test and which passages are being used. To answer your question directly, in order to determine the best length of cloze test for your students, given a particular passage, you might want to take the following steps:

* Develop a cloze passage of say 50 items that is at the correct level of difficulty (using the hit-or-miss, modification, or well-tailored cloze approach described in Brown, 1988).

* Administer the test to students of the same abilities and range as those for whom the test will eventually be used to make decisions.

* Score the test using whichever scoring method makes more sense in your situation (for instance, the exact-answer scoring may make more sense if you are doing large scale placement testing and the students will never see the tests again, or the acceptable-answer scoring method may make more sense if you are going to give the tests back to the students and go over the answers).

* Enter the results as 1s and Os (1 for correct answers, O for incorrect or blank answers) in your spreadsheet program.

* Use a K-R20 formula to calculate the reliability of the entire test with all items included (after Brown, 1996, p. 199-202). Then copy and modify that formula so that you figure the reliabilities for decreasing lengths of the cloze test (by excluding the last item from the previous analysis each time you do so). In other words, figure out the reliability for the entire 50 item cloze, then for a 49 item cloze (by excluding the last item), then for a 48 item cloze, a 47 item cloze, etc.

* Examine the reliability estimates that you get for the various lengths of cloze test (either by simply inspecting the values or by using the spreadsheet to plot reliability vs. length in a scatter plot). Determine the reliability at which it is no longer worthwhile to add more items. In other words, find the reliability above which little more reliability is gained by adding more items.

* Retype the cloze test so that it has only the number of blanks you have determined to be best.

Following the above steps will allow you to create a cloze test that is maximally efficient for your particular group of students. Good luck with it!

References

Brown, J. D. (1978). Correlational study of four methods for scoring cloze tests. Unpublished master's thesis, University of California at Los Angeles, Los Angeles, CA.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal, 64* (3), 311-317.

Brown, J. D. (1984). A cloze is a cloze is a cloze? In J. Handscombe, R. Orem & B. Taylor (Eds.) *On TESOL '83: The question of control. Selected papers from the 17th Annual TESOL Convention, Toronto* (pp. 109-119). Washington, DC: TESOL.

Brown, J. D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5* (1), 19-31.

Brown, J. D. (1993). What are the characteristics of natural cloze tests? *Language Testing, 10* (2), 93-116.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Rand, E. (1978). The effects of test length and scoring method on the precision of cloze test scores. *UCLA Workpapers in Teaching English as a Second Language,* 12, pp. 62-71.

HTML: http://www.jalt.org/test/bro_3.htm / PDF: http://www.jalt.org/test/PDF/Brown3.pdf