DOCUMENT RESUME

ED 360 662 CS 508 246

AUTHOR Willmington, S. Clay; Steinbrecher, Milda M.
TITLE Assessing Listening in the Basic Course: The University of Wisconsin-Oshkosh Listening Test.
PUB DATE Apr 93
NOTE 20p.; Paper presented at the Joint Meeting of the Southern States Communication Association and the Central States Communication Association (Lexington, KY, April 14-18, 1993). Legibility is poor.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Communication Research; Higher Education; Introductory Courses; *Listening Comprehension; *Listening Comprehension Tests; *Speech Communication; Test Bias; *Test Construction; Testing; Test Reliability; Test Validity
IDENTIFIERS Listening Research; *University of Wisconsin Oshkosh

ABSTRACT

A "Fundamentals of Speech Communication" course is required of all college students, and upon completion of such a course students should possess those basic speaking and listening skills necessary to complete successfully their college educations. With a view toward developing a new, more effective listening test, a study examined research in listening test development. The study also explained how the University of Wisconsin-Oshkosh Listening Test was developed and reported what was learned regarding the properties and performance of the test. The validity of the test was assessed using three procedures: the content procedure, the predictive procedure, and the known-groups method. Validity was also promoted by implementing suggestions and findings reported by listening assessment theorists. Two typical kinds of reliability tests were conducted in the test: test-retest and the Kuder-Richardson #20. Bias in regard to gender did not appear to exist in the test. No claims can be made concerning race bias since few minority students have taken the test. The test was subjected to item analysis to check for difficulty and discriminating power. The test can be administered successfully to up to 30 students in a single classroom by use of a one-half inch VHS video playback unit and monitor. Of 916 students tested, scores ranged from a low of 15 to a high of 52. The mean score was 34.77 with a standard deviation of 5.55. (A profile of the questions on the test and 24 references are attached.) (RS)

Assessing Listening in the Basic Course: The University of Wisconsin - Oshkosh Listening Test

By
S. Clay Willmington
Milda M. Steinbrecher
University of Wisconsin - Oshkosh

Submitted for presentation at the 1993 Joint Convention of the Central States Communication Association and the Southern Speech Communication Association, Lexington, KY

THE VALIDATION OF A LISTENING TEST FOR COLLEGE/UNIVERSITY STUDENTS

INTRODUCTION

The decade of the 1980s witnessed an unprecedented recognition of the vital role of listening in all aspects of a person's life. The failure to listen effectively has been advanced as the primary reason for all kinds of problems, ranging from relational problems to mistakes and inefficiencies in the workplace (Steil, Barker, & Watson, 1983; Wolff, Marsnik, Tacey, & Nichols, 1983; Wolvin & Coakley, 1992).

The 96-111 "Fundamentals of Speech Communication" course is required of all students, and it is charged with the responsibility of insuring that upon completion of the course students will possess those basic speaking and listening skills necessary to successfully complete their college education and to perform as effective communicators in their careers. A major mission of the university involves teacher preparation. Those of us in the Communication Department were particularly interested in a Department of Public Instruction rule adopted in 1987 which required "Demonstrated proficiency in speaking and listening as determined by the institution (preparing teachers)"

As a result of the conditions described above, the 96-111

instructional staff took the initiative to strengthen the listening dimension of the course. First, we identified the content to be taught and some learning activities and exercises designed to develop listening skills. Next, we searched for a way to assess the skills. After revieving the available instruments, we selected the video version of the Watson-Barker Listening Test as best meeting our immediate need. We discovered, however, that even with some obvious strengths of the test, we still were in need of a more appropriate and effective instrument to assess students in our basic course. So two of the staff, were encouraged to develop a listening test suitable for our use. Three years later, this test became a reality and was named the University of Wisconsin-Oshkosh Listening Test.

The purpose of this report is to review research in test development, listening test development in particular; explain how we used this research in the development of our test; and to report what we have learned regarding the properties and performance of the test.

VALIDITY

The first concern in the development of a test is its validity: does it indeed measure what it purports to measure? We have assessed the validity of the test by three procedures described by Smith: the content procedure, the predictive procedure, and the known-groups method (Smith, 1988).

Content validity, sometimes called face validity, asks if

the instrument measures a representative sample of the skills that comprise effective listening. This sample should be consistent with listening literature in general and the textbook for the course which was Communication Works, 3rd edition by Gamble and Gamble. A study of both sources shows that comprehension, defined as "...to understand the message in order to retain, recall, and - possibly - use the information at a later time" (Wolvin & Coakley, 1992) is the most basic purpose of listening. Early listening tests such as the Brown-Carlson and the STEP tests focused on comprehension. Thus, most of the questions should address comprehension, which is true of 38 of the 55 questions (69%) on the test.

Another purpose of listening is called critical or evaluative listening. The critical listener evaluates what is heard on the basis of sound logic or reasoning (Brownell, 1986). While not used as frequently as basic comprehension, critical listening is recognized by many experts as an important listening purpose in the wide-spread attention now being given to the development of critical thinking skills across all higher education curricula. Thirteen of the questions (24%) involve this kind of listening.

A final purpose of listening as explained in our textbook and taught in the course is empathic listening. Previous listening tests have not attempted to directly assess this kind, but it has been addressed in listening literature for some time (Wolff, Marsnik, Tacey, & Nichols, 1983), and has received

increased attention in recent years (Bruneau, 1989; Thomlison, 1990). It is sometimes treated as part of therapeutic listening (Wolvin & Coakley, 1992). Consequently, four of the fifty-five questions (7%) address empathic listening. (A list of the questions and what each one tests appears on the last page of this report.)

The second kind of validity study done was predictive validity, defined as comparing a behavior that is an important manifestation of the construct being measured with scores on an instrument designed to measure the same construct. For this purpose, we compared scores on our test with those on the 1991 Watson-Barker Listening Test (WBLT). The WBLT was developed in 1991 as a revision of the original Watson-Barker test of 1982. It is viewed as both a training tool and standard testing instrument (Watson, Barker, Roberts, & Johnson, 1991). Of the standardized listening tests now available (Rhodes, Watson, & Barker, 1990), our test appears to match most closely with the Watson-Barker, which is the only commercially available test that seeks to test listening skill of college students with a video format. Certain claims of validity have been reported for the original video version of the test which was produced in 1987 (Rubin & Roberts, 1987; Watson & Barker, 1988; Roberts, 1988).

Both tests were administered to selected sections of the 96-111 course. The Pearson product-moment correlation coefficient comparing the scores for a class of 23 was .65 ($p>.001$), for a class of 20 it was .61 ($p>.01$) and for a total of 62 students in

three classes taught by a single instructor it was .61 (p>.001). This method as an attempt to establish validity has been used for earlier tests (Applegate & Campbell, 1985; Rubin & Roberts,1987). Some experts in testing refer to this technique as supporting construct validity (Popham,1990). The assumption underlying this exercise is that if two tests correlate highly, as was found by these two tests, whatever validity is present in either is at least somewhat shared by the other.

The third kind of validity test used, the known groups method, compares the scores of two groups, one of which is known to possess higher levels, and one of which possesses lower levels, of the properties of the construct being tested. Validity is suggested if the group identified as possessing higher listening skills performs better on the test than the group possessing lower skills. The developers of the Kentucky Comprehensive Listening Test used this method in attempting to show validity of their instrument. They compared test scores for three groups: university students, high school students, and army colonels (Bostrom & Waldhart, 1988).

We used the known groups methods in two ways. First, we administered the test to two classes taught by the same instructor on the same day near the end of the semester. One was a regular class; the other was an honors class. It seems reasonable to assume that one of the unique characteristics of a group of honors students is that they are better than average listeners. It is unlikely that a poor listener could become an

honors student. The mean score for the regular class was 34.83, compared to a significantly higher mean of 39.82 for the honors class. ($t=3.10$, $p>.01$). The comparison of these two classes, therefore, support test validity.

The second known groups method was used in another way. One instructor taught a small group communication unit and administered a test based solely on classroom lectures, discussions, and activities to three sections of the course. There was no reading assignment in the unit. The test required that students remember, understand, and apply principles of small group communication in a real-life group of which they are a member. It is reasonable to believe that students possessing better listening skills would perform better on the test than students less skilled at listening. Out of the 3 classes, 35 students received either an A or a B on the test. Out of the same 3 classes, 22 students received either a D or an E. The mean score on the (Name of the test) for the 22 who received the D or E on the small group communication test was 36.05, while the mean score for the 35 who received As or Bs was significantly higher at 38.91 ($t=2.4$, $p>.05$). Thus, both applications of the known groups method, first with the honors students and secondly with the high and low small group communication test performers, supported the validity of the listening test.

Validity was also promoted by implementing suggestions and findings reported by listening assessment theorists. One example of this is the claim that a listening assessment should not

honors student. The mean score for the regular class was 34.83, compared to a significantly higher mean of 39.82 for the honors class. ($t=3.10$, $p>.01$). The comparison of these two classes, therefore, support test validity.

The second known groups method was used in another way. One instructor taught a small group communication unit and administered a test based solely on classroom lectures, discussions, and activities to three sections of the course. There was no reading assignment in the unit. The test required that students remember, understand, and apply principles of small group communication in a real-life group of which they are a member. It is reasonable to believe that students possessing better listening skills would perform better on the test than students less skilled at listening. Out of the 3 classes, 35 students received either an A or a B on the test. Out of the same 3 classes, 22 students received either a D or an E. The mean score on the (Name of the test) for the 22 who received the D or E on the small group communication test was 36.05, while the mean score for the 35 who received As or Bs was significantly higher at 38.91 ($t=2.4$, $p>.05$). Thus, both applications of the known groups method, first with the honors students and secondly with the high and low small group communication test performers, supported the validity of the listening test.

Validity was also promoted by implementing suggestions and findings reported by listening assessment theorists. One example of this is the claim that a listening assessment should not

depend on reading and writing skills (Caffrey, 1955; Backlund, Brown, Gurry, & Jandt, 1982). Many previous tests in communication have required the student to read and/or write to the point that reading and writing skills levels may have contaminated the purported purpose of the test. In this test, reading skill is not vital to success because everything printed on the video screen is also presented orally. The only written response necessary is making marks on a computer-scored answer sheet.

We noted the advice that the methods of presentation should be controlled (Caffrey, 1949), which is best accomplished by videotape (Backlund, Brown, Gurry, & Jandt, 1982). Research on methods of presentation found that students score significantly higher on the Brown-Carlson and STEP listening tests when administered by an "effective" speaker than when administered by an "ineffective" speaker (Barker, Watson, & Kibler, 1984). Consequently, the (Name of the test) has been placed completely on videotape, thus controlling for methods of presentation.

Most previous listening tests have used audio-only stimuli. But as Roberts points out, "listeners generally do not 'listen' with just their ears. Listening typically takes place while the listener is hearing and viewing the sender of the message." He suggests that for a listening test to be useful in terms of applying results to everyday encounters, the respondent must be able to respond to a speaker's entire communication code, both verbal and nonverbal (Roberts, 1988). Consequently, in this

test all messages are presented by people who are seen as well as heard in various settings.

Validity can also be influenced by the content of the stimulus material. One suggestion is that the material should be interesting and meaningful to those taking the test (Backlund,Brown, Gurry, & Jandt, 1982). A further concern of the developers was that the test material be of somewhat equal interest, meaningfulness, and familiarity to test-takers to reduce the chances of any of these elements giving an advantage or disadvantage to certain persons. While recognizing that no single piece of material can totally meet these criteria, we tried to minimize differences by using material that should at least be somewhat interesting, meaningful, and familiar to college/university students. Some of the scenarios present situations that are oriented uniquely to higher education and life as a student.

Another suggestion is that a listening test should differentiate among three kinds of listening as defined partially by the time between the stimulus and the response. One kind is called short-term listening which calls for a response within 15 seconds. Another is short-term with rehearsal, which calls for a response within 40 seconds. The third is called lecture listening, where the response comes at least one minute after the presentation (Bostrom & Waldhart, 1988). The (Name of the test) includes all three of these types of listening. The most short-term questions are those that ask for a response to one or two

sentence statements that respondents are supposed to identify as either acceptable or unacceptable examples of evidence or as either sound or unsound examples of reasoning. The longest stimulus is a four-minute speech about which 11 questions are asked.

Finally, if the listening instruction is part of a broader communication course, it seems reasonable that some of the questions can assume a knowledge of basic communication principles. The test includes 17 questions that require some of this kind of knowledge to respond correctly. But the inclusion of these questions raises a validity question: Is the validity of the test restricted to students who have received instruction in basic communication principles? To find the answer to this question, we administered the test to several sections of the course prior to basic communication instruction and to several sections after the instruction. The fact that there was no significant difference between those tested before and those tested after communication instruction suggests that the knowledge of communication principles necessary for performing well on the test is so basic that this knowledge does not influence test performance. The validity of the test, therefore, is independent of knowledge of communication principles as taught in the course.

RELIABILITY

The next major concern in developing an assessment instrument is its reliability. Reliability, or consistency,

refers to the extent to which individual items on the test function in the same way.

Two typical kinds of reliability tests were conducted on the test: test-retest and the Kuder-Richardson #20 (K-R20). A total of 49 students took the test two times within a 10-day period of time. A Pearson product-moment correlation coefficient of .68 ($t=4.37$, $p>.001$) was calculated comparing the two sets of scores.

The K-R20 test gives an overall reliability score indicating the average correlation obtained from all possible split-half reliabilities (Kuder-Richardson, 1937). For a group of 916 students taking the test in a single semester, the K-R20 reliability score was .67 ($t=20.24$, $p>.001$).

Opinions vary in regard to what the reliability of a test should be to be deemed satisfactory or exemplary. The only item of agreement is "the higher the better." One expert argues that a test should have a reliability coefficient of at least .65 to be considered satisfactory (Cangelosi, 1982). This test meets the .65 minimum, but does not exceed it by much. Two factors operate to limit the reliability coefficient of the test compared to standardized tests boasting of higher figures. One factor is the relative shortness of this 55-question test; the more questions included on a test, the greater the potential for high reliability. Another factor is the relatively homogeneous population used for the reliability studies. Again, the potential for higher reliability would be increased by administering to a more diverse population than found in the

group of students on a single university campus who have taken the test.

BIAS

A third subject to be addressed in evaluating a test is possible sources of bias. Bias occurs when questions are more easily or less easily answered because of experience which is unique to a particular group. Gender and race are often cited as possible sources of bias.

In regard to gender, bias does not appear to exist to any significant extent in this test. Women students on our campus average approximately one more correct answer than men of the 55 questions asked. While this is not a statistically significant difference, if the test is used so that a single point difference determines a student's grade, or whether a student is admitted into a professional program of study, one more right or wrong answer can make a profound difference. In this case, a closer look is warranted to make sure that the additional correct answer given by a woman is a reflection of her listening ability and not the result of gender bias in the test.

No claims can be made at present concerning race bias. Although 3 minority students appear as talent in the test, the overwhelming majority of the students who have taken the test are Anglo-Saxon, young people, born in this state, between the ages of 18 and 21. The few minority students who have taken the test constitute an insufficient number for any analysis of race bias. Any institution or group using the test with members of a

minority might conduct their own analysis to determine possible race bias.

ITEM ANALYSIS

The test has been subjected to item analysis to check for difficulty and discriminating power. In developing this test we made sure that the questions were difficult enough that scores did not cluster at the high end of the scale, but not so difficult that they clustered at the low end of the scale.

The mean item difficulty score for 916 students taking the test was 63.29. This is a satisfactory score because while it keeps scores from clustering at the top end, which would reduce discrimination power, it is not so difficult that students become demoralized at its difficulty. Out of the total of 208 possible responses to the 55 questions, practically all of the options receive at least some "bites" when administered to a class of 20-28 students.

The mean item discrimination score for 916 students mentioned above was 22.97. With a score above 20, we are satisfied with the ability of the test items to discriminate between highly skilled and less skilled listeners. It should be noted that the reliability and discrimination scores are somewhat related, moving up or down together.

ADMINISTRATION

The test can be administred successfully to up to 30 students in a single classroom by use of a one-half inch VHS video playback unit and monitor. Care should be taken to see

that all students are positioned so they can see the picture clearly on the monitor and hear the audio portion of the tape. Computer-scored answer sheets providing for at least 4 options to 55 questions can be marked with a pencil and machine scored.

The test takes 45 minutes to administer. Test administrators can choose to run straight through the test without a break. But unlike pencil and paper tests where students can look ahead at questions, because this test is on videotape, breaks can be taken at any time, or the test can be administered in parts in as many blocks of time as desired. In fact, because of the sustained concentration necessary in taking the test, even a pause for a few seconds some time during the test is recommended.

NORMS

Of 916 students tested, the scores ranged from a low of 15 to a high of 52. The mean score was 34.77 with a standard deviation of 5.55. The mode was 36.

The percentile ranks corresponding to raw scores are shown below:

Percentile	Raw Scores
90	42
80	39
70	38
60	36
50	35
40	34
30	32
20	30
10	28

These figures show that a score of 42, for example, is

higher than 90% of the total scores. Likewise, a score of 30 is higher than 20% of the total scores.

CONCLUSION

To summarize this report, first, we explained the need we experienced for a standardized listening test. Then we identified some of the literature in test construction, especially relating to listening assessment. Next, we explained how we applied information in the literature to the development of the test. Finally, we reported data on the nature and effectiveness of the test as an instrument to assess the listening skills of students in our basic speech communication course.

UW-Oshkosh Listening Test - Profile of Questions

	Question Number	Listening to Comprehend	Listening Critically/Evaluatively	Listening Empathically	Recognizing Comm. Principles
	1		X		X
	2	X			X
	3	X			X
	4	X			X
Four Minute Speech	5	X			X
	6	X			
	7	X			
	8	X			
	9	X			
	10	X			
	11	X			
	12	X			
Student Council Announcement	13	X			
	14	X			
	15	X			
	16	X			
Open Meeting Announcement	17	X			
	18	X			
	19	X			
Air Fare Announcement	20	X			
	21	X			
	22	X			
Directions for Chest X-Ray	23	X			
	24	X			
	25	X			
	26	X			
Description of a State Park	27	X			
	28	X			
	29	X			
	30	X			
	31				
Use of Room for Weekend	32		X	X	X
	33	X			X
	34		X		
	35	X			
Bad Relationship With Third Party	36	X			
	37		X		
	38	X			
	39				
Problem Taking Test	40		X	X	X
	41				X
	42				
Problem With Grade on Paper	43		X	X	X
	44	X			
	45		X		X
	46				
Job Interviews	47	X		X	X
	48	X			
	49	X			
	50				
Assessing Use of Evidence	51		X		X
	52		X		X
	53		X		X
Assessing Use of Reasoning	54		X		X
	55		X		X

REFERENCES

Applegate, J. & Campbell, J. (1985). A correlation analysis of overall and sub-test scores between the Watson-Barker and the Kentucky Comprehensive listening tests. Paper presented at the Annual Meeting of the International Listening Association, Orlando, FL.

Backlund, P., Gurry, J., Brown, K, & Jandt, F. (1982). Recommendations for assessing speaking and listening skills. Communication Education, 31(1), 9-18.

Bostrom, R.N. & Waldhart, E.S. (1988). Memory models and the measurement of listening. Communication Education,37, 1-13.

Barker, L.L., Watson, K.W., & Kibler, R.J. (1984). An investigation of the effect of presentation by effective and ineffective speakers on listening test scores. Southern Speech Communication Journal,48, 309-318.

Brownell, J. (1988). Building Active Listening Skills (pp. 166-169). Englewood Cliffs, N.J.: Prentice-Hall.

Bruneau, T. (1989). Empathy and listening: A conceptual review and theoretical directions. Journal of the International Listening Association, 3, 1-20.

Caffrey, J. (1949, Nov. 12). The establishment of auding-age norms. School and Society, 70, 310-12.

Cangelosi, J.S. (1982). Measurement and Evaluation (p.292). Dubuque, IA: Wm C. Brown.

Gamble, T.K. & Gamble, M. (1990). Communication Works, 3rd ed. New York: McGraw-Hill.

Glenn, E.C. (1988). Gender differences in listening tests. Paper presented at the Annual Meeting of the International Listening Association, Scottsdale, AZ.

Kuder, G.F. & Richardson, M.W. (Sept. 1937). The theory of the estimation of test reliability. Psychometrika, 151-60.

Popham, J. (1990). Modern Educational Measurement (p. 109). Englewood Cliffs, N.J.: Prentice-Hall.

Rhodes, S.C.,Watson, K.W., & Barker, L.L. (1990). Listening assessment: Trends and influencing factors in the 1980s. Journal of the International Listening Association, 4, 62-82.

Roberts, C.V. (1988). The validation of listening tests: Cutting the Gordian Knot. Journal of the International Listening Association, 2, 1-19.

Rubin, R. & Roberts, C. (1987). A comparative examination and analysis of three listening tests. Communication Education, 36, 142-153.

Smith, M.J. (1988). Contemporary Research Methods (pp. 49-50). Belmont, CA: Wadsworth.

Steil, L.K., Barker, L.L. & Watson, K.W. (1983). Effective Listening (pp.36-43). Reading, MA: Addison-Wesley.

Thomlison, T.D. (1990). Teaching empathic listening within the speech curriculum. Paper presented at the Annual Meeting of the Central States Communication Association, Detroit.

Watson, K.W., Barker, L.L., Roberts, C.V., & Johnson, P.M. (1991). Development and administration of the Watson-Barker Listening Test (p.1). New Orleans: Spectra, Inc.

Watson, K.W. & Barker, L.L. (1988). Listening assessment: The Watson-Barker Listening Test. Journal of the International Listening Association, 2, 20-32.

Watson. K.W. & Rhodes, S.C. (1988). A preliminary study of the effects of gender and a video tape instructional strategy on listening effectiveness as measured by the Watson-Barker Listening Test. Paper presented at the Annual Meeting of the International Listening Association.

Wolff, F.I., Marsnik, N.C., Tacey, W.S., & Nichols, R.G. (1983). Perceptive Listening (pp. 39-40; pp. 170-172). New York: Holt, Rinehart and Winston.

Wolvin, A.D. & Coakley, C.G. (1992). Listening, 4th ed (pp.27-31; pp. 295-320). Dubuque, Ia: Wm C. Brown.

Wisconsin Department of Public Instruction (1987). State of Wisconsin teacher education program approval rules and certification standards, PI 4.06(6)(a)2. Madison, WI: Wisconsin Department of Public Instruction.