

A Comparison of Three Modes of
Administering Listening
Tests

By

Deborah Roach
Department of Communication
University of Oklahoma

and

Margaret Fitch Hauser
College of Business Administration
University of Oklahoma

Presented to

The International Listening Association
Fifth Annual Meeting
Scottsdale, Arizona
March 9, 1984

(The authors wish to acknowledge the contributions of Dr. Wayland Cummings to this paper. His support and guidance through the Statistical Analysis is most appreciated. We also are grateful to Joe Hukels, who did an excellent job on the narrations, and to Don and Lisa in the Instructional Services Department for going beyond the call of duty to produce the tapes for us.)

A COMPARISON OF THREE MODES OF PRESENTING LISTENING TEST RESPONSES

An examination of various listening tests reveals the existence of very few commonalities among the current tests available on the "market." This lack of commonality among existing listening tests includes differences in the dimensions of listening operationalized by listening test authors, differences in modes of presenting item or message stems, and differences in the modes of administering item responses. Although the literature in the area of listening addresses the first two of these three areas, the third area and the area in which these authors are interested has not been addressed in the current literature--namely, the possible differences which might emerge when mode of presenting message responses varies from the way in which the item or message stems initially are presented. Because the many listening tests available vary with regard to presentation of mode of listening responses, an exploratory study in this area of listening research seems warranted. Given the paucity of literature in this area, a study of this phenomenon was undertaken by the authors.

Literature Review

Concern regarding the operationalization of listening dimensions has been expressed by many researchers in the field of listening. More recently, this concern has been detailed by Brown et. al, (1979), in their comprehensive review of listening literature. These authors concluded that many listening tests tap different dimensions than those tapped by alternative listening examinations. For example, many appear to tap such dimensions as comprehension, memory and ability to follow directions while others tap such dimensions as

verbal abilities or skills. These researchers believed that such a lack of consistency among listening tests, in part, emerged from our inability to separate theoretically the concepts of listening and reading and, hence, our subsequent inability to operationalize these concepts. Other researchers, such as Lundsteen (1971), however, have argued that the two skills are inexorably tied together. Her argument is supported conceptually by Kintsch and Kozminsky (1977), who argued that the link connecting listening and reading is a common core of comprehension processes which underlie the two abilities or skills. Kintsch and Kozminsky argued that the link did not imply that listening and reading are the same. However, the relationship for these researchers appeared to be validated by research reporting significant correlations between listening and reading tests. For example, Brown (1965) had found correlations ranging from .76 to .82 between listening and reading test scores. However, other researchers (e.g., Mead, 1978) have also linked listening scores with measures of verbal ability.

Perhaps, the strongest link between the skills of listening and reading is, in fact, their common link to general verbal skills, as Mead (1978) intimates. Brown, et.al., (1979a) likewise support this conceptual link in their break down of verbal skills into two general categories: expressive and receptive skills. Expressive skills, for these authors, included speaking and writing skills, while receptive skills included listening and reading. The researchers further detailed areas of commonality between listening and reading. These areas included recognizing and identifying the meaning of words and phrases; identifying and understanding main ideas; associating important details with main ideas; determining communication purposes; and following directions. Although the authors believed that listening and reading share these skills, they likewise postulated that any assessment of these two variables must account for significant differences between oral and written

language. Hence, skills that are unique to reading and to listening must equally be accounted for.

Many listening tests now in existence have not paid heed to this directive regarding the operationalization of listening and reading skills. For example, a review of listening tests by Daly, Neville, and Pugh (1975) reveals that many of the passages used in listening tests utilize language much more like written language than spoken language. This indictment that some listening tests do not use typical "spoken language" is a valid one and suggests that the responses to the messages may be influenced by the type of language used. Thus, the correlation between listening and reading tests necessarily will be high.

Another area in which little commonality exists among listening tests is the manner in which the stems of messages are administered to subjects. Two comprehensive reviews of listening tests by Brown, et.al., (1979b) and Daly, Neville, and Pugh (1975) reveal two predominant methods by which current tests present a stimulus message. One method, used in tests such as the Peabody Picture Vocabulary Test, the STEP Listening Test, and some versions of the Brown-Carlson Listening Comprehension Tests, directs test administrators to read the passages aloud. This method of presentation has received much criticism due to the possible influence of administrators on the interpretation of stimulus messages presented. In turn, such modes of presenting listening tests might influence Ss' responses.

In order to eliminate any source or experimenter effects, other listening tests have been developed which present all messages and test item stems on audiotapes or phonograph records. Certain versions of the Brown-Carlson test, Nancy Wallner's test for Listening Comprehension, and the Xerox Effective Listening Test are a few examples of tests utilizing this mode of presenting

message stems.

Just as the mode of presentation of message stems is not consistent across existing and current listening tests, the mode of administering test item responses also differs across tests. Unlike the other two areas, however, differences in administering test responses have not been addressed in this literature. Currently, however, three basic methods appear to be prevalent. The first method, and seemingly the most widely used method of administering responses, allows the respondents to read or look at item responses following the presentation of a message and a corresponding item stem. Tests which use this mode of presentation include the Cooperative Primary Test, the Durrell-Sullivan Reading Capacity and Achievement Tests, the Jones-Mohr test and Fristoe and Woodcock Listening Test.

A second mode of administering test responses, and used in a number of listening tests currently available, allows the respondents to listen as well as to read the possible responses. Two examples of this type of listening test include the Learning through Listening Test and the STEP Listening Test.

The third mode, and seemingly the purest mode of presenting listening responses, administers both the item responses and message stems orally. In this type of testing situation, Ss listen to the stems, then listen to the responses, and finally write down or select the correct (or "best") choice among responses. One version of the Xerox Effective Listening Test is administered in this fashion.

This lack of consistency among modes of administering listening tests has not been addressed by researchers in the area of Listening. If one accepts Brown, et al.'s notion that a distinction must be made between the testing of oral and written language, and that reading and listening tests should tap different skills, one must also consider the possibility that a response to a question will be influenced by whether the answers are heard, read, or both.

Based on this concern, the following research question will be addressed:
Does the mode of administering test responses to subjects influence the outcome
of listening test scores?

METHODS

Design

In order to explore the possible differences among modes of administering listening test responses, a 1 X 3 analysis of variance design was employed. The independent variable, modes of presentation, was operationalized via the newly developed Watson-Barker Listening Test. This measure is an audiotaped, 50 item listening test designed to tap five dimensions of listening. Due to its very recent development, no validity and reliability estimates have been established. In part, this study was designed to assess these estimates.

To operationalize the three levels of the independent variable, three variations of the Watson-Barker Listening Test were constructed. Condition A included test instructions and original multiple choice items, with both item stems and responses presented orally. Condition B included test instructions and original multiple choice items, but with only the item stems presented orally. In this condition, Ss read the item responses from the original W-B test item response booklet. Finally, Condition C included test instructions and multiple choice items, with item stems and responses to each item presented orally, but also included written responses via the W-B test item response booklet from which Ss could read the responses as they were presented on the tape. In order to reduce systematic error across conditions, the original stimulus messages employed in the Watson-Barker Listening Test were retained, altering only the content of the test instructions. To reduce any additional error that might be introduced by a new narration of instructions, a graduate student trained in voice and diction was employed to narrate the new instructions and test item responses (the latter narration was altered to maintain consistency across all narrations). All three "new" versions of the Watson-Barker test were pre-recorded on

audiotape.

The dependent variable in this study was operationalized as total score on the Watson-Barker Listening Test. The maximum possible score on the test is a score of fifty (50).

Subjects

One hundred and eighty four subjects enrolled in an upper division business communications course at a major midwestern university participated in the study. Intact sections of the business communications classes were assigned to one of three conditions, via the arbitrary assignment of one morning class and one afternoon class to each condition. One section was omitted from the final analysis due to poor testing conditions (e.g., noise, room temperature, etc.). In total, 132 subjects' tests were scored for the final data analysis.

