DOCUMENT RESUME

ED 267 597 FL 015 491

AUTHOR Barnwell, David
TITLE The Audiolingual Tradition in Foreign Language Testing.
PUB DATE 86
NOTE 19p.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Audiolingual Methods; *Contrastive Linguistics; *Language Proficiency; Language Research; Language Skills; Language Tests; Second Language Instruction; Structural Grammar; *Test Format; *Testing

ABSTRACT

Approaches to the testing of foreign langauge proficiency have tended to mirror prevailing philosophies in foreign language teaching, and for many years, no serious effort was made to devise oral proficiency measures. However, after World War II, structural linguistics applied to the classroom produced audiolingualism, which was a method heavily influenced by behavioral psychology. New interest in estimating the quantity of sounds and structures learned as habit led to discrete-point testing, a trend reflected in the American emphasis on standardized testing and a variety of discrete-point test formats. Audiolingualism never delivered the results its theoreticians promised, either empirically or theoretically. Contrastive analysis, while useful in examining learner errors arising from differences between languages, will not predict errors from irregularities within a language, a major source of learner problems. In addition, language points are not really discrete, and are not entirely subject to the relatively inflexible discrete-point measurement. Language proficiency is increasingly viewed as a global skill, and language testing has become more integrative. (MSE)

The Audiolingual Tradition in Foreign Language Testing

by David Barnwell

I.1. Approaches to the testing of foreign language proficiency have tended for many years to mirror prevailing philosophies in foreign language teaching. Thus, during the decades of grammar-translation dominance in language teaching, no serious effort was made to devise means of gauging oral proficiency. For instance, of the nineteen tests of foreign languages reviewed in Buros (1949), only three tested auditory comprehension and/or perception, and none sought to measure oral production. Translation was the favourite device for assessing grammar and vocabulary, though in the opinion of those who wrote the reviews in Buros, the type of language production called for was often unnatural,and the vocabulary seemed arbitrarily selected.

Further evidence for the undeveloped state of foreign language testing as late as the 1940s can be found in Agard and Dunkel's major investigation of the situation of foreign language teaching. Reviewing the position as they found it, they concluded that there was no consensus among teachers regarding what to test or how to test it: "as for actual tests of oral production, we know of none in published form for general use" (1948,55).

Some years later, Furness (1953), in reporting on then available auditory tests in Spanish, observed that there was not yet a single aural-oral test which had been validated; the position was no better in the case of other languages.

Agard and Dunkel put forward several reasons for the failure to teach or test the ability to speak a foreign language. They pointed out that the large class-size of the time, often as many as forty students in a class, was a major obstacle to a concentration on oral ability. This factor was aggravated by the short length of language courses, and by the low level of oral skills among the teaching population itself. In addition, since international transport and communication were still difficult, there seemed no need to teach or test an ability which might never be used. Far wiser, it seemed, was it to view language learning as a mental training, and as a means of introduction to the great literatures of the world. If such were the goals, then there was no need to test oral proficiency.

1.2. In the decades after World War II, however, a number of factors combined to create an utterly new view of language testing. From theoretical linguistics came the ideas of American Structuralism, characterized by a belief that languages should be seen as independent structures which were best studied through analysis of their component parts. Though often putting forward quite disparate views, adherents of the Structuralist philosophy shared a faith in the methodology of analysis and classification, and sought to isolate and list the discrete elements of the languages they studied. To such as extent was this the case that Chomsky and Halle (1968, 402n) rather disparagingly labelled this approach as 'taxonomic linguistics'.

Structural linguistics applied to the classroom produced Audiolingualism. Its teaching practices were heavily influenced by applications of the Behaviourist model of learning. Robert Lado, the great apostle of Structuralism in foreign language testing, put it in these terms:

> The lowly power of habit is the support of the distinctively human gift of language. We can in this sense speak of language as a conventionalized, highly complex system of habits which functions as a human instrument of communication (1964,4).

Lado's concept of language is typical of that which dominated the literature on foreign language teaching and testing methodology in the 1960s. In Stockwell and Bowen's words:

> The process of language learning can be viewed as an experience by means of which new habits of sentence formation are acquired and used for communication. (1965,295)

Logically, therefore, the difficulty in learning a foreign language results from the very newness of these habits--the fact that they differ from the 'old' habits learned as first language. A second language is difficult because it is different. Lado in fact dubbed as 'non-problems' for the learner those structural areas which do not differ between first and second language. For him, these elements "are transferred from the native language, and since they function satisfactorily, they do not have to be learned anew". The 'problems' were thus the true domain of the language test. In Lado's aphorism, "testing the problems is testing the language" (1964,20).

If learning a foreign language entailed mastering a finite list of 'problems' until the entire set of sounds and structures had been acquired, it followed that the learner's proficiency could best be gauged by estimating the quantity of these sounds and structures that he had at any given time learned as habit. In 1961 John Carroll coined the term by which this procedure has since been known - the 'discrete-point' approach. Taxonomic linguistics had given birth to taxonomic testing.

The creation of lists of likely 'problem' areas arising from the learner's need to acquire new habits provided fertile ground for applied linguists in the 1960s. Techniques which involved a contrastive analysis of English and Spanish, for example, were used by Politzer and Staubach (1961), Bull (1965), and Stockwell and Bowen (1965). Not surprisingly, they provided the basis for drawing up areas and elements to be covered in testing.

The educational climate of the United States, with its stress on standardized testing, was reflected in Structuralist testing practices. It is for this reason that Spolsky (1978) juxtaposes the terms 'psychometric-Structuralist' when categorizing the foreign language testing of the 1960s. The need to set examinations for large numbers of candidates, in conjunction with technological advances which permitted these tests to be scored quickly and cheaply, caused language tests to evolve towards an objective, multiple-choice format. This tendency formed part of the twentieth century tradition of standardized testing, even of such an elusive construct as intelligence.

In the late 1950s and early 1960s, input from theoretical linguistics and standardized testing became dominant. As John Oller puts it:

> In the early tradition of Structural linguistics, particularly the Bloomfieldian variety of Behaviorism, psychometry found a willing partner, and the conjugal result was ... the discrete-point approach to language testing (1976, 142).

1.3. The conceptual framework of the classical discrete-point test can be thought of as being composed of two axes. One axis represents the skill to be tested - broadly speaking, the four principal skills of listening, speaking, reading, writing. The other represents the four major classificatory components used by the Structuralists: phonology/orthography, morphology, syntax, lexicon. Within this axis, there is room for ever-greater subdivision into the discrete points of the language. The domain of the test can be considered as the intersection of lines drawn on each axis, where a specific kind of knowledge shows itself in a specific kind of behaviour. Thus, one could not properly describe a test of this kind as a test of Spanish, to take an example. One could only speak of measuring a particular language element as evinced in a particular skill - trill /r/ in speech, for example, or the imperfect/preterite distinction in reading. One could only 'test one thing at a time'.

An inevitable result of this is a very high rate of specialization in the tester's focus. "When judging the student's pronunciation, the teacher should grade only one sound per utterance. When evaluating the use of grammatical forms, he should grade only the grammar, and not focus upon the pronunciation and the vocabulary" (Chastain 1971, 332).

The tendency to subdivide shows itself equally in the range of testing methods used. The MLA Testing Handbook (Paquette and Tollinger 1968) contains a grid in which one axis is formed by the four principal skills. The points on the other axis consist of eleven different types of test format, e.g., multiple-choice, completion, matching, true-false, etc. Theoretically, any particular skill could be measured by almost any of these methods.

Different applications of the discrete-point philosophy can be found in Lado (1964), Harris (1969), and Valette (1967,1977). Descriptions of the MLA tests are given in Clark (1965) and Bryan (1966). A brief example from Pimsleur (1966, 204) may suffice to show the theory in practice:

> Language Area: 1/ Phonemes 2/ Intonation
> Skill: Speaking
> Testing Device: Mimicry
>
> The examinee repeats eight sentences after a tape-recorded native voice. In each sentence, one phoneme or combination of phonemes is scored right or wrong; in addition, several sentences are scored for intonation, again, right or wrong.

It would be incorrect to think that such testing methods died with the 1960s. As late as 1977 one finds the author of a handbook for foreign language teachers stating dogmatically: "Modern evaluation methods call for testing procedures that measure each skill directly and as a separate entity" (Grittner 1977, 341). Few data are available on testing practices in high school or university in the 1980s, but it would be surprising if strong elements of the discrete-point heritage did not persist. After all, any teacher who

was trained in the 1960s or early 1970s is unlikely to have been exposed to anything other than discrete-point testing. Thus, in the light of this resilience, it is not otiose to examine the theoretical and empirical bases of the testing practices of the 1960s.

1.4. Before assessing the evidence concerning the several components of the testing methodology associated with Audiolingualism, it is worth mentioning in passing that Audiolingualism itself never delivered the results which some of its theoreticians had promised.

On the theoretical level, the Structuralist and Behaviourist underpinnings of Audiolingualism received a lethal blow from Chomsky's new paradigm (1959). On the practical level, studies such as the Pennsylvania Project (Smith 1970), showed that Audiolingualism achieved results no better than a more traditional methodology. While recognizing the problems in research design encountered by ambitious studies such as Smith's, it is still striking that they yielded little or no empirical backing to the claims of the Audiolingualists.

Research has also shown that Audiolingual testing tenets are similarly weak in their foundations. Let us first deal with contrastive analysis. This in its 'strong version' (Wardhaugh 1970) purports to predict learner behaviour. There is however no body of research evidence to support this claim. Indeed, as early as 1966, at the Northeast Conference, a traditional stronghold of Audiolingualism, widespread dissatisfaction with the empirical basis of contrastive analysis was expressed (Ferguson 1966). In 1968, for example, John Carroll described two tests designed for a language proficiency programme in the U.S. Army. One test was explicitly drawn

up on the basis of contrastive analysis, while the other was composed of 'a more or less random assortment' of 'intuitively good' items. The 'intuitively' designed test proved to be the more valid; its validity was deemed equal, if not even superior to that of the 'contrastive' test. More rigorous studies such as those in George (1972), Oller and others (1972), Whitman and Jackson (1972), and Wilson (1977), to mention but a few, have shown that the 'strong version' of contrastive analysis is untenable.

In fact, the very evidence put forward by Lado as the empirical proof of his theory was later subverted by Dulay and Burt's reappraisal of his data. Lado's claims were based to a large extent on studies of patterns in the learning of English by Norwegian immigrants in the United States (1957). However, in their fresh analysis of the data, Dulay and Burt demonstrated that interlingual interference affected these speakers' use of Norwegian to a much greater extent than it caused errors in their use of English as a second language. Dulay and Burt concluded that contrastive analysis had relied for empirical support on impressionistic observation and intuition (1974, 104).

All of this does not serve to deny the utility of contrastive analysis in designing instructional materials, or ignore the insights it may yield in the study of learners' errors. This is the 'weak version', to use Wardhaugh's phrase, and it is a version few could dispute. But it leaves contrastive analysis with the status of just one of a number of useful methodologies, not an indispensable predictive theory. In fact, contrastive analysis has never been shown to be a prerequisite for the creation of valid foreign language tests.

It is somewhat surprising that no significant empirical validation studies of contrastive analysis were carried out in the 1960s, given its adherents' pride in the scientific status of their methodology. It is fruitless to search the contrastive literature of the time for anything more convincing than anecdote or impression. Further, aside from its non-existent empirical basis, the theory suffered from two inter-related logical flaws. Firstly, if languages are different, as the Structuralists had stressed, how could they be legitimately compared or contrasted ? In fact, the more different are two languages, the harder it is to make meaningful comparisons or contrasts, and so the theory copes worst with those cases with which it should cope best. We can fairly fruitfully compare English and a Romance language, but on what terms can we compare English and an American Indian language, for example ?

The second thoretical anomaly in contrastive analysis was pointed out as early as 1962 by John Upshur. He noticed that once the individual begins to learn, he is no longer the 'pure native speaker' demanded by the theory: "all of what he has learned will have facilitation or interference effects upon what has not yet been taught" (1962,116). To use a term coined later, the student now possesses an 'interlanguage' (Selinker 1974), composed of sets of hypotheses about the target language, based upon observations of both it and the native language. Contrastive analysis, if it were to succeed, would have to refer to interlanguage as well as native and target language. Yet each individual's interlanguage is different and constantly changing, so the task is impossible. Contrastive analysis is static, while language learning is dynamic.

Arising from the new perspectives introduced by Chomsky, research in the 1970s turned towards the discovery of what languages have in common rather than what sets them apart. Within the areas most germane to language testing, the focus for study became that of the common patterns of language acquisition shared by both natives and nonnatives. An impressive body of work was produced which showed that all learners have a lot in common (Dato 1970, Ervin-Tripp 1974, Boyd 1975). Though error analysis has not yet permitted us to quantify the importance of different sources of learner errors, it has shown that a great many learning problems arise from irregularities within a language rather than differences between languages. Contrastive analysis will not predict errors in these areas.

1.5. The other great principle of testing in the Structuralist tradition appears to be equally untenable. Discrete-point testing is even more fundamental to Structuralist approaches to language testing than is contrastive analysis, but its foundations are equally weak. This is so despite the 'scientific' aura which accompanies many discrete-point tests. Indeed, it can be argued that discrete-point testing is actually much less sophisticated than the technology and techniques which often accompany it. It would clearly be foolish to suppose that a person who knows a set of vocabulary items, be it a

hundred, a thousand, or n words, could ipso facto be considered to know a particular language. Yet discrete-point testing assumes that a person knows a language when he has acquired a finite set of structural items - an assumption that may be no more tenable than the previous one.

If language points were really discrete, then it would of course make sense to measure them discretely. However, despite the injunctions not to contaminate observations of one skill with another, discrete-point testing has never devised procedures for actually isolating discrete points for measurement. Indeed, it could be said that the only time at which a pure discrete-point test could be administered would be at the end of day one in the language classroom. As proficiency increases, it becomes impossible to test points discretely. In fact, the more the candidate exhibits the construct to be measured, i.e. proficiency, the less able is the discrete-point test to measure it. This is surely a crippling weakness in any test.

All tests of anything but the simplest construct must make do with a mere sample of the behaviour or construct they wish to measure: they cannot hope to cover every single instance and element. This is true of any test of language proficiency that is likely to be devised. It can therefore be admitted that a selection of 'points'

has to be made when testing. Clearly it is the business of the tester to select that sample, based upon a theory, practical experience, pragmatic constraints, or whatever preconceptions he may bring to the task. In discrete-point tests, however, the evaluation of candidates' responses is left to a machine, or at least can be performed mechanically. The simplified scoring mechanisms which are compatible with an 'objective' format are inconsistent with fine discrimination as to quality of response.

Actually, discrete-point testing deliberately cultivates an inflexible mode of scoring. In Lado's words, the best tactic is "to list for the examiner the specific point in the problem which decides whether the response is right or wrong, and to instruct the examiner to disregard everything else" (1964). Proficiency is thus gauged on the sum of the elements that are scored as correct, not on how those elements are combined in normal language. It is the examiner who arrives at this sum, not the examinee. The reductio ad absurdum of this tendency can be seen in Lado's unconsciously ironic statement:

> We are thus able to break away from having to ask the student to speak when we test his ability to speak, since this process is inaccurate and uneconomical (1960,160).

The validity of the discrete-point hypothesis has never been proven empirically. If language proficiency can be subdivided into

abilities in different skills, it is obvious that tests of these skills should not inter-correlate too well. Theoretically, in fact, if the abilities are truly discrete, they should not inter-correlate at all, though no discrete-point advocate has ever adopted this position. On the other hand, if tests of supposedly discrete abilities inter-correlate at a high level, this suggests that it is idle to think of these abilities as separate entities. Rather should they be viewed as different manifestations of perhaps only one underlying factor of proficiency.

There is abundant evidence in the literature that separate tests on areas of language such as grammar, vocabulary, listening comprehension, etc., correlate very highly with each other, or, put another way, they load heavily on a common factor. In the case of the MLA Proficiency Tests, for example, Paquette and Tollinger (1965) calculated that between .80 and .90 of the variance could be ascribed to one general factor. Myers and Melton (1964), in a study of the MLA Cooperative Tests, found that there was no pattern by which scores on particular subskills correlated better with each other than with scores on different subskills. In the case of the TOEFL, Hosley and Meredith (1979) showed that the component subtests all correlated with the total at around .80, a high figure given the great

disparities among the population which takes the TOEFL. Upshur (1971) found a higher correlation between an oral communication test and a written composition than between the oral test and a discrete-point speaking test. Rand (1972) and Stubbs and Tucker (1974) produced parallel findings for tests of English as a foreign language, again showing discrete subtests loading heavily on a common factor. More recently, Oller and Perkins (1980) edited a large number of studies by themselves and others, whose results would be anomalous were the discrete-point hypothesis correct.

The significance of these data is all the greater when it is remembered that, given the imperfect state of the testing art, there is an inbuilt tendency for divergence between any two measurements. Error of measurement will always prevent inter-correlations between language tests from approaching too near to 1.00. In this light, intercorrelations in the .70 to .90 range, which are typically produced in language tests, are very impressive. This is even more so when one makes allowance for the diverse learning backgrounds of those who take language tests. This aspect of the TOEFL has already been referred to, but it is almost equally relevant to foreign language tests for English speakers. Some students are taught in ways that f ster oral and aural abilities, while others still concentrate

on reading and writing. It does not weaken the case against the discrete-point hypothesis to admit that students are more likely to learn what they are taught rather than what they are not taught.

Given findings and considerations such as those that have been discussed, researchers within the past decade or so have increasingly operated within a construct of proficiency as a global skill. As John Carroll, who for long had worked within the discrete-point tradition, was forced to admit:

> We have the paradox that the more we attempt to measure different skills, and the better our measurements of these skills, the higher the correlations among the skills, and thus the more they appear to converge towards the measurement of a single all-embracing skill (1973,11).

None of this overlooks the fact that individuals exhibit differing patterns of strengths and weaknesses, often quite striking. Nor does it preclude the possibility that future researchers will be able to isolate more than one factor of proficiency. But testers are no longer striving to smash the mosaic of language - rather are they seeking ways which enable the examinee to put all the pieces together. In short, we have reached the era of integrative foreign language testing.

REFERENCES CITED

Agard, Frederick, and Harold Dunkel. 1948. An investigation of second-language teaching. Chicago: Ginn & Co.

Boyd, Patricia A. 1975. The development of grammar categories in Spanish by Anglo children learning a second language. TESOL Quarterly, 9,2: 125-35.

Bryan, Miriam. 1966. Tests with a new look and a new purpose: the MLA Cooperative FL tests. Washington, D.C.: National Education Association. ED 012154.

Bull, William. 1965. Spanish for teachers: applied linguistics. New York: Ronald Press.

Buros, O.K. 1949. Third mental measurements yearbook. Highland Park, New Jersey: Gryphon Press.

Carroll, John B. 1961. Fundamental considerations in testing for English language proficiency of foreign students. In John B Carroll ed. Testing the proficiency of foreign students. Washington, D.C.: Center for Applied Linguistics, 30-40.

Chastain, Kenneth. 1971. The development of second language skills. Philadelphia: Center for Curriculum Development.

Chomsky, Noam. 1959. Review of Skinner's 'Verbal behaviour'. Language, 35,1:26-58.

and Morris Halle. 1968. The sound pattern of English. New York: Harper and Row.

Clark, John L.D. 1965. MLA Cooperative Foreign Language Tests. Journal of Educational Measurement, 2,2: 234-44.

Cooper, Robert L. 1968. An elaborated language testing model. Language Learning, special issue no. 3, 57-72.

Dato, Daniel P. 1970. American children's acquisition of Spanish syntax in the Madrid environment: preliminary edition. Washington, D.C.:Institute of International Studies.

Dulay, Heidi, and Marina Burt. 1974. You can't learn without goofing; an analysis of children's second language 'errors'. In Richards, 95-123.

Ervin-Tripp, Susan. 1974. Is second language learning like the first ? TESOL Quarterly, 8,2:111-27.

Ferguson, Charles. 1966. Applied linguistics. In Robert Mead ed., Reports of the Northeast Conference. Manasha, Wisconsin: George Banta, 50-58.

Furness, Edna L. 1953. Historical background of audition testing of Spanish. Modern Language Journal, 37:23-7.

Grittner, Frank M. 1977. Teaching foreign languages. New York:Harper and Row.

Harris, D.P. 1969. Testing English as a foreign language. New York: McGraw-Hill.

Ingram, Elizabeth. 1968. Attainment and diagnostic testing. In Davies, 70-97.

Jones, Randall, and Bernard Spolsky. 1975. Testing language proficiency. Arlington:Center for Applied Linguistics.

Lado, Robert. 1957. Linguistics across cultures. Ann Arbor: Univ of Michigan Press.

1960. English language testing: problems of validity and administration. English Language Teaching, 14:153-61.

1964. Language testing. New York:McGraw-Hill.

Myers, Charles T., and Richard S. Melton. 1964. A study of the relationship between scores on the MLA Proficiency Tests and ratings of teacher competence. Princeton:ETS.

Oller, John Jr. 1972. Cloze tests in English, Thai and Vietnamese; Native and Nonnative performance. Language Learning, 22, 1:1-15.

Oller John. 1976. A program for language testing research. Language Learning, special issue No. 4: 141-65.

Paquette, F. Andre, and Suzanne Tollinger. 1968a. A handbook on foreign language classroom testing. New York: M.L.A.

Paquette, F. Andre, and Suzanne Tollinger. 1968b. A handbook on the MLA Foreign Language Proficiency Tests for teachers and advanced students. New York: M.L.A.

Pimsleur, Paul. 1966. Testing foreign language learning. In Albert Valdman ed. Trends in language teaching. New York:McGraw-Hill , 175-214.

and C.N. Staubach. 1961. Teaching Spanish: a linguistic orientation. New York: Blaisdell.

Richards, Jack C. 1974. Error Analysis. London: Longmans.

Selinker, larry. 1974. Interlanguage. In Richards 31-54.

Smith, Phillip D. 1970. A comparison of the cognitive and audiolingual approaches to foreign language instruction; the Pennsylvania Foreign Language Project. Philadelphia: Center for Curriculum Development.

Spolsky, Bernard ed. 1978. Approaches to language testing. Arlington: Center for Applied Linguistics.

Stockwell, Robert, and J.D Bowen. 1965. The grammatical structures of English and Spanish. Chicago: University of Chicago Press.

Upshur, John. 1962. Language proficiency testing and the contrastive analysis dilemma. Language Learning, 12,2: 123-7.

Valette, Rebecca. 1967 & 1977. Modern language testing; a handbook. New York: Harcourt Brace Jovanovich.

Wardhaugh, Ronald. 1970. The contrastive analysis hypothesis. TESOL Quarterly, 4, 2: 123-30.

Whitman, Randal, and Kenneth Jackson. 1972. The unpredictability of contrastive analysis. Language Learning, 22,1: 29-41.

Wilson, Craig B. 1977. Can ESL cloze tests be contrastively biased ? Southern Illinois University, Occasional Papers in Linguistics.