*Annual Review of Applied Linguistics* (1998) **18**, 171–191. Printed in the USA.


# ASSESSING LISTENING ABILITIES

**Geoff Brindley**

## INTRODUCTION

Over the last two decades, research has highlighted the important role that listening plays in language acquisition (Brown and Yule 1983, Ellis, *et al.* 1994, Faerch and Kasper 1986, Feyten 1991, Long 1985), and listening comprehension skills have begun to receive a lot more systematic attention in language teaching classrooms. A wide range of books, articles, and materials aimed at assisting teachers to develop learners' listening skills are now available, and a variety of comprehension-based methodologies have been proposed (see, for example, Anderson and Lynch 1988, Courchene, *et al.* 1992, Rost 1990; 1994, Underwood 1989). However, although many of the tasks used for teaching listening are virtually identical to those which appear in tests, assessment of listening ability has received relatively limited coverage in the language testing literature. As a partial contribution towards the correction of this imbalance, this chapter has three aims: 1) to survey some of the issues and challenges involved in assessing second language listening ability, 2) to discuss and evaluate some of the assessment methods and techniques used, and 3) to consider potential applications of new technology. Avenues for further research will also be suggested.

## VIEWS OF SECOND LANGUAGE LISTENING

### 1. Current models of listening

The relatively low profile of listening assessment may reflect the inherent difficulties involved in describing and assessing an invisible cognitive operation. In this regard, a number of overviews of listening comprehension have identified the lack of empirically sound models of listening comprehension which could be used to guide testing (Brindley and Nunan 1992, Brown and Yule 1983, Buck 1990, Dunkel 1991a, Dunkel, *et al.* 1993, Rost 1990).

2. Skills hierarchies

Despite the perceived inadequacies of current models of listening comprehension, a number of common points of consensus on the nature of listening processes emerge from the language testing literature. First, an assumption is frequently made by test developers that there are identifiable listening skills which language learners deploy in order to comprehend aural texts and that these can be arranged in a hierarchy from 'lower order' (involving understanding of utterances at the literal level) to 'higher order' (involving inferencing and critical evaluation) (Buck 1990; 1991, Rost 1990, Weir 1993). Numerous taxonomies of both general and specific skills involved in listening have been proposed (see, for example, Boyle 1984, Munby 1978, Powers 1986, Richards 1983, Rost 1990, Weir 1993), and some of these have been used as a basis for identifying the language operations to be sampled in listening tests such as understanding main ideas, listening for specific information, inferring the speaker's meaning, etc. These skills are sometimes presented in test specifications according to their hypothesized levels of difficulty. A postulated developmental hierarchy of skills also forms the basis of the levels of listening ability which appear in proficiency rating scales such as the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines (ACTFL 1986).

3. Listening as an interactive process

A second common theme which emerges from the testing literature is the move away from the notion of listening as auditory discrimination and decoding of decontextualized utterances towards a much more complex and interactive model which reflects the ability to understand authentic discourse in context (Buck 1990, Douglas 1988, Hendrickson 1992, Thompson 1995, Weir 1993). This model, which has had considerable influence on communicative testing practice, is described thus by Buck:

> ...any model of the normal process of listening comprehension must allow the sum total of the listener's knowledge, past experience, current thoughts, feelings, intentions, personality and intelligence to interact freely with the acoustic input and with each other, to create the interpretation of a text. Processing has to be massively interactive and parallel (1990:409).

## ISSUES AND CHALLENGES IN THE ASSESSMENT OF LISTENING

1. Assessing higher level listening skills

Testing an expanded construct of listening which attributes an active role to the listener poses a number of serious problems for developers of listening tests. If processing is interactive and parallel, listeners may use higher level and lower

level processing simultaneously to interpret a text. This makes it very difficult either to distinguish between different levels of processing or to attribute test responses to any one skill (Brindley forthcoming a, Buck 1990; 1991). It also casts doubt on descriptions of listening ability which are couched in terms of skill progression such as proficiency rating scales (Brindley forthcoming b, Douglas 1988). At the same time, any attempt to assess the ability to relate utterances to context raises the question of how feasible it is to test text interpretation, since interpretations may differ between individuals according to a multiplicity of cognitive and affective factors (Buck 1990; 1991, Nissan, *et al.* 1996).

Given the complexities involved in assessing higher order skills, it could be argued that the emphasis in listening assessment should be on direct meaning comprehension, rather than on the ability to draw inferences (Buck 1991:86). The case for such a course of action is strengthened by the findings of some studies which have shown items that test propositional information to be better discriminators of listening ability (de Jong and Glas 1987, Henning 1991).

Despite the numerous difficulties associated with constructing items aimed at tapping non-literal comprehension, many language testers recommend that such items should be included in listening tests (Boyle and Suen 1994, Buck 1990, Burger and Doherty 1992, Thompson 1995, Weir 1993), presumably on the basis that to omit them would lead to an unacceptable narrowing of the construct being assessed. Also, as Buck (1990:417) demonstrates, "the difference between language processing and inferencing is not always so clear," with higher level processing often being used to make sense of explicitly stated information. However, a number of writers emphasize that considerable care needs to be taken in constructing inferencing items. Since text interpretation is subject to considerable individual variation, items need to be designed in such a way as to constrain the range of possible responses (Buck 1990, Nissan, *et al.* 1996, Weir 1993). In this regard, Nissan, *et al.* (1996:29), noting the particular difficulties in testing interpretive inference via the multiple-choice format, call for further research into the relationship between test taker inferences and those intended by the item writers.

2. Confounding of skills

The act of assessment complicates the comprehension process considerably, as Spolsky demonstrates:

> To a complex enough model of speaker/writer, text and understander, we add several more critical parts: a tester who becomes speaker/writer and creator of a second text that sets tasks for the understander, a third text produced by that understander, a reader/interpreter of that new text (the tester or marker), and a fourth text, a mark or score or grade that awaits the interpretation of an additional participant, the test user (1994:147).

The validity of listening tests is further threatened by the fact that the tasks that are set and the texts that are produced will often require the use of language skills other than listening. Candidates may have to read written stimuli (sometimes at the same time as they are listening to an aural text) and provide oral or written responses to test questions. In such cases, it is easy to see how what is intended to be a listening test can easily end up assessing another ability, thus introducing what Messick (1989:34) refers to as "construct-irrelevant variance." This requirement to elicit some kind of "product" means that it is extremely difficult—some would say impossible—to construct a "pure" test of listening uncontaminated by some other skill (Boyle and Suen 1994, Buck 1990). The situation is further complicated by the potential confounding effects of a range of individual factors including memory capacity (Call 1985) and topic familiarity (Long 1990, Schmidt-Rinehart 1994).

### 3. Assessing listening in oral interaction

Many standardized listening tests tend to focus on non-participative listening tasks which require candidates to listen to pre-recorded texts and respond through such activities as ticking boxes, circling alternatives, or writing short answers. A good deal of listening, however, happens in the context of oral interaction where listening and speaking ability are closely interconnected—a person cannot carry on a conversation effectively without understanding what the interlocutor is saying. But this type of listening is rarely sampled in language tests (Schrafnagl and Cameron 1988:88).

In order to address this shortcoming, some testers have argued that listening tests should attempt to incorporate opportunities for the negotiation of meaning. Ross and Langille (forthcoming), for example, suggest that the inclusion of passages containing negotiated discourse would improve construct and content validity by making the test content more like naturally occurring language use. Berne (1995:326) recommends that tests should allow repetition of a passage for lower level learners if they are to reflect authentic listening behavior. Brindley and Ross (1997) suggest that there may be a case for conducting separate assessments of listening in interactive methods such as oral interviews.

### 4. Dealing with authenticity

A persistent theme in language testing over the past two decades has been the importance of trying to construct test tasks and items which replicate authentic language in a given domain of language use (Bachman 1990). However, it is not easy to design listening comprehension tests which mirror the purposes of real-life listening. This difficulty is partly because a great many of the listening tasks people undertake in everyday life (e.g., listening to radio or television programs) do not require a specific response—the listener simply processes the information and stores it until it is needed. Nor would people usually find themselves in the

situation of having to carry out the sorts of tasks which are commonly used in language tests, such as deciding whether what they have just heard corresponds to one of four written or pictorial alternatives.

The use of authentic samples of naturally-occurring speech for listening assessment can also be problematic. Authentic texts are frequently unsuitable for use in tests because of factors such as poor sound quality, lack of adequate contextualization, or overdemanding processing load, and may therefore need to be edited and re-recorded (Brindley forthcoming a, Weir 1993).

For all of these reasons, listening test designers usually have to make some compromise on authenticity. In terms of text selection, while unedited samples of natural language may be difficult to use, experience indicates that it is possible to devise test tasks which engage the same kinds of processes candidates would use in real life and reflect the kinds of aural texts they would encounter (UCLES/RSA 1990). Examples of the types of "authentic" text types used in communicative listening tests would include conversations, announcements, service encounters, answering machine messages, directions, lectures, narratives, anecdotes, personal reports, news broadcasts, interviews, advertisements, announcements, debates, and talkback exchanges (Dunkel, *et al.* 1993, Hughes 1989, UCLES/RSA 1990, Weir 1993).

## CONSTRUCTING TASKS FOR THE ASSESSMENT OF LISTENING

### 1. Factors affecting test performance

A wide range of variables which may affect listening text and task difficulty have been identified by researchers (Brindley and Nunan 1992, Buck 1990, de Jong 1987, Dunkel, *et al.* 1993, Henning 1991, Nissan, *et al.* 1996, Rost 1990, Rubin 1994, Thompson 1995). All of these variables need to be considered by test writers when preparing test specifications and grading listening passages. In this regard, Buck (1990:96) suggests that "care should be taken by testers to ensure that the variables which influence the final score are those the test maker intended." However, this is easier said than done, given the multiplicity of factors involved and the complexity of the interactions between them. Among the key variables are the nature of the input (speech rate, length, background, syntax, vocabulary, noise, accent, register, propositional density, amount of redundancy, etc.), the nature of the assessment task (amount of context provided, clarity of instructions, availability of question preview, whether the task calls for recognition only or synthesis, etc.), and individual listener factors (memory, interest, background knowledge, motivation, etc.). However, only a small number of studies have been conducted into the role played by these variables, either singly or in combination, in listening test performance (see Berne 1993, Buck 1990, Henning 1991, Nissan, *et al.* 1996, Sherman 1997, Shohamy and Inbar 1991), and a good

deal of further research remains to be done before their specific effects can be gauged.

## 2. Practical issues in listening test construction

Undertaking a listening test in a second language requires intense concentration on the part of candidates and can be very stressful, particularly if only one hearing of a text is allowed and the recorded input cannot be stopped, as is the case with many large-scale high-stakes tests (Brindley, *et al.* forthcoming, Hughes 1989). For this reason, it is important to minimize the possible effects of extraneous factors such as test presentation and administration on candidates' performance. Ways in which this can be done are outlined in a number of test construction handbooks and practical articles (see, for example, Carroll and Hall 1985, Cohen 1994, Hughes 1989, Shohamy 1985, Thompson 1995, Weir 1993) and will therefore simply be presented in summary here.

Test instructions. It is important that candidates are provided with clear, simple, and explicit instructions on how to do the test. A number of testers suggest that instructions can be given in the first language of the learners in order to minimize the possibility of confusion (Hughes 1989, Thompson 1995, Weir 1993), although this is clearly not a feasible option in cases where the test candidates do not share a common language. Nevertheless, it is important to make sure that the language of instructions is not more complex than the language used in the test passages. Adequate time should be allowed for candidates to familiarize themselves with the item formats and task requirements, and clear examples of each new item type should be given. It is conventional to allow item preview, which has been found both to motivate testees (Buck 1990, Sherman 1997) and to facilitate comprehension (Berne 1995), although Buck (1990) found that it did not make a great deal of difference to item difficulty, and Sherman (1997:185) suggests that previewed questions "seem more helpful than they really are."

Item development. Three important points need to be stressed with respect to item development. First, many testing manuals emphasize that items should be based on the recorded text itself, not on a transcript, and should follow the sequence of presentation of information in the text (Thompson 1995, Weir 1993). Second, the questions should be based only on the key information which listeners could be expected to extract from the text, not on trivial details (Shohamy and Inbar 1991, Weir 1993). Third, sufficient space needs to be provided between test questions to enable candidates to finish answering a question before they hear the information containing the answer to the next one—otherwise, there is a chance that candidates will become disoriented and "lose their place," thus missing a succession of questions that they might be quite capable of understanding in non-test situations (Brindley, *et al.* forthcoming, Hughes 1989).

Conditions of administration. Since many listening tests are administered to large groups via a tape-recorder, the acoustics of the testing room are of particular importance, as is the sound quality of the recordings used (Alderson, *et al.* 1995, Heaton 1988, Weir 1993). It is important to ensure that all technical equipment is carefully tested before use.

3. Item formats

The relative merits of various item formats for assessing listening are discussed in detail by a number of writers, including Boyle and Suen (1994), Heaton (1988), Hughes (1989), Rost (1990), Shohamy (1985), Thompson (1995) and Weir (1993) and will not be reiterated here. However, certain issues and debates surrounding some of the more commonly used formats are canvassed below.

Short answer questions. Production-based response formats such as the short answer question allow the tester to determine fairly clearly whether the testee has understood the spoken text. Hughes (1989:137) suggests that short answer questions can work quite well in listening tests provided that the answers are kept very short and thus do not depend too heavily on candidates' writing skills. However, open-ended short answer questions usually have to be clerically marked and are thus more resource-intensive than other objectively scored item types such as multiple-choice or picture identification, which can be mechanically scored. Short answer questions also require a detailed scoring key containing a list of acceptable responses. Although the list can be developed on the basis of an analysis of the range of responses given in piloting, it is likely that further plausible answers will emerge during the first administrations of the test. In high-stakes testing situations, markers can be asked to identify these and to bring them to the attention of a supervisor or chief examiner who can adjudicate on their acceptability, if necessary in consultation with other members of the group responsible for test development. In this way, a comprehensive set of marking guidelines can be built up. The fact that a test item may produce such a wide range of responses strengthens the case for using procedures for item development such as those suggested by Weir (1993:111), who recommends that items should be based only on those main points and details identified by native and non-native listeners through mind-mapping and note-taking.

True-false. Opinion is divided as to whether true-false items are an appropriate way to assess listening comprehension. From a validity perspective, it could be argued that evaluating whether a proposition is true or false is a legitimate and quite common purpose for listening. True-false items are used in a number of well-known language tests, such as the Cambridge Certificate of Proficiency in English. The obvious disadvantage of this format is that candidates have a fifty percent chance of getting the right answer, but some test manuals suggest that this problem can be addressed by including a third option of "no information

available" or "can't tell from text" (Carroll and Hall 1985). Another way of reducing the possibility of the candidate getting the right answer by chance is to increase the number of such items. However, Burger and Doherty (1992:315) report that the true-false-not given format did not work well in an English as a Second Language listening test they developed at the University of Ottawa "because of the fleeting nature of the spoken word and the natural and desired fact that listeners focus on what is said and not on what is not said."

Multiple-choice. Many major international tests of listening comprehension (such as the TOEFL) use multiple choice questions which offer ease of scoring and high internal consistency reliability (Henning 1991). Some language testers, however, advise against the use of the multiple-choice format for the testing of either listening or reading on the grounds, *inter alia*, that answering multiple-choice questions does not resemble normal language use, that such items are too open to guessing, and that these items are too difficult to develop (e.g., Hughes 1989, Weir 1993). Hansen and Jensen (1994:250-251) point out that multiple-choice questions make considerable processing demands on testees—not only do they have to pay attention to the aural input but they also have to read and retain four propositions in working memory before matching them with the aural text. At the same time, they have to be listening for the next item. There is also some evidence to indicate that there may be a method effect with multiple choice questions in listening tests. Berne (1993) found that subjects performed better on multiple-choice questions than on either an open-ended or cloze task, suggesting that items requiring only recognition are easier than those requiring retrieval and production of a correct answer. In a study of factors contributing to the difficulty of TOEFL listening comprehension items, Nissan, *et al.* (1996) found that the multiple choice format, when used to test inferencing, may engage candidates in an extra step of re-adjusting their inference if it does not correspond to one of the response options.

On the other hand, while recognizing the potential harmful effects of this item format, some testers have argued that it is possible to devise multiple-choice items which test the meaningful use of language in context (Boyle and Suen 1994, Rea 1985). In this context, it is worth noting that the developers of a number of recent communicative tests still see a place for such items. The revised Cambridge First Certificate in English (UCLES 1995), for example, includes some multiple choice questions as part of the listening paper, although a range of other item types are also used.

Summary cloze. This item format requires candidates to fill in gaps in a summarized version of the text they have heard as they listen. On the basis of the results of a listening summary cloze test used to test the English language proficiency of Chinese tertiary students, Lewcowicz (1991) concludes that it has a number of advantages: It allows a good deal of flexibility in the texts and topics chosen; a large number of items can be developed; and marking is objective but

not restricted to exact words and phrases. However, she notes that the summary cloze is not easy to set and requires careful pretesting and moderation of the marking scheme. It should also be pointed out that this format is potentially very cognitively demanding since it requires candidates to read, listen, and write simultaneously.

Dictation. A number of writers suggest that dictation can be used as a measure of general listening ability (Cohen 1994, Coniam 1996, Hughes 1989). Errors made by testees provide some insight into the processing strategies they use and quite high correlations have been found between dictation and other more direct measures of listening (Bacheller 1980). However, since dictation clearly involves skills other than listening, including auditory memory, spelling, and grammatical and lexical knowledge, it would be inadvisable to use it as a surrogate for a listening test. Also, as Weir points out, "the conditions under which this task is conducted only in a very limited sense reflect the normal conditions for the spoken language" (1993:124).

The need for variety. Although some high profile language tests favor the use of a single item format such as multiple-choice for the assessment of listening, there is a growing body of evidence which suggests that candidates' test performance may vary according to the type of response required (Burger and Doherty 1992, Nissan, *et al.* 1996, Shohamy and Inbar 1991). In the light of these findings, it would seem prudent for test developers to try to include a variety of item formats in a listening test which tap a range of listening purposes (Berne 1993).

## THE PROMISE OF NEW TECHNOLOGY

### 1. Computer-adaptive assessment

Recent technological developments such as multi-media, interactive video, and computer-generated speech have opened up a wide range of possibilities for the development of new forms of language assessment (Alderson 1988, Corbel 1993). In recent years, language testers have begun to explore the feasibility of creating computer-adaptive tests with interactive capabilities which would include the contextual elements now lacking in audio-based listening tests (Burstein, *et al.* 1996, Dunkel 1991a; 1991b; 1992). Dunkel (1991b; 1992) describes a prototype development effort to design a computer-adaptive test of listening based on the ACTFL Guidelines. The testee is presented with speech samples via a headset and responds to different types of questions which appear on the computer screen in multiple choice form. Text and graphics are used to contextualize the aural input. The researcher concludes that this type of computer-adaptive test is a viable alternative to the traditional audio-mediated test using pencil and paper, and she foresees the development of tests based on the speech-digitizing and advanced graphics capabilities which are available with the computer systems of the 1990s.

However, she notes that a major interdisciplinary effort will be needed to achieve this goal.

Coniam (1996) describes a computer-based dictation used to test English listening skills of Hong Kong secondary students which involves direct recording of sound files onto a hard disk. Candidates hear and see the first part of a dialogue and are required to type the second part of a dialogue onto the screen. The author suggests that this technique has the potential to avoid some of the problems of group audio-based testing. However, although this kind of test appears to offer a convenient way of conducting individualized assessments, dictation involves elements other than listening, as noted above, and it would not be appropriate to interpret the scores as indicators of listening ability.

2. The use of video

In the teaching of listening comprehension nowadays, a good deal of emphasis is given to the importance of providing adequate visual support to learners so as to enable them to activate their content schemata and to assist them in making predictions and inferences when a text has only been partially understood (Kang 1995, MacWilliam 1986, Mueller 1980, Ruhe 1996). However, although video has been widely used for this purpose in language teaching, video-mediated language testing remains relatively unexplored. It would clearly be advantageous to listening test designers to be able to incorporate video elements into test design as a way of providing the context that is often difficult to convey in audio-based tests. The results of a study by Progosh (1996) suggest that such a move would also be welcomed by learners. The vast majority of the 62 Japanese learners of English whom he surveyed expressed a preference for video-mediated listening tests over audio-recorded tests.

With the very rapid advances in instructional technology that are currently taking place, it can be anticipated that computer-adaptive tests of listening ability incorporating video and digitized speech will become available in years to come (Dunkel 1992). However, according to Burstein, *et al.* (1996), the capabilities of new technology remain to be fully realized owing to limitations in computer memory and difficulties of speech recognition. In the case of listening, this means that testees still have to type in responses even though they may have been presented with an aural prompt and visual stimulus. Burstein, *et al.* (1996:245) comment that this creates a situation in which "a relatively rich language presentation is followed by a limited productive assessment." They argue, however, that the principal barriers to the use of technology are conceptual rather than technical and highlight the need to develop explicit definitions of performance as a basis for fully functional communicative tests.

## CONCLUSION

Current views of listening suggest that it is a highly complex, individual, and interactive process in which listeners use a wide variety of verbal and non-verbal cues to interpret messages. This survey has demonstrated some of the difficulties involved in describing and assessing listening abilities, given the limitations of our knowledge concerning the nature of the construct itself and the various practical constraints imposed by the test situation (including the item formats that are used and the technology which is currently available).

Although some progress has been made in understanding the interaction among listeners, texts, and assessment tasks, a good deal of research remains to be conducted in order to place listening assessment on a more secure theoretical foundation. First, and probably most importantly, it is imperative that research continues into the further development and testing of explicit models of the listening construct such as that proposed by Dunkel, *et al.* (1993). Second, in order to understand better the way in which second language listeners arrive at interpretations of a text, there is an urgent need for more introspective studies of the kind conducted by Buck (1990). Both of these issues are crucial in establishing the validity of listening tests. Third, if testers are to have any empirical basis for designing test specifications and reporting listening test scores and levels, much more information is needed on factors affecting task, text, and item difficulty in listening tests and how these variables work singly and in combination. Fourth, different options for incorporating the notion of negotiated interaction into non-participative listening tests need to be explored and ways investigated of assessing interactive listening ability in direct tests of oral interaction such as interviews. Finally, since assessment practice appears to be lagging somewhat behind language teaching in its use of new technology, more collaborative development efforts among computer specialists, educationists, and linguists are needed in order to explore ways in which the recent rapid developments in electronic communication can be harnessed to develop listening tests which incorporate a much wider range of contextual elements and response modes than has hitherto been the case.

As Dunkel (1991a) points out, listening is becoming an increasingly important skill in contemporary postliterate society. In the future, it is to be hoped that this change in emphasis will bring about a concomitant increase in research activity into second language listening and ultimately lead to better procedures for assessing listening abilities.

## ANNOTATED BIBLIOGRAPHY

Buck, G. 1990. The testing of second language listening comprehension. Lancaster: University of Lancaster. Ph.D. diss.

This thesis, winner of the TOEFL Outstanding Dissertation Research Award for 1993, is a mine of information on second language listening comprehension and its assessment. It includes a thorough critical discussion of the nature of listening and an evaluative overview of methods and item formats used in tests of listening comprehension. The results of a number of carefully-conducted research studies into listening processes and listening test-taker behavior are reported, including a multi-trait multi-method study which sets out to establish whether listening and reading could be identified as separate traits (also described in Buck 1991), along with an investigation of test-taker behavior in listening tests using introspective techniques. The effects of item preview on test performance are also examined. On the basis of his research, the author concludes that listening comprehension is a highly complex and individual process involving the whole of the listener's knowledge and experience, and he explores the consequences of this finding for listening test design. In the process, he raises some fundamental questions concerning the nature of the listening trait and the appropriacy of applying latent trait methods of test analysis to a complex and multidimensional skill such as listening. A range of useful suggestions are given for further research into listening processes and listening assessment, along with an extensive list of references. This is a work of major significance which should be read carefully by anyone embarking on the development of second language listening tests.

Douglas, D. 1988. Testing listening comprehension in the context of the ACTFL proficiency guidelines. *Studies in Second Language Acquisition*. 10.245–261.

This article summarizes a range of tasks commonly used in listening comprehension testing and critically reviews the descriptions of proficiency which figure in the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines (ACTFL 1987), posing a number of questions concerning the validity of the proposed task and skill hierarchies. The author considers four key issues affecting the testing of listening: the feasibility of criterion-referencing, the lack of information on the influence of context on performance, the problems of defining specific domains of language use, and the potential for using new forms of technology in listening assessment. The article highlights a number of important questions concerning listening test validity which are in need of further research and provides helpful insights into the challenges

involved in attempting to develop listening tests which are based on purposeful language use.

Dunkel, P., G. Henning and C. Chaudron. 1993. The assessment of a listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal*. 77.180–191.

This article reports on an attempt to design a detailed framework for describing and assessing second language listening comprehension. The proposed framework contains a range of components which need to be taken into account when developing tests of listening comprehension. These include the purpose and context of assessment, the person domain (cognitive factors, affective factors, etc.), the cognitive operations required of the testee, and the texts, tasks, items, and scoring methods used. The authors provide a helpful set of "leveling variables" according to which listening tasks can be made to vary in difficulty (such as cultural proximity, length, dialectal variation, degree of contextual support provided, etc.). A number of suggestions are made concerning ways in which the framework could be expanded and used for practical test development purposes. The model is soundly based in the research literature and provides a useful starting point for test design, particularly in the context of computer-adaptive testing, but, as the authors point out, it needs to be tested "against real world situations and purposes."

Henning, G. 1991. *A study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on TOEFL listening comprehension item performance*. Princeton, New Jersey: Educational Testing Service. [TOEFL Research Report 33.]

This research study was commissioned in response to the following concerns expressed about the TOEFL listening comprehension component: 1) it makes excessive demands on short-term memory; 2) it relies on reading responses, invalidating the test as a test of listening; and 3) it tests too many trivial details which are not essential for overall passage comprehension. In relation to the first question, the researcher examined the effects of both repeating and lengthening the stimulus passage and found no effects on either item quality or task validity. He concludes that there is no evidence to suggest that the TOEFL item formats over-burden candidates' memory capacity. Concerning the second question, Henning found that items accompanied by shortened options showed a non-significant tendency to demonstrate greater discriminability and a significant tendency to demonstrate greater response validity, suggesting that response lengths could profitably be shortened. The examination of processing hierarchy showed a significant effect on item discriminability: Interestingly, 'lower level' items proved to be better discriminators than

'higher level' items, recalling the findings of de Jong and Glas (1987). On the basis of this finding, Henning concludes that there is no reason why the test should contain a majority of items requiring top-down strategies and, in fact, suggests that including more lower order items might result in an increase in the test's construct validity. The latter finding, however, as the author concedes, needs to be treated with caution due to potential problems with the way in which processing level is operationalized in the study. Here, in order to avoid the difficulties experienced in previous studies in which expert judges were unable to agree on the classification of higher and lower order items (cf. Alderson 1990a, Alderson and Lukmani 1989), the notion is defined simply in terms of the amount of discourse which candidates have to process in order to respond to the item (i.e., whether information had to be extracted from one, two, or three sentences of the stimulus passage). This seems a rather simplistic definition which fails to take into account a variety of other factors which may affect processing difficulty, including, *inter alia*, background knowledge, the nature of the test question, and the type of information required.

Nissan, S., F. DeVincenzi and K. L. Tang. 1996. *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension.* Princeton, New Jersey: Educational Testing Service. [TOEFL Research Report 51.]

This report describes a study which set out to investigate factors contributing to item difficulty in the dialogue section of the TOEFL listening comprehension component. Of the seventeen variables hypothesized as likely to have an effect, five were found to be significant: 1) the presence of infrequent oral vocabulary, 2) the sentence pattern of the utterances in the stimulus, 3) the presence of negatives in the stimulus, 4) the requirement of an inference for an item, and 5) the role of the speaker in the stimulus. The researchers point out, however, that these findings are to some extent counterintuitive since the remaining twelve variables had previously been used, reportedly with some success, to increase the number of difficult items in the TOEFL pool. They hypothesize that these findings may be due to the methodology used which analyzed the effects of the item features in uncontrolled combination and thus did not allow the specific effects of individual features to be clearly identified. The researchers recommend that future studies need to be conducted in which all characteristics of the item should be kept constant except the one under investigation. They also suggest that further research needs to be conducted into the functioning of questions testing interpretive inferences. Although this study is only concerned with the multiple-choice item format, the list of variables affecting item difficulty provided by the authors offers a useful basis for further experimental studies in which these variables can be systematically manipulated. The researchers' discussion

of the complexities involved in testing inferencing via the multiple-choice format raises the question of the validity of such items and should be noted by listening test designers. Given the uncertainties surrounding the relationship between item responses and cognitive processes (cf. Brindley forthcoming a, Buck 1990;1994), their call for research into the relationship between multiple-choice and constructed responses to inferencing items is also worth taking up.

Rost. M. 1990. *Listening in language learning*. London: Longman.

This is a thoroughly researched and detailed overview of second language listening processes. It contains chapters covering the role of listening in verbal communication, the nature of processing and inferencing in listening performance, and listening in transactional (non-collaborative) discourse. Listening skills and strategies employed by both first and second language learners are described and analyzed. The author also considers the place of listening in the language curriculum and provides a useful set of principles to guide listening task design. Assessment issues are addressed in a chapter which includes examples of various types of listening tests and item types. The chapter on listener performance which is accompanied by examples of authentic conversational data provides some interesting insights into the part which listening skills play in collaborative discourse and would be of particular interest to designers of tests of oral interaction.

Shohamy, E. and O. Inbar 1991. Validation of listening comprehension tests: The effect of text and question type. *Language Testing*. 8.23–40.

This study set out to investigate the effect of text and item type on listening test scores. The researchers administered listening tests to 150 Israeli secondary school EFL learners based on three different types of texts: a news broadcast, a lecturette, and a consultative dialogue. They found that the texts ranged in difficulty along a continuum corresponding to the degree of 'orality' of the text in question. They attribute this effect to the greater redundancy and interactivity of the dialogue and lecturette as opposed to the propositional density and conciseness in the news broadcast. The researchers also compared subjects' performance on global items, which require the testee to synthesize information, make inferences, and draw conclusions, with their performance on local items, requiring them to locate details, understand words with contextual support, paraphrase, and recognize facts. At the same time, they examined the effects of trivial questions testing examinees' ability to recall unimportant factual details from memory. They found that the global questions were more difficult than the local ones and that trivial questions showed inconsistent results, suggesting that the memory load they impose on

testees may interfere with processing of more important information. On the basis of their findings, the authors recommend that listening tests should sample a range of genres with different degrees of 'listenability,' that both local and global question types should be included, and that trivial questions should be avoided in listening comprehension tests. Emerging from this study are a number of principles concerning listening task/text difficulty and item construction which could profitably be drawn on by test designers.

Thompson, I. 1995. Testing listening comprehension. *AATSEEL Newsletter*. 37.24–31.

This article provides an excellent practical summary of the issues involved in testing listening comprehension. The author begins by considering the special qualities of the aural medium which make the testing of listening different from testing of reading. She then goes on to describe a range of factors which need to be taken into account in selecting suitable passages, making the important point that the level of difficulty of an item in a given passage is not just a function of the text itself but lies in the interaction of text, task, background knowledge, memory, and inferencing ability. Various sources of measurement error in listening tests are then identified. In particular, readers are alerted to the need to consider the demands made by test tasks on candidates' memory capacity and inferencing ability. A brief discussion of the advantages and disadvantages of commonly used response types follows. These include multiple-choice items, true-false items, open-ended questions, recall protocols, and non-verbal responses. Issues of test presentation and administration such as the role of question preview and contextualization cues, the language of instructions, and the need for uniformity of presentation are then discussed and illustrated. The author concludes with some brief practical suggestions on the need for piloting and item analysis.

Weir, C. J. 1993. *Understanding and designing language tests*. London: Longman.

This book is a useful practical manual which contains chapters on testing spoken interaction, reading, listening, and writing, prefaced by a discussion of issues in language testing and a set of general guidelines for test construction. The author highlights the necessity for test developers to begin with a theoretical model of the skill being assessed and organizes each chapter on skills assessment around a three-part framework which provides a checklist of the operations (skills) involved, performance conditions (factors which may affect testees' performance), and criteria for assessing the quality of output. The chapter on listening comprehension testing contains a clear discussion of issues and problems in listening test

Berne, J. E. 1995. How does varying pre-listening activities affect second language listening comprehension? *Hispania*. 78.316–329.

Boyle, J. 1984. Factors affecting listening comprehension. *ELT Journal*. 38.34–38.

_____ and D. L. K. Suen. 1994. Communicative considerations in a large-scale listening test. In J. Boyle and P. Falvey (eds.) *English language testing in Hong Kong*. Hong Kong: The Chinese University Press. 32–55.

Brindley, G. Forthcoming a. Investigating second language listening ability: Listening skills and item difficulty. In G. Brindley and G. Wigglesworth (eds.) *ACCESS: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

_____ Forthcoming b. Describing language development? Rating scales and second language acquisition. In L. F. Bachman and A. D. Cohen (eds.) *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

_____, S. Hood, C. McNaught and G. Wigglesworth. Forthcoming. Issues in test design and delivery. In G. Brindley and G. Wigglesworth (eds.) *ACCESS: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

_____ and D. Nunan. 1992. Draft bandscales for listening. Sydney: National Centre for English Language Teaching and Research, Macquarie University. [IELTS Research Project 1.]

_____ and S. Ross. 1997. Trait-method comparisons across three language test batteries using exploratory, MTMM and structural equation modelling approaches. Paper presented at 19th Annual Language Testing Research Colloquium. Orlando, Florida, March 1997.

Brown, G. and G. Yule. 1983. *Teaching the spoken language*. Cambridge: Cambridge University Press.

Buck, G. 1991. The testing of listening comprehension: An introspective study. *Language Testing*. 8.67–91.

_____ 1992. Listening comprehension: Construct validity and trait characteristics. *Language Learning*. 42.313–357.

_____ 1994. The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*. 11.145–170.

Burger, S. and J. Doherty. 1992. Testing receptive skills within a comprehension-based approach. In R. J. Courchene, J. I. Glidden, J. St. John and C. Thérien (eds.) *Comprehension-based second language teaching*. Ottawa: University of Ottawa Press. 299–318.

Burstein, J., L. T. Frase, A. Ginther and L. Grant. 1996. Technologies for language assessment. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 16. Technology and language*. New York: Cambridge University Press. 240–260.

Call, E. M. 1985. Auditory short-term memory, listening comprehension and the input hypothesis. *TESOL Quarterly*. 19.765–781.

design, illustrated with examples from the research literature. The author provides a useful set of practical guidelines for developing test tasks which cover text selection, recording, item construction, trialling, and marking. Numerous examples of different test types and item formats are given, together with a discussion of their advantages and disadvantages. Indirect techniques such as dictation and listening recall are also described. Although oriented towards academic listening and somewhat lacking in concrete detail on procedures for test analysis, this chapter gives a comprehensive and accessible summary of what is involved in listening test construction, particularly at the item development stage.

## UNANNOTATED BIBLIOGRAPHY

Alderson, J. C. 1988. *Innovation in language testing: Can the micro-computer help?* Lancaster: University of Lancaster. [Special Report No 1: Language Testing Update.]

_____ 1990a. Testing reading comprehension skills (Part One). *Reading in a Foreign Language*. 6.2.425–438.

_____1990b. Testing reading comprehension skills (Part Two). *Reading in a Foreign Language*. 7.1.465–503.

_____, C. Clapham and D. Wall. 1995. *Language test construction and evaluation*. Cambridge: Cambridge University Press.

_____, and Y. Lukmani. 1989. Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*. 5.2.253–270.

American Council on the Teaching of Foreign Languages. 1986. *ACTFL Proficiency Guidelines*. Hastings-on-Hudson, NY: ACTFL.

Anderson, A. and T. Lynch. 1988. *Listening*. Oxford: Oxford University Press.

Bacheller, F. 1980. Communicative effectiveness as predicted by judgements of the severity of learner errors in dictation. In J. W. Oller Jr. and K. Perkins (eds.) *Research in language testing*. Rowley, MA: Newbury House. 66–71.

Bachman, L. F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baltova, I. 1994. The impact of video on the comprehension skills of core French students. *Canadian Modern Language Review*. 50.507–521.

Berne, J. E. 1993. The role of text type, assessment task and target language experience in L2 listening comprehension assessment. Paper presented at the annual meetings of the American Association for Applied Linguistics and the American Association of Teachers of Spanish and Portuguese. Atlanta, GA, March 1993, and Cancun, Mexico, August 1992. [ED 358737.]

Hughes, A. 1989. *Testing for language teachers*. Cambridge: Cambridge University Press.

Kang, S. 1995. The effects of a context-embedded approach to second-language vocabulary learning. *System*. 23.143–155.

Kelly, P. 1991. Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL*. 29.135–149.

Lewkowicz, J. 1991. Testing listening comprehension: A new approach? *Hong Kong Papers in Linguistics and Language Teaching*. 14.25–31.

Long, D. R. 1990. What you don't know can't help you. *Studies in Second Language Acquisition*. 12.65–80.

Long, M. 1985. Input and second language acquisition theory. In S. M. Gass and C. G. Madden (eds.) *Input in second language acquisition*. Rowley, MA: Newbury House. 377–393.

MacWilliam, I. 1986. Video and language comprehension. *ELT Journal*. 40.131–135.

Messick, S. 1989. Validity. In R. Linn (ed.) *Educational measurement*. Washington: American Council on Education and National Council on Measurement in Education. 13–103.

Mueller, G. A. 1980. Visual contextual cues and listening comprehension: An experiment. *Modern Language Journal*. 64.335–340.

Munby, J. 1978. *Communicative syllabus design*. Cambridge: Cambridge University Press.

Powers, D. 1986. Academic demands related to listening skills. *Language Testing*. 3.1–38.

Progosh, D. 1996. Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal*. 14.1.34–44.

Rea, P. M. 1985. Language testing and the communicative teaching curriculum. In Y. P. Lee, R. Lord, A. C. Y. Y. Fok and G. D. Low. *New directions in language testing*. Oxford: Pergamon. 15–32.

Richards, J. C. 1983. Listening comprehension: Approach, design, procedure. *TESOL Quarterly*. 17.219–240.

Ross, S. and J. Langille. Forthcoming. Negotiated discourse and interlanguage accent effects on a second language listening test. In G. Brindley and G. Wigglesworth (eds.) *ACCESS: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Rost, M. 1994. *Introducing listening*. London: Penguin.

Rubin, J. 1994. A review of second language listening comprehension research. *Modern Language Journal*. 78.199–221.

Ruhe, V. 1996. Graphics and listening comprehension. *TESL Canada Journal*. 14.1.45–60.

Schmidt-Rinehart, B. C. 1994. The effects of topic familiarity on second language listening comprehension. *Modern Language Journal*. 78.179–189.

Schrafnagl, J. and D. Cameron. 1988. Are you decoding me? The assessment of understanding in oral interaction. In P. Grunwell (ed.) *Applied linguistics*

Carroll, B. J. and P. Hall. 1985. Make your own language tests. Oxford: Pergamon.

_____ and R. West. 1989. *ESU framework*. London: Longman.

Cohen, A. D. 1994. *Assessing language ability in the classroom*. Boston: Heinle and Heinle.

Coniam, D. 1996. Computerized dictation for assessing listening proficiency. *CALICO Journal*. 13.3.73–85.

Corbel, C. 1993. *Computer-enhanced language assessment*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Courchene, R. J., G. I. Glidden, J. St. John and C. Thérien (eds.) 1992. *Comprehension-based second language teaching*. Ottawa: University of Ottawa Press.

de Jong, J. H. A. L. 1987. Defining tests for proficiency in listening comprehension: A response to Dan Douglas's "Testing listening comprehension." In A. Valdman (ed.) *Proceedings of the symposium on the evaluation of foreign language proficiency*. Bloomington, Illinois: Indiana University Press. 115–124.

_____ and C. Glas. 1987. Validation of listening comprehension tests using item response theory. *Language Testing*. 4.170–192.

Dunkel, P. 1991a. Listening in the native and second/foreign language: Toward an integration of research and practice. *TESOL Quarterly*. 25.431–457.

_____ 1991b. Computerized testing of nonparticipatory L2 listening comprehension proficiency: An ESL prototype development effort. *Modern Language Journal*. 75.64–74.

_____ 1992. The use of PC-generated speech technology in the development of an L2 listening comprehension proficiency test: A prototype design effort. In M. C. Pennington and V. Stevens (eds.) *Computers in applied linguistics: An international perspective*. Clevedon, Avon: Multilingual Matters. 273–293.

Ellis, R., Y. Tanaka and A. Yamazaki. 1994. Classroom interaction, comprehension and the acquisition of L2 word meanings. *Language Learning*. 44.449–91.

Faerch, K. and G. Kasper. 1986. The role of comprehension in second-language learning. *Applied Linguistics*. 7.257–274.

Feyten, C. M. 1991. The power of listening ability: An overlooked dimension in language acquisition. *Modern Language Journal*. 75.173–180.

Hansen, C. and C. Jensen. 1994. Evaluating lecture comprehension. In J. Flowerdew (ed.) *Academic listening: Research perspectives*. Cambridge: Cambridge University Press. 241–268.

_____ 1995. The effect of prior knowledge on EAP listening-test performance. *Language Testing*. 12.99–119.

Heaton, J. B. 1988. *Writing English language tests*. London: Longman.

Hendrickson, J. M. 1992. Creating listening and speaking prochievement tests. *Hispania*. 75.1326–1331.

*in society*. London: Centre for Information on Language Teaching and Research. 88–97.

Sherman, J. 1997. The effect of question preview in listening comprehension tests. *Language Testing*. 14.185–213.

Shohamy, E. 1985. *A practical handbook in language testing for the second language teacher*. Experimental edition. Tel Aviv: Tel Aviv University.

Spolsky, B. 1994. Comprehension testing, or can understanding be measured? In G. Brown, K. Malmkjaer, A. Pollitt and J. Williams (eds.) *Language and understanding*. Oxford: Oxford University Press. 141–152.

Thompson, I. 1996. Assessing foreign language skills: Data from Russian. *Modern Language Journal*. 80.47–65.

Underwood, M. 1989. *Teaching listening*. New York: Longman.

University of Cambridge Local Examinations Syndicate (UCLES)/Royal Society of Arts (RSA). 1990. *Certificates in communicative skills in English*. Cambridge: RSA/UCLES.

University of Cambridge Local Examinations Syndicate (UCLES). 1995. *First Certificate in English*. Cambridge: UCLES.