Procedures

Upon arriving to the experimental session, Ss were presented with test materials. For subjects in Condition A, these materials consisted only of IBM computer answer forms. Subjects in Conditions B and C received both IBM computer answer forms and test item booklets containing responses to the item stems to be presented on the tape. All Ss were asked by the experimenter to listen carefully to the audiotape, which had been announced a day prior to its presentation as a test of listening ability, and to place their answers to the test on the IBM answer form. Prior to administering the tests in Condition C, the experimenter reminded the Ss that the "answers" were to be presented on both the tape and in their answer booklets.

Results

In order to explore the overall significance of differences among modes of administering listening test responses, a 1×3 analysis of variance

was performed. An alpha of .05 was set as the overall level of significance. Results of the statistical analysis revealed no significant differences for modes of presentation ($F = 1.674$; $df = 131$; $p < .19$). Based on this finding no further comparisons between conditions were conducted. Table 1 provides a summary of the mean scores for each mode of administering listening test responses.

MODE OF PRESENTATION	\bar{X}
Condition A	32.00
Condition B	31.61
Condition C	30.39

TABLE 1 Mean Values by Mode of Administering Listening Test Responses

Discussion

In turning to possible explanations for the insignificant findings revealed in this study, at least three possible explanations emerge. First, in actuality, no differences, in fact, may exist between modes of presenting listening responses--we may simply do well or poorly on listening tests despite the method by which we are presented a set of item responses.

Some research in the area of recall and recognition tends to support this particular claim. Fitch Hauser (1984) and others, for example, have shown that the structure of a message, itself, may influence the recall of a given message (Bartlett, 1936; Bransford and Frank, 1972; Rumelhart and Ortony, 1977; Fitch Hauser, 1982). Smiley, et al. (1977) further verified this notion when they found that the structure of the stimulus message influenced their Ss' responses

on reading and listening tests. This finding was especially significant for subjects who were classified as "good readers." Poor readers, on the other hand, did not seem to be as aware of the message structure and scored equally poorly on both the reading and listening portions of the tests. These findings, in turn, point to a second possible facet of this explanation -- that is, that listening and reading are such similar processes that the mode by which information is presented does not influence the recall of that message. Mead (1978) and others have reported that a clearly established link exists between listening and reading abilities. To support this assertion, Mead cites Crook's (1957) report of a .70 correlation between listening and comprehension, and Haberland's (1959) finding of high positive correlations between listening and several verbal ability measures. This link between listening and reading skills has been further confirmed by findings of high correlations between reading and listening scores reported by Brown (1965), Ducker (1965), and Smiley, et al. (1977).

The significant correlations between the two skills, however, has led some to believe that research has simply failed to adequately determine whether or not listening test scores were a result of listening as a unique factor, or the result of a more general factor such as intelligence. To support this claim Kelly (1965) pointed to the high correlations (.83) between the Brown-Carlson and STEP tests with the OTIS test of mental ability. Possible explanations stemming from these results are that: (1) listening and reading are so closely related as to be inseparable skills, or (2) that listening research has not yet been able to theoretically or operationally distinguish listening and reading skills. Either of these two explanations could explain why no significant differences were found in this study among modes of presenting test item responses.

Although as the literature cited suggests, this initial explanation may be quite viable, an alternative suggests that no differences may have been found due to systematic error produced by our testing procedures. In order to assess the viability of this explanation, a manipulation check was conducted via a 1 x 6 analysis of variance across sections of students participating in the study. Results of this statistical procedure revealed an overall significant F-test ($F=5.126$; $df = 131$; $p < .0003$) across sections of subjects. Thus, systematic error, in fact, may have been introduced via the testing procedure.

To determine the specific differences that existed between sections of Ss employed in this study, a post hoc Tukey B multiple range test was employed. Results of this analysis revealed the following significant comparisons: Section 2 with Section 5 ($p < .05$), Section 3 with Section 5 ($p < .05$), Section 2 with Section 4 ($p < .05$), and Section 3 with Section 4 ($p < .05$). Table 2 provides the mean scores for each set of Ss by section:

Section	\bar{X}
1	30.6522
2	
3	33.4762
4	
5	33.8333
6	
3	30.0769
4	
5	28.8261
6	
5	32.0952
6	

TABLE 2: MEAN SCORES BY SECTION

A retrospective analysis of the actual testing situation uncovered several classroom variables which may have impacted on the test results. One factor could have been the furniture in the rooms. The section that produced the lowest mean score (Section 5: $\bar{X} = 28.8261$) met in a biology lab room. Although they had plenty of room and sat as close to the front of the room as possible, the students could not adjust their seats or desks to become more comfortable or to hear more easily. All other groups met in regular classrooms with standard desks, allowing them to lean forward or even to move the desks closer in order to hear better.

Another physical factor was the temperature of the rooms. All but one of the rooms used were excessively hot. The two sections with the highest mean scores (Section 3: $\bar{X} = 33.8333$ and Section 2: $\bar{X} = 33.4762$) were in a classroom with open windows and a relatively comfortable temperature.

The fact that a lot of extraneous noise existed outside of the classroom also may have affected the test scores. Only in one section was noise not a problem. Section 2, which had the second highest mean score ($\bar{X} = 33.4762$) had no problems with such noise. Section 3 with the highest mean ($\bar{X} = 33.8333$) had to cope with a little noise, but the noise was momentarily distracting. It did not mask the test items as did the noise in some of the other sections. The group with the lowest mean ($\bar{X} = 28.8261$) had the worst noise problem of the six groups. During the middle of the test for Section 5, especially during part II and much of part III, a turf vac was operating outside of the classroom. Therefore, not only did the students have to cope with excessive heat and lab tables, they also had to attempt to combat the outside noise.

Individual characteristics of the groups may have also had some effect on the test results. Students in Section 4, who were in the read only mode,

were observed reading the answers ahead of time even though they had been instructed to keep the answers covered until time to mark the correct answer. Since their scores were relatively low ($\bar{X} = 30.0769$) they may have been reading instead of listening to item responses. The students in section 2 (listen only mode) actually physically moved their desks closer to the message source in order to hear more clearly. This sections' mean score was 33.4762. Section 5 Ss (listen and read mode), on the other hand, seemed to display signs of impatience. A number of individuals were observed marking their answers before all of the answers had been given. This behavior could be interpreted in one of two ways. The students either read the answers at their own pace and chose their answers while ignoring the taped answers, or the students simply chose the first answer that sounded acceptable and chose to ignore the rest. In either case they seemed to ignore most of the tape presented item responses.

Given the significant statistical analysis by section and the potentially corresponding procedural problems which arose during the experimental sessions, we might attribute the lack of significance regarding presentational mode to explanation two. However, a third set of analysis conducted on the Watson-Barker instrument itself revealed a possible third interpretation of the results. It is to an explication of these analyses that the authors now turn.

As stated earlier, the primary purpose of this study was an exploration of possible differences regarding presentational modes. However, a tangential purpose that we had for this study was to present an initial assessment of test reliabilities for the Watson-Barker Listening Test. In turning to this specific task, several analyses were conducted. These included (1) the calculation of internal-consistency reliabilities on the Watson-Barker Listening Test overall; (2) the calculation of the same reliabilities on each separate

section of the measures; (3) the calculation of internal consistencies by mode of presentation; and (4) the construction of correlation matrices both by items and by sections of the test. Each of these will be discussed respectively as follows.

The first set of analyses that were conducted on the Watson-Barker Listening Test included the calculation of internal consistency reliabilities on (1) the test overall, (2) on the test by mode, and (3) on the test by instrument sections. These calculations were made via the Kuder Richardson 20, an index designed to calculate reliabilities when items are scored either 0 or 1 (1 = correct answer; 0 = incorrect answer). The formula for the Kuder Richardson 20 is as follows:

$$r_{tt} = \left(\frac{n}{n-1}\right) \left(\frac{S_t^2 - \sum pg}{S_t^2}\right)$$

where n = number of items in the test

p = proportion passing an item

q = $1-p$ (Guilford and Fruchter, 1978, p. 427)

Results of the analyses are summarized in Table 3.

Kruder Richardson 20 Tests	KR-20 Values	Strength of Coefficient
W-B test (all items)	0.587**	Moderate
MODE A	0.602**	Moderate
MODE B	0.636**	Moderate
MODE C	0.542**	Moderate
PART I (W-B)	0.292**	Weak
PART II (W-B)	0.272**	Weak
PART III (W-B)	0.457**	Moderate
PART IV (W-B)	0.011	Weak
PART V (W-B)	0.319**	Weak

**denotes significance at the .01 level

TABLE 3: INTERNAL CONSISTENCY RELIABILITIES OF THE WATSON-BARKER LISTENING TEST

Although, as the reader will note, all but one reliability coefficient reached the .01 level of significance, an interpretation of the strength (i.e., the "meaningfulness") of the reliability coefficients suggests that the internal consistency of the Watson-Barker Listening Test is, at best, only moderate.¹ Only five of the nine coefficients fell within the ranges of .40 to .80 (a moderate range of strength) with four of the coefficients accounting for less than 10% of the variance, respectively. For the overall test, only 35% of the variance was accounted for. These findings suggest the possible existence of problems regarding the intercorrelations among the test items. For this reason, the decision was made to construct correlation matrices both by items and by sections (5) of the listening test.

Because the Watson-Barker Listening Test consists of fifty (50) test items and their responses, time and space limitations do not allow for a complete presentation of the item intercorrelation matrices. However, Table Four presents summary ranges of intercorrelations for items 1 through 50 with items 1 - 50, and of intercorrelations among items by instrument part or section (Parts I - V).

Item Intercorrelations	Ranges of Coefficients
Items 1 - 50 with Items 1 - 50 (W-B test overall)	-0.2706 - 0.2763
Items 1 - 10 with Items 1 - 10 (Part I)	-0.2600 - 0.2235
Items 11 - 20 with Items 11 - 20 (Part II)	-0.1970 - 0.2311
Items 21 - 30 with Items 21 - 30 (Part III)	-0.1882 - 0.2242
Items 31 - 40 with Items 31 - 40 (Part IV)	-0.1676 - 0.1513
Items 41 - 50 with Items 41 - 50 (Part V)	-0.2592 - 0.1860

TABLE 4: ITEM BY ITEM INTERCORRELATION RANGES

As the above ranges in Table 4 seem to indicate, intercorrelations among the items in the Watson-Barker Listening Test, at its present stage of development, are exceptionally weak, although, as with the internal consistency reliabilities previously reported, many of the item intercorrelations reach the .01 level of statistical significance.² These findings suggest that the

instrument, at present, operationalizes not five independent and internally consistent dimensions of listening comprehension/ability, but instead represents a series of fifty items with very little inter-item consistency either within the parts of the test or within the instrument as a whole. Although as Table Five indicates, the five sections of the listening test seem to tap five independent dimensions of listening, a finer analysis of the intercorrelation matrices by items reflects a more realistic picture of the internal consistencies of the test parts and the test as a whole.

	Part I	Part II	Part III	Part IV	Part V
Part I	--	0.0930 p=0.115	0.1774 p=0.010	0.1075 p=0.081	0.0642 p=0.202
Part II		--	0.3464 p=0.001*	0.2456 p=0.001*	0.3228 p=0.001*
Part III			--	0.0331 p=.330	0.3438 p=0.001*
Part IV				--	0.1740 p=0.01*
Part V					--

* denotes a significant intercorrelation at the .01 level or greater.

TABLE 5: INTERCORRELATION MATRIX BY
W-B LISTENING TEST SECTION

These findings support our previous argument that the instrument, in its current form, may have contributed to the overall nonsignificance of the results.

Conclusions

As reflected in the previous discussion, no firm conclusions may be drawn regarding differences which may exist among modes of administering listening test responses. However, several viable research directions did emerge as a function of this analysis.

First, any further examination of modes of presenting test responses should attempt to minimize random or systematic error through the careful monitoring of testing conditions. Indeed, many listening tests are administered in precisely the conditions that we encountered, i.e., different environmental conditions that exist in all typical college classroom settings.

Second, should further attempts be made to examine modes of presenting test item responses, previously validated instruments with strong reliabilities should be used as independent or dependent measures of listening comprehension. Although the decision was made in this study to use an instrument without such prior testing, our decision, in part, included our desire to establish such indexes for the Watson-Barker Listening Test.

A third and, perhaps, more important research direction which emerged focused on the necessity of creating much more valid and reliable instruments tapping listening comprehension. Such instruments, however, require more adequate definition of listening as a theoretical construct as well as more distinctive delineation among listening and other verbal skills. This task is, perhaps, the most difficult task of all.

Finally, this study has produced some directions for further development of the Watson-Barker Listening Test. Hopefully, with such further development by the two test authors, studies such as ours may become much more viable research projects in the future.

ENDNOTES

1. The criteria of "strength" on which this analysis is based is as follows: 0 - .40 = weak internal consistency; .40 to .80 = moderate internal consistency; and .80 - 1.00 = strong internal consistency (Cummings, 1984).
2. For more specific information concerning the item intercorrelations, the reader may contact either of the authors.

BIBLIOGRAPHY

- Bartlett, F. C. Remembering, Cambridge, England: Cambridge University Press, 1932.
- Bransford, J. D. and Franks, J. J. The abstraction of linguistic ideas: A review. Cognition, 1972, 211-249.
- Brown, C. T. Three studies of the listening of children. Speech Monographs, 32, 1965, 129-138.
- Brown, K. L.; and Others. Assessment of Basic Speaking and Listening Skills: State of the Art and Recommendations for Instrument Development, Vol. I. Massachusetts State Department of Education, Bureau of Research and Assessment. Boston, September, 1979a.
- Brown, K. L.; and Others. Assessment of Basic Speaking and Listening Skills: State of the Art and Recommendations for Instrument Development, Vol. II. Massachusetts State Department of Education, Bureau of Research and Assessment. Boston, September, 1979b.
- Crook, F. E. Interrelationship among a group of language arts tests. Journal of Educational Research, 51, 1957, 305-311.
- Cummings, W. C. Personal Interview, March 5, 1984.
- Daly, B.; Neville, M. H.; and Pugh, A. K. Reading while listening: An annotated bibliography of materials and research. Unpublished Document. The University of Leeds Institute of Education, 1975.
- Ducker, S. Listening and Reading. Elementary School Journal, 65, 1965, 321-329.
- Faires, C. L. The development of listening tests. A paper presented to the Annual Meeting of the Mid-South Educational Research Association, New Orleans, November 2, 1980.
- Fitch Hauser, M. The effect of message structure on recall. Paper presented to the 3rd Annual Meeting of the International Listening Association, March, 1982.
- Fitch Hauser, M. The effect of message structure on inference making in recall. In Robert Bostrom (ed), Communication Yearbook 8. Beverly Hills: Sage Publications, in press.
- Guilford, J. P. and Fruchter, G. Fundamental Statistics in Psychology and Education (sixth ed.). New York: McGraw Hill, Inc., 1978.
- Haberland, J. A. A comparison of listening tests with standardized tests. Journal of Educational Research, 52, 1959, 299-302.
- Kelly, C. M. An investigation of the construct validity of two commercially published listening tests. Speech Monographs, 1965, 32, 139-143.

- Kintsch, W. and Kozminsky, E. Summarizing stories after reading and listening. Journal of Educational Psychology, 1977, 69, 491-499.
- Kintsch, W. and Van Dijk, T. A. Toward a model of text comprehension and production. Psychological Review, 85, 5 September, 1978, 363-394.
- Lundsteen, S. W. Listening: Its Impact on Reading and Other Language Arts. Urbana, Illinois: Clearinghouse on the Teaching of English, 1971.
- Mead, N. A. Issues related to assessing listening ability. Paper presented to the Annual Meeting of the American Educational Research Association in Toronto, Canada, March, 1978.
- Rogers, J. R. A formula for predicting the comprehension level of material to be presented orally. Journal of Educational Research, 56, 1962, 218-220.
- Rumelhart, D. E. and Ortony, A. The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, and W. E. Montague (Eds.), Schooling and the Acquisition of Knowledge. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977, 137-165.
- Smiley, S. S., et al. Recall of thematically relevant material by adolescent good and poor readers as a function of written versus oral presentation. Journal of Educational Psychology, 69, 1977, 381-387.
- Weisberg, R. A comparison of good and poor reader's ability to comprehend explicit and implicit information in short stories based on two modes of presentation. Research in Teaching English, 18, December, 1979, 337-351.