Association of Recognized English Language Schools Oral Examinations
Australian Second Language Proficiency Ratings
Basic English Skills Test
Basic Inventory of Natural Language
Bilingual Syntax Measure I
Bilingual Syntax Measure II
Bilingual Vocational Oral Proficiency Test
Cambridge First Certificate in English
Cambridge Certificate of Proficiency in English
Cambridge Preliminary English Test
Comprehensive English Language Test
Delta Oral Placement Test
English Language Battery
English Language Skills Assessment
English Language Testing Service
English Proficiency Test Battery
Royal Society of Arts: Examinations in the Communicative Use of English as a Foreign Language
General Test of English Language Proficiency
Henderson-Moriarty ESL Placement Test
Idea Oral Language Proficiency Test
Idea Proficiency Test II
Ilyin Oral Interview
Interagency Language Roundtable Oral Proficiency Interview
John/Fred Test
Language Assessment Battery
Language Assessment Scales
Listening Comprehension Picture Test
Listening Comprehension Written Test
Maculaitis Assessment Program
Michigan Test of English Language Proficiency
Michigan Test of Aural Comprehension
The Oxford Examination in English as a Foreign Language
Oxford Placement Test
PRE-LAS
Quick Language Assessment Inventory
Second Language Oral Test of English
Selection Test
ndary Level English Proficiency Test
ture Tests-English Language
of Ability to Subordinate
Test in English (Overseas)
Test in English for Educational Purposes
Test of English as a Foreign Language
Test of English for International Communication

Reviews of English Language Proficiency Tests

Edited by
J. Charles Alderson
Karl J. Krahnke
and
Charles W. Stansfield

Reviews of English Language Proficiency Tests

J. Charles Alderson
Karl J. Krahnke
Charles W. Stansfield
Editors

Teachers of English to Speakers of Other Languages

Staff Editor: Julia Frank-McNeil
Editorial Assistants: Juana E. Hopkins
Christopher R. Byrne

Library of Congress Catalog No. 87-050894
ISBN 0-939791-31-5

Typeset in Linotype Caledonia by
Graftec Corporation, Washington, DC
and lithographed by
Pantagraph Printing, Bloomington, IL

Acknowledgements

The editors would like to express their gratitude to a number of people for making this publication possible. Most importantly, Donna Ilyin, who originally conceived the project. Without her energy and dedication the project could never have been undertaken. Early in the history of the project, a number of committees around the United States participated in choosing and evaluating tests. While there are too many who participated to mention here, their contributions were invaluable, even though their work may not appear in the final form of the publication. Maria Parish-Johnson and Sayuri Madusa assisted in correspondence and typing. Educational Testing Service provided support for the typing of the manuscript as well as substantial phone and postage expenses. James E. Alatis, a number of recent TESOL Presidents and Executive Boards, TESOL Publication Committee Chairs, H. Douglas Brown, and Julia Frank-McNeil have given unflagging support to the project.

Preface

This work provides descriptive and evaluative information on the major English as a second language and English as a foreign language (ESL/EFL) tests in current use throughout the world. Tests were selected for review on the basis of two criteria: (a) The test must be commercially available, and (b) the test must be relatively widely used. The first criterion is fairly straightforward. One exception is the Interagency Language Roundtable Oral Proficiency Interview (ILR), which is not actually a published test but a carefully defined and widely used procedure. Commercial availability does not apply to many British tests, as Charles Alderson notes in his introduction to British tests. In addition, the criterion of commerical availability excludes many tests that have been developed and used in local programs or school districts in the United States and only occasionally made available to others.

The second criterion was more difficult to apply since it was impossible to define "widely used." The editors gradually compiled a list of tests on the basis of informal discussions with a number of test users. The list was circulated among testing professionals, in particular those active in ESL/EFL testing, over a period of several years to determine whether tests should be added or deleted. The final result is the product of this process, and although several tests may have been added or omitted, the selection represents the major published tests available and in use at the time of publication of this collection.

A number of compromises were inevitable. The first is with the *number of tests included*. Far more ESL/EFL tests are in use than are reviewed in this volume. The selection criteria excluded all but a fraction of the tests actually in use in the world.

Second, the *quality and content of the reviews* is variable. Reviewing a test is not a precise procedure. While some features of a test are easily observable and confirmable, others are matters of opinion or depend on extensive experience. Reviewers have individual strengths, weaknesses, and interests, and their reviews cannot help but reflect them. To achieve the highest possible quality reviews, the editors have solicited qualified and objective reviewers, and have subjected the reviews themselves to further review.

A third concern is the *timeliness of the reviews*. This collection includes tests that were known to the editors at the time of the completion of the project. New tests or new editions of tests may have become available since that time, and it may be possible to correct this problem in a future edition of this collection. Because tests are frequently being revised and updated, users are urged to contact test publishers for the most recent information on the tests.

The Editing Process for this Volume

The reviews in this volume have been carefully prepared. Reviewers were cautioned to examine the test materials in depth. Many reviewers already had experience with the test. Others gained experience by administering it to one or more examinees on a trial basis, although they were not required to do so.

The editors generally divided their work as follows. Karl Krahnke took charge of obtaining the North American tests from the publishers. He selected competent reviewers and invited them to write reviews, sent out test materials, and received the completed reviews. Charles Alderson performed these tasks for the British and Australian tests. In addition, he critiqued and edited the reviews submitted to him. Charles Stansfield received the reviews from Krahnke and Alderson, edited them, and verified the information they contained, according to the procedures outlined next.

A unique aspect of this collection, compared with similar collections, such as the test reviews that appear in the *Mental Measurements Yearbook,* is that the reviewers' opinions or comments underwent independent evaluation. A review written by one author was sent to another author for comment. These comments resulted in challenges to statements in a number of cases. Similarly, each review, sometimes in revised form, was sent to the test author or publisher, who was invited to comment also. This produced many points of contention, with some of the publishers providing written replies longer than the reviews themselves. Much of the information supplied by test authors and publishers simply corrected minor inaccuracies in price or other information that is subject to frequent change. However, in other cases, the publishers expressed fundamental differences of opinion with reviewers. While a serious effort was made to resolve these disputes by examining test materials and other supporting documentation, it was not possible to do so to everyone's satisfaction in all cases. In such cases, Stansfield decided on the final version of the review, after considering both the reviewer's freedom to make critical statements about a test and the publisher's desire to have accurate, appropriate statements about the test published. While this process resulted in more balanced and more accurate reviews, the reader should keep in mind that different reviewers could reach different conclusions about the validity of a test. Thus, while the reviews may be helpful, they do not necessarily reflect the conclusions about a test's validity that any user would draw after examining a set of test materials or administering a particular instrument to a group of students.

Organization of the Volume

This volume is organized into separate reviews of the major tests of ESL and EFL that are currently used in the United States, Canada, the United Kingdom, and Australia—the major population centers of the English-speaking world. The tests are of two kinds: off-the-shelf and secure. Off-the-shelf tests are available for purchase from publishers by a teacher. Such tests are used on multiple occasions and whenever needed. Secure tests are returned to the publisher after use. Instead of purchasing the test, the institution or the examinee may pay a fee for the right to administer it. The test reviews are arranged in alphabetical order.

Each review begins with a Synopsis that presents basic information on the test in summary format. The reader should read the Synopsis first, since the reviewers were encouraged not to repeat in their reviews information that was included in the Synopsis. The information is presented in telegraphic form in the following order:

Complete title
Acronym, or commonly used short form of title
Date of publication
Intended examinee population
Intended purpose of test
Scoring method (by parts or sections of test)
Type of administration (individual or group)
Length of test
Test components (i.e., date of publication and number of pages)
Cost of components (1985-1986)
Author(s)
Publisher, including complete address and phone number

Although the information contained in the Synopsis has been verified by the publisher or by comparing it with test publications, certain types of information (i.e., product availability, components, price, and publisher's address and phone number) change frequently. The reader will want to take this into account before placing an order for a test based on information obtained from this volume.

Following the Synopsis, some reviews contain one or more entries in a Test References section. These are references added by Stansfield and are not mentioned by the reviewer, nor are they included in the technical or administrative manual for the test. Rather, they are intended to provide more comprehensive information about the test and often refer to other published reviews or studies of the test that the reader may consult. The fact that less than half of the reviews contain a Test References section is indicative of the paucity of independent analyses of language proficiency tests.

A description of the test and a discussion of its reliability, validity, and related issues are contained in the Review. Each Review is from 900 to 2,500 words in length; the average length is approximately 1,500 words.

Following the Review are the Reviewer's References. As the name implies, these are references cited by the reviewer in the body of the Review. The reader is encouraged to consult both the Test References and Reviewer's References sections for additional information relevant to each test.

How to Use this Volume

This volume may be used in several ways. One is to provide information to test users about tests they may already be using, and the second is to provide test users with information to assist them in choosing tests to use for their specific testing needs. Of course, a collection of reviews such as this may be put to a number of other uses. For example, the work provides feedback to test writers and publishers, informs test users about the qualities and criteria they should consider in choosing and using tests, and informs the consumers of test results about the quality and significance of the information they are being provided by tests. In general, the editors hope that the reviews contained in this volume will lead, at least in some small way, to an improvement in both the tests available for ESL/EFL assessment, and the uses made of such tests and their results.

Regarding the primary purposes for using this volume, however, actual or prospective test users should first carefully characterize the type of test takers they will be testing, their purposes for testing them, and the use that will be made of the results of the testing. Only then can they consult specific reviews to determine if a specific (or any) test is appropriate to their needs.

When consulting specific reviews, the test user should first refer to the information on intended examinee population to determine if the stated population for which the test is designed is identical or similar to the population the user wishes to test. The user should also refer to the information on intended purpose of test to determine if the application suggested by the publisher is similar or identical to the user's purpose for testing. (The usual purposes for testing are reviewed later in this Introduction for those who are not familiar with them.)

After determining whether the test is appropriate to the user's purpose and students, the body of the review may be examined to verify whether the intended purpose and population for the test are legitimate. This information may be found in comments about the population on which the test was normed and on the statistics generated during the norming process. In general, the quality of a test can only be determined for the type of test takers on which the test was normed. Tests should be applied with caution to other populations. Test users who cannot interpret the technical information relating to a test are encouraged to consult a specialist in testing (usually available in any school system, college, or university) for assistance.

Next, the test user should consider the practical aspects of the test: administration time and type, train-

ing requirements of administrators, cost, and so on. Close attention should be paid to any experience in actually using the test the reviewer reports. Such experience can often uncover strengths or weaknesses in a test that strictly objective information may not reveal.

Test users are encouraged to correspond with publishers to obtain additional information about a test or to inform the publisher of problems in administering a test or interpreting its results. Without constructive feedback, publishers may be unable to improve their products. Test users should also share experiences and discuss problems with each other.

Finally, potential users should examine a copy of a test before making a commitment to use it. Special attention should be given to the accompanying documentation. A good test should have extensive supporting documentation, such as a user's manual or a technical manual. While there may be good tests that lack such documentation, it is difficult to determine their quality or how best to interpret their results.

Uses and Misuses of Testing

Testing plays a major and sometimes dominant role in language teaching. Most second or foreign language teaching involves some sort of test or examination. Tests are used at the beginning of instruction to determine readiness, during instruction to determine student progress, and at the end of instruction to determine its effectiveness. Tests may be used for research, diagnosis, or even for practice in taking tests. Tests are even rightly or wrongly used as the basis for curriculum, such as when instruction is keyed to preparing students to perform well on a specific test.

Although tests are a universal ingredient in ESL/EFL teaching, they are often poorly understood, misused, or misapplied. One of the most frequently asked questions of second language teaching professionals is "What test do you use for . . . ?" or "What is the best test for . . . ?" Many ESL/EFL teachers seem to believe that one or more tests must exist that will solve all their administrative and instructional problems, but that they simply have not heard about them. This sentiment reflects the difficulty teachers and administrators have in finding, understanding, and evaluating information on ESL/EFL tests. It also reflects the understandable lack of training and experience that many language teaching professionals have in the field of testing and test use.

The complexities of English language testing and the limited expertise of many users of language tests have lead to the misuse (and occasional abuse) of tests, especially of commercially available tests. Out of a well-meaning desire to solve their testing and teaching problems as efficiently as possible, test users can make the mistake of employing the wrong test for their students or purpose, misinterpreting the results, or misapplying the results.

An example of the first kind of mistake is using a test designed to measure students' success at learning a specific instructional set of materials (achievement) to measure students' overall ability in English (proficiency). Another example is using a test designed for and normed on children learning English as a second language to measure the success of adults in a basic education program. In both cases, using the test will provide the testers with scores, but the utility of the scores will be low because the tests are not valid for the population being tested. Another example of misuse is to apply a test designed for native speakers of English to nonnative speakers.

An example of misinterpretation or misapplication of test results is making a decision about a student's progress or readiness for advancement (e.g., mainstreaming in a school system) on the basis of a small difference in scores on a single English language test. This type of misapplication is especially serious when the test does not discriminate between more and less proficient students, when it measures only a narrow range of language behaviors (e.g., only sound-letter correspondence or only reading), or when performance on the test is generalized to broader aspects of language behavior. Another example of misapplication of test results might be the exclusive use of English language test scores to predict success or failure in some other effort, or in school work in general.

The misuse of language tests is minimized when the test user has two kinds of knowledge. One is a basic knowledge of testing principles, knowledge that is available in a number of publications (see Appendix), from teacher education courses and from workshops. The second kind of knowledge concerns the features and quality of the tests that are available.

This volume addresses a knowledge gap in ESL/EFL testing by providing basic descriptive and evaluative information on many commercially available tests. By having this information available in a consistent format, the informed test user can choose appropriate tests if they exist. This volume cannot, of course, substitute for a course on language testing. The test user is encouraged to obtain that knowledge elsewhere or to consult a language testing specialist when choosing a test for a specific testing situation.

Purposes of Testing

To assist the user of this collection in choosing appropriate tests, a brief introduction to the most common purposes for testing is provided here. Although space permits only a summary of the purposes, teachers and nonspecialists should find this information useful.

Purposes for language testing fall into four major types: placement, measuring achievement, diagnosis, and measuring proficiency.

Placement

A test is used for placement when the results determine the level of instruction for which a student is ready. A good placement test should: (a) contain the

same types of knowledge or skills that are taught in the instruction program in which the student is being placed, and (b) not include tasks that are so unusual or unfamiliar to the test takers that they may negatively affect the students' performance.

Measuring Achievement

An achievement test measures a student's success in learning some specific instructional content and is given after the instruction has taken place. A good achievement test should contain only material that was actually taught. Thus, commercially available tests do not serve as achievement tests, except those that have been prepared to accompany some specific instructional material. None of the tests reviewed in this collection do. Achievement tests are normally prepared by the staff of the instructional program in which they are used.

Diagnosis

A diagnostic test measures specific aspects of second language ability, usually for the purpose of determining what the test taker knows and needs to learn. Few truly diagnostic language tests have been published, and none are included in this collection. Tests that measure more broadly defined aspects of second language behavior (e.g., listening comprehension vs. reading ability) may claim to be diagnostic, but they are really tests of relative proficiency in different communicative skills.

Measuring Proficiency

Proficiency tests measure the test taker's overall ability in English along a broad scale. Proficiency is usually defined independently of any instructional program: Proficiency is not easily taught since it is a global construct. Proficiency tests may help determine whether the test taker is ready for a job or task requiring English (e.g., working as a government official or entering higher or secondary education), or they may be used to compare the overall success of different instructional programs. Proficiency tests are sometimes subdivided into subskills or modes of language, including speaking, listening, reading, writing, vocabulary, grammar, and sociolinguistic, strategic, and discourse competence, among others. Relative ability in these areas can be determined by a proficiency test.

A number of the tests included in this collection are intended to be proficiency tests. It should be noted, however, that proficiency tests are often poor tests of achievement, since the content of a proficiency test usually has little or no relationship to the content of an instructional program. Proficiency tests may not be appropriate for placement, since they may provide an overall stratification of students, but do not specify their abilities according to the specific instructional content of the courses in which they are placed. The global nature of the construct of proficiency also makes many such tests poor diagnostic instruments because the results do not specify the knowledge a student has or is lacking. Test users should also avoid using proficiency tests to make strong predictions about a test taker's eventual success in some other endeavor (e.g., an academic program), since language is but one element among many that contribute to success.

A Final Word

No single publication can answer all the questions that test users have about testing, any more than a single test can serve all testing needs. This collection is a tool that can assist test users in choosing tests that are appropriate to their needs and in evaluating the quality of those tests. Nevertheless, for the testing process to function well, it must be carried out by people with expertise in using a particular test. We hope that this volume will contribute to increased awareness on the part of test users, test writers, and test publishers of the need for quality in ESL/EFL testing. If this goal is realized, the ESL/EFL teaching profession will benefit from an improvement in one of the most important tools used in the ESL/EFL educational arena.

The Editors

Appendix

Following is a selected list of reference books on ESL/EFL and second language testing.

Allen, J. B. P., & Davies, A. (1977). *Testing and experimental methods: Vol. 4.* London: Oxford.

Carroll, B. J. (1980). *Testing communicative performance.* Oxford: Pergamon.

Carroll, B. J., & Hall, P. J. (1985). *Make your own language tests: A practical guide to writing language performance tests.* Oxford: Pergamon.

Cohen, A. D. (1980). *Testing language ability in the classroom.* Rowley, MA: Newbury.

Harris, D. P. (1969). *Testing English as a second language.* New York: McGraw-Hill.

Heaton, J. B. (1975). *Writing English language tests.* London: Longman.

Jones, R. L., & Spolsky, B. (1975). *Testing language proficiency.* Arlington, VA.: Center for Applied Linguistics.

Maculaitis, J. D. (1982). *The MAC checklist for evaluating, preparing, and/or improving standardized tests.* San Francisco: Alemany

Madsen, H. S. (1983). *Techniques in testing.* Oxford: Oxford.

Oller, J. W., Jr. (1979). *Language tests at school.* London: Longman.

Valette, R. M. (1977). *Modern language testing* (2nd ed.). New York: Harcourt Brace Jovanovitch.

Table of Contents

A Brief Introduction to ESL Proficiency Testing in North America 1
An Overview of ESL/EFL Testing in Britain 3
Association of Recognised English Language Schools Oral Examinations 5
Australian Second Language Proficiency Ratings 7
Basic English Skills Test 9
Basic Inventory of Natural Language 10
Bilingual Syntax Measure I 12
Bilingual Syntax Measure II 14
Bilingual Vocational Oral Proficiency Test 17
Cambridge First Certificate in English 18
Cambridge Certificate of Proficiency in English 20
Cambridge Preliminary English Test 21
Comprehensive English Language Test 22
Delta Oral Placement Test 24
English Language Battery 25
English Language Skills Assessment 27
English Language Testing Service 28
English Proficiency Test Battery 31
Royal Society of Arts: Examinations in the Communicative Use of English as a Foreign Language 32
General Test of English Language Proficiency 34
Henderson-Moriarty ESL Placement Test 35
Idea Oral Language Proficiency Test 37
Idea Proficiency Test II 39
Ilyin Oral Interview 41
Interagency Language Roundtable Oral Proficiency Interview 43
John/Fred Test 47
Language Assessment Battery 49
Language Assessment Scales 51
Listening Comprehension Picture Test 53
Listening Comprehension Written Test 55
Maculaitis Assessment Program 57
Michigan Test of English Language Proficiency 58
Michigan Test of Aural Comprehension 60
The Oxford Examinations in English as a Foreign Language 61
Oxford Placement Test 62
PRE-LAS 64
Quick Language Assessment Inventory 65
The Second Language Oral Test of English 67
Secondary Level English Proficiency Test 68
Short Selection Test 70
Structure Tests - English Language 72
Test of Ability to Subordinate 73
Test in English (Overseas) 76
Test in English for Educational Purposes 77
Test of English as a Foreign Language 79
Test of English for International Communication 81
Test of English Proficiency Level 83
Test of Spoken English 84
Test of Written English 86

A Brief Introduction to ESL Proficiency Testing in North America

Commercially published ESL proficiency tests in the United States may be characterized, when comparing them with tests in the United Kingdom, by the fact that most are designed for elementary and secondary school children rather than adults entering higher education institutions. This situation is basically a response to a U.S. Supreme Court decision (*Lau vs. Nichols*, 1974) in which the court ruled that publicly funded schools must provide special instructional programs for non-English-speaking (NES) and limited-English-speaking (LES) students. (The second group is also sometimes called limited English proficient [LEP].)

As a result, in May 1975 the U.S. Office of Civil Rights of the Department of Education issued a set of guidelines for local school districts enabling them to comply with the decision. These guidelines called for the assessment of language dominance (a comparison of proficiency in two or more languages) of all students with linguistic backgrounds involving a language other than English. One programmatic outcome of the guidelines was the widespread implementation of bilingual education programs in the U.S. Prior to the Lau guidelines, bilingual education was common only in church-operated school settings, in cities (such as Miami) containing a large number of refugees, and on American Indian reservations.

As a result of the requirement that proficiency be assessed, dozens of ESL proficiency tests for school children were developed. Some of these tests included parallel versions in one or more other languages. In most of these cases a Spanish language version was developed and published first, while unpublished versions in other languages were developed by researchers or educators. While a number of the early tests were naively constructed and are out of print today, others have survived critical evaluation and enjoy widespread use. New tests for children at the kindergarten through 12th grade level (K-12) continue to be developed and published commercially. Approximately half of all tests reviewed in this volume fall into this group. Typically, at the kindergarten and early elementary school levels (K-2), these tests assess speaking proficiency only (e.g., the PRE-LAS). Tests used with children in subsequent grades (3-12) may exhibit greater variety in the skills assessed, with some (e.g., the LAB) assessing receptive as well as productive skills and others assessing receptive skills only (e.g., the SLEP).

At the higher (tertiary) education level in the U.S. and Canada, students may take one or more of three common instruments: the TOEFL, one of The University of Michigan tests (e.g., MELAB), or CELT. The TOEFL, a secure test, is accepted as evidence of English proficiency by 2,500 universities in the U.S., Canada, and other countries. Similarly, the MELAB (another secure test) is also accepted as evidence of English proficiency by many universities. Both of the these testing programs make available previously used versions of the test or sections of it to English language teaching programs that wish to use the test in order to determine initial placement within their instructional sequence. In some cases however, scores on these less secure, institutional forms of the test may be accepted by a university admissions officer. A third alternative, the CELT, is a nonsecure instrument that is sold to educators and institutions for ESL program placement. These tests have traditionally involved the assessment of listening, reading, vocabulary, and grammar. More recently, the TOEFL moved toward the inclusion of a direct measure of writing, while the MELAB has included one for some time.

Another test of academic English that might be included in this group is the ALIGU, published by the American Language Institute of Georgetown University. It is given only to applicants for scholarships awarded by the Agency for International Development of the U.S. Department of State. At the request of the program, the ALIGU is not reviewed in this volume.

A number of tests of nonacademic English have been developed for adults with limited educational backgrounds in North America. The development of many of these was spurred by the influx of Asian refugees to the the U.S. after the Vietnam War. Nonetheless, it would be inappropriate to assume that these tests are suitable for Asians only, as this was frequently not the authors' intent. Among these tests of nonacademic English for adults are the HELP, the John and Fred tests, the BEST, DOPT, and so forth.

A number of other tests of general English for adults have appeared recently for use in academic situations that are less demanding than traditional universities. These include the tests developed by Donna Ilyin and her colleagues in the San Francisco Community College District.

As a parallel development, the Interagency Language Roundtable (ILR) procedures produced by agencies of the U.S. government have grown increasingly popular and have influenced the development of other tests in the U.S. and outside the U.S. Many of the oral language tests used in North America today exhibit some characteristics of the ILR scale and procedures. The ILR's influence is also felt in the testing of foreign languages in the U.S. The American Council on the Teaching of Foreign Languages (ACTFL) has taught the ILR procedures to over 1,000 foreign language

teachers and has added three additional points to the scale at the lower levels.

Sources of Information on North American ESL Tests

Readers of this volume are encouraged to consult other sources of information on ESL tests. In the North American context, a number of publications publish reviews or studies of ESL tests. The following brief annotated bibliography is provided to assist the reader in this endeavor.

Mitchell, J. V., Jr. (Ed.). (1985). *The ninth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements & the University of Nebraska.

The *Mental Measurements Yearbook* (MMY) is generally considered the most authoritative source of reviews availabie anywhere. Unfortunately, it does not ordinarily publish reviews of tests outside of North America. Nine MMYs have been published since 1936. Many of the tests reviewed in this volume are also reviewed in the seventh, eighth, and ninth MMYs. A tenth MMY is scheduled to be published in 1990. Although the MMY reviews sometimes assume special training in tests and measurements, the series remains a principal reference on tests of all kinds.

Dieterich, T., & Freeman, C. (1979). *A linguistic guide to English proficiency testing in schools*. Washington, DC: Center for Applied Linguistics.

This volume contains short reviews and information on many of the tests included in this volume. It also contains a detailed discussion of different approaches to language testing that are frequently exhibited by ESL tests.

Stansfield, C. W. (1981). The assessment of language proficiency in bilingual children: An analysis of theories and instrumentation. In R. V. Padilla (Ed.), *Bilingual education technology: Vol. 3. Ethnoperspectives in bilingual education research series*. Ypsilanti, MI: Eastern Michigan University.

This article is a cogent introduction to the field of second language proficiency test construction. After reviewing the choices available to a test developer, the author illustrates how the various approaches are reflected in an instrument. A number of ESL tests are reviewed briefly, with a focus on the contrasting approaches they use.

Erickson, J. G., & Omark, D. R. (Eds.). (1981). *Communicative assessment of the bilingual/bicultural child*. Baltimore, MD: University Park.

This volume contains a number of articles on second language tests for children at the K-12 levels. Chapters 7 and 8 include short reviews or analyses of a number of tests, that is, many of the K-12 North American tests reviewed in this volume.

The Modern Language Journal (MLJ). The 1976-1980 volumes of this well-established professional journal contain reviews of language proficiency tests that were available at that time, including some that were not reviewed in this volume because they are currently out of print. *The MLJ* is widely available in academic libraries. Information on subscriptions may be obtained by writing: Journals Divisions, University of Wisconsin Press, 114 North Murray Street, Madison, WI 53715, USA.

In addition to the above sources of information specifically on North American tests, the following publications provide further news and information on tests in the U.S., England, and elsewhere.

Language Testing. This new professional journal promises to be a useful source of reviews of tests from around the world. Information on subscriptions may be obtained by writing: Edward Arnold Publishers, 41 Bedford Square, London WC1B 3DQ, England. In the U.S. and Canada, subscription information may be obtained from the North American distributor, Cambridge University Press, 32 East 57th Street, New York, NY 10022, USA.

Language Testing Update. This newsletter provides information on new research and developments in the language testing world and occasional reviews of second language tests. Information on subscriptions may be obtained by writing: Institute for English Language Education, University of Lancaster, Lancaster, LA1 4YT, England.

Charles Stansfield

An Overview of ESL/EFL Testing in Britain

In many respects there are considerable differences in the way British and American tests are produced and validated. In addition, there are also differences in the way tests on the two sides of the Atlantic are made available to the teaching profession and the general public. It seems useful, then, to present a brief overview of British testing as background for the British tests reviewed in this collection. We emphasize that this overview cannot be a full account of British testing practices. Omitted, for example, are the very interesting developments in the graded objectives movement in modern languages in the United Kingdom known as Graded Tests. Instead, we confine ourselves to issues relevant to an understanding of those British tests reviewed in this volume. In addition, it should be noted that very interesting changes are about to take place in British school-leaving examinations (to be known as the General Certificate of Secondary Education [GCSE]) which may eventually have considerable impact on the way ESL/EFL tests are produced, administered, and validated in the U.K.

Most of the British tests reviewed in this collection are produced by examination boards, whose main activities relate to the development and administration of secondary school examinations given at ages 16 and 18, approximately. These examinations are known as O (Ordinary) and A (Advanced) Level examinations. These examination boards also produce ESL/EFL examinations to determine ESL/EFL proficiency and readiness for entry to British universities. The examination boards are usually associated with one or more universities, for example, the University of Cambridge Local Examinations Syndicate, the Oxford Delegacy of Local Examinations, the Joint Matriculation Board (of the northern universities of Manchester, Liverpool, Leeds, Sheffield, and Birmingham). There are also other examining bodies roughly similar to the O and A Level Boards: the regionally organized Certificate of Secondary Education Boards; and vocationally oriented boards, such as the Royal Society of Arts, which has recently produced the highly influential Examination in the Communicative Use of English as a Foreign Language. This overview will concentrate on the latter.

Perhaps the most important point to note is that the examinations produced by these exam boards are not standardized or normed in the usual sense, but are produced and administered for one occasion only. Thus, tests (often known as *papers*) produced for administration in Spring 1987 will never be used again. (There are exceptions to this general rule, but they do not affect the principle.) Indeed, past papers are often made publicly available, sometimes for a small fee.

Due to the constant need to produce new examinations and the lack of emphasis by exam boards on the need for empirical rather than judgmental validation, these examinations are rarely, if ever, tried out on pupils or subjected to the statistical analyses of typical test production procedures. Examination boards do not see the need to pretest and validate their instruments, nor conduct posthoc analyses of their tests' performance. Although the objective items in the tests are usually pretested, the statistics are rarely published. There are significant exceptions to this statement: the joint UCLES/British Council's ELTS test and the Associated Examining Board's Test of English for Educational Purposes (see reviews in this volume). It is hoped that publications such as this might induce examination boards to conform more closely to accepted test production practices. This is not to say that the tests produced are not valid and reliable, but that we have very little empirical evidence of their characteristics.

Rather, the exam boards lay great store by the asserted validity of their examination construction procedures, which rely almost exclusively upon "expert" judgments. The production of a test for any occasion is the responsibility of a chief examiner, selected by the board for "proven" qualities of judgment and track record of reliability of marking the production of sample test questions in past years. This chief examiner will also have recent, if not current, experience teaching the subject for which he or she is producing a test. The chief examiner is aided by a set of assistant examiners and a moderating committee, who produce, scrutinize, edit, and finalize the tests (a process known as *moderation*). In addition, the chief examiner produces marking criteria (sometimes known as *mark schemes*) and is responsible, with senior examiners, for the training and standardizing of markers, and the checking of interrater reliability after the examination has been administered. Even in this process of the checking of interrater reliability, it is extremely unusual for an exam board to calculate or publish statistics of reliability of its markers. It would be a simple matter for the board to calculate and provide the data, and we believe they should be encouraged to do so. The exam boards place great faith in the qualities of their chief examiners and in their selection, moderation, standardization, and grade-awarding procedures. They should, however, be prepared to produce the evidence that their examinations are valid and reliable. (As Gary Buck put it: "It is, after all, normal practice to count one's change in the grocery store, even if the cashier looks honest.")

Finally, it should be noted that most, although not all, of the British tests in this collection are not avail-

able commercially for administration at any point in time. The normal procedure is for a school to register its pupils with a particular board for a particular exam on the published date. In EFL, it is usually possible for individual students to enter for examinations on prespecified occasions. There are, however, exceptions to this practice, and these are noted in the relevant synopses of the test reviews.

In conclusion, although we have been critical of British tests, we should emphasize that many of these tests are highly innovative in content and format, and they should not be dismissed lightly by the test-producing or test-using fraternity. Indeed, we believe that other tests could benefit greatly, both from greater attention to the actual examples of tests produced by some exam boards and by attention to the content validation procedures they use. It is also true to say that many tests would benefit from greater attention to the relationship between testing and teaching, for which the British Exam Boards are particularly noted. Some combination of British judgmental validation and American empirical validation seems required.

Charles Alderson

Association of Recognised Language Schools Oral Examinations in Spoken English

Reviewed by

Terry Tony
British Council
London

Synopsis

Association of Recognised Language Schools Oral Examinations in Spoken English. ARELS Oral Examinations. 1967-84. All foreign students of English at three levels: the Preliminary Certificate (AP) for any age, the Higher Certificate (AH) for over 16 years, and the Diploma (AD) for over 18 years. Designed to measure skills in the use and comprehension of spoken English, especially in everyday, realistic situations. Taped verbal responses scored, by appointed examiners only, on scales which vary according to specified criteria/performance descriptions; total score translated to one of three pass grades or two fail grades for AP, and one of six pass grades or three fail grades for AH and AD. Individual and group according to number of tape recorders available. 40 minutes (AP); 45 minutes (AH & AD). New papers compiled for each series of examinations: three AP and AH, and two AD scheduled examinations per year; other than these, Opportunity Examinations are available. *A Guide for Examination Centres* (booklet, 1978, 18 pp.); *Regulations and Outline Syllabuses* (brochure, 1984, 5 pp.); *Rationale, Development and Methods* (booklet, 1983, 12 pp.). Examinations available to recognized centers only; charge per candidate made up of a basic fee (AP, £12; AH, £17; AD, £21) and a local fee (between £3 and £15 depending on center and number of candidates). Past examinations are available in sets (1 master cassette, 5 keys, 20 candidate papers) at £13 for AP, and £10.50 for AH and AD; or individually at £6 for 1 master cassette, £0.35 for each key, and £0.35 (AP) and £0.18 (AH & AD) for each candidate paper. Reel tapes are available at additional cost; booklets and pamphlets listed above are free of charge; postage and overseas bank charges extra. ARELS Examination Trust, 113 Banbury Road, Oxford OX2 6JX, England, telephone: (0) 865-514272.

Review

The ARELS Oral Examinations are designed to test a candidate's oral production and listening comprehension in everyday English communication. The focus on these two skills is an attempt to counterbalance the emphasis on reading and writing skills found in most other EFL examinations. A consequence of this focus is the manner in which the examinations are conducted. The instructions, text input, questions and candidate responses are all recorded on tape. This means that the number of candidates able to take the examination at any one time is restricted only by the facilities (number of booths in a language laboratory or number of tape recorders) available. The candidates are only required to have minimal reading and writing skills, as the exam is comprised of visual stimuli, texts for reading aloud, grids for marking answers, and spaces for filling in short written answers, with variations according to level.

The three levels of examinations in order of increasing difficulty are: the ARELS Preliminary Certificate (AP), the ARELS Higher Certificate (AH), and the ARELS Diploma (AD).

The AP is meant to show that a candidate has "sufficient skill to survive in an English-speaking environment." The content is based on the Waystage proposals of the Council of Europe and usually takes the form of an extended role play on one central theme (e.g., a wedding, a visit to a holiday camp, etc.) in which the candidate participates from time to time. There are 15 questions testing three areas of competence.

The first of these is *Social English*. Social English is examined in tasks which require the candidate to read (aloud) and write numbers, letters, and common abbreviations. Candidates must also take the role of one participant in a short conversation, give appropriate responses to situations described on the tape, and finally ask questions. The second area of competence tested is *Audial Comprehension*. The word *audial* is used in place of *aural* because of the possible confusion of the latter with the word *oral*. Tasks require candidates to make a short written response (tick, cross, one word, etc.) or select an appropriate visual response. They may also have to say whether there is any connection between what they hear and what they see; and finally, they must give short verbal answers to questions on the tape. The third element in the test is *Extended Speaking*. Here the candidates must speak for 45-60 seconds. Pictures are normally provided in the candidate's paper as a stimulus for the talk. As can be gathered from this description, visual stimuli are used extensively in the AO.

The other two levels of examination, the AH and AD, have much in common with each other in both content and format. They both examine six areas of competence and are divided into six corresponding sections.

The first section examines *social responses* and is divided into three parts with 20 items in all. In Part 1 the candidate must give appropriate replies to a series of unconnected questions or comments. In Part 2 a situation is described on the tape, and the candidate takes the second part in a conversation on the basis of the information given. In Part 3 the candidate must make a natural comment after listening to a short description of a situation on the tape. In all parts there is a time limit of about 8-10 seconds for responses.

Section 2 in both examinations is devoted to *intelligible speech* and focuses on a candidate's intonation, stress, rhythm, and pronunciation. The task requires

the candidate to read a part in a conversation written in the candidate's paper. The other part(s) is recorded on the tape.

Section 3 examines *audial comprehension* (see above). This is done through a variety of tasks requiring candidates to answer questions on spoken texts (interviews, reports, descriptions) by explaining what has happened in a picture, to explain the meaning of words or phrases, to complete sentences started by speakers on the tape, or to mark in the candidates' paper the meaning corresponding most closely to a statement on the tape. This often depends on interpreting intonation or context.

Section 4 requires candidates to produce samples of *sustained speaking*. Stimulus is provided: in AH, a picture story for the candidates to tell after being given the background on tape; in AD, a spoken text (e.g., radio interview) is heard twice, the candidates may make notes, and then retell what they have heard adding comments as they wish.

Oral accuracy forms the focus of Section 5. The items are in the form of grammar drills in which the candidates must transform, complete with the correct grammatical form, or provide short answer forms to sentences they hear on the tape.

Section 6 in both examinations gives an opportunity for *free oral expression*. Candidates have 1 minute in AH and 1 1/2 minutes in AD to talk on a topic chosen from a list given to them 20 minutes before the examination. They are not allowed to read from notes.

While AP is clearly different from the other levels in both rationale and content, the distinction between AH and AD is not as clear. The authors write that both AH and AD test the above six skills, but AD does this "to a higher standard." A further difference is that AD tests "the use of English as a medium of abstract and intellectual thought."

After each series of examinations a key is produced for each level. This includes a full tape script and a marking guide, which gives sample answers where possible. For those questions requiring extended speaking, the key includes a list of the criteria used to assess the speech.

Background information on the examinations is provided in two booklets and a brochure. The first booklet, *A Guide for Examination Centres*, gives details of the facilities (language laboratory, tape recorders, etc.) needed to administer the examinations and instructions on how to run a center and deal with administrative aspects of the examinations. The *Guide* is mainly of interest to potential centers. The second booklet, *Rationale, Development and Methods*, provides a brief history of the examinations, their overall aim, an outline of the skills tested at each level, and a brief guide to marking procedures. "Regulations and Outline Syllabuses" covers the dates of examinations, the fee for candidates, a list of centers, and a brief outline of the aim of the examinations and the content at each level. A list of publications (including past papers) is also available. At present the ARELS Examination Trust (AET) does not publish a report on the examinations, and no statistics on candidate performance are issued.

Comments

One of the characteristic features of the ARELS Orals, namely, that they are conducted wholly on tape, gives rise to the difficult question of whether such an interaction can possibly be a natural communication situation. When compared with the inequality of the usual oral interview used for this kind of assessment, the issue appears largely academic. There are several advantages to having a sample of the candidate's performance on tape. It allows borderline cases and random samples to be re-marked, and in this way can foster greater reliability and fairness. The tapes can be sent to a limited number of trained assessors, who can facilitate greater agreement on standards, assuming, of course, that they are consistent in the criteria they use to assess the tapes. This is not easy to do in some parts of the examinations. The difficulty of establishing clear criteria is particularly apparent in the Social English/ Responses sections of the papers. The variety of possible responses to the decontextualized statements made in these sections is so wide, taking into consideration the attitudes and personalities of the candidates, that it is difficult to see what criteria might be relevant. This is least problematic in AP where an overall context and role are established for the candidates. This problem of decontextualization emphasizes how difficult it is to test communication skills. Indeed, in Sections 2 (Intelligible Speech) and 5 (Oral Accuracy) of AH and AD, purposeful communication is abandoned altogether. The emphasis is on the skills of verbal articulation and structural manipulation.

No matter which skills are tested, one of the most useful aspects of an examination is the feedback it provides to teachers (and candidates) on what has been taught well and what needs further work. The results of individual candidates, while being of interest to those individuals, are not very helpful in influencing course design. The most useful feedback is a report on overall performance in the examination. Such a report from the AET could make clear to teachers the deficiencies of the candidates (and the teaching) and in this way influence future teaching objectives. In short, it would be one of the best ways for the AET to achieve their stated objective of improving "the focus and quality of language teaching in general." A further element in this campaign would be to provide statistical data showing the tests to be well designed and relatively good measures of the language competence they set out to assess. This sort of information is necessary to assist in the development of the examinations and to give test users confidence in the validity and reliability of the examinations.

A number of issues have been raised here concerning the deficiencies of the ARELS Orals, but it must be remembered that they are also innovative and imaginative examinations offering a focus and a methodology found in no other EFL examinations. This flexibility

places them high on the list of measures of the oral and listening skills.

Reviewer's Notes

The AET has recently established a link with the Oxford Delegacy, which produces examinations of reading and writing with similar communicative aims. Between them, their examinations cover all four English communication skills.

Australian Second Language Proficiency Ratings

Reviewed by

T. J. Quinn and T. F. McNamara
University of Melbourne

Synopsis

Australian Second Language Proficiency Ratings. ASLPR. Adolescent and adult learners or speakers of a second or foreign language. Specific versions related to ESL, French, Italian, and Japanese are available (with versions in Chinese and Spanish in preparation), but the basic instrument (the version using ESL examples) is applicable to any language. Designed to measure general proficiency in a language learned as a second or foreign language. Learners are rated on a descriptive (criterion-referenced) scale in which the proficiency levels are specified by a number, title, and behavioral description. Each macroskill is rated separately, for example, S:l; L:1+; R:l; W:l-. Individual administration for S, L, R, but W can be done in group. Interview for S, L, R, and W may vary from 10 to 25 minutes depending on interviewee and skill of interviewer. Administration Manual in preparation. This Manual will include sample reading texts, writing scripts, an introductory paper on the ASLPR, and a copy of the scale. A set of nine videos and accompanying kits has been produced by Film Australia, the Australian Department of Immigration and Ethnic Affairs (DIEA), and the Australian Government Publishing Service (AGPS). Technical report: *Report on the Formal Trialling of the Australian Second Language Proficiency Ratings* (ASLPR) (1984, 136 pp.) published by AGPS and DIEA; information booklet, *Australian Second Language Proficiency Ratings* (1984, 58 pp.), available from AGPS and DIEA through Australian government bookshops, Australian embassies, and Australian trade commissions. Mimeo copy of ASLPR available from authors. Videos available from Film Australia and Australian embassies, trade commissions, and Film Australia outlets. Manuals, $2.50 (Australian) each; videos, $12.50 (Australian) full set; individual videos available; sample interviews, $65 (Australian) per tape; mimeo copies of ASLPR, free of charge from authors. D. E. Ingram, Brisbane College of Advanced Education, Mt. Gravatt Campus, PO Box 82, Mount Gravatt 4122, Australia, telephone: (07) 343 0611. Elaine Wylie, Department of Education, Brisbane, Qld 4000, Australia. Australian Government Publishing Service and Australian Department of Immigration and Ethnic Affairs, PO Box 25, Belconnen, A.C.T. 2617, Australia, telephone: (062) 64 1111.

Review

The ASLPR is a structured interview procedure and rating scale designed to provide a measure of language proficiency in the four basic communicative skills. According to the Manual (p.9), it is based on the *absolute proficiency ratings* of the U.S. Foreign Service Institute, otherwise known as the ILR scale (see this volume for review of the ILR). However, the ASLPR differs from the ILR; while the latter is designed for use with well-educated, civil and foreign service personnel, the ASLPR is designed for use with learners whose education and employment is more diverse. The ASLPR scale also permits greater differentiation at the lower end of the scale than does the ILR.

The ASLPR scale has gained wide acceptance in Australia, particularly in the federally funded Adult Migrant Education Program, the dominant area of adult ESL in Australia.

The scale consists essentially of descriptions of language behavior at nine levels ranging from O to native-like proficiency. The nine levels are: 0, 0+, 1-, 1, 1+, 2, 3, 4, and 5. The wider differentiation at the lower levels reflects that most learners in the Adult Migrant Education Program are clustered at the very early stages of proficiency. There is some limited provision for the use of three further levels, 2+, 3+ and 4+, thus allowing the possibility of 12 levels, although there is no description of these last three levels. Each of the four traditional macroskills (speaking, listening, reading, and writing) is described separately, so that a learner's proficiency profile will consist of four components, for example, S1, L1+, R2, W1-. The nine proficiency levels are also given brief descriptive titles, as follows:

0: zero proficiency
0+: initial proficiency
1-: elementary proficiency
1: minimum survival proficiency
1+: survival proficiency
2: minimum social proficiency
3: minimum vocational proficiency
4: vocational proficiency
5: native-like proficiency

The ASLPR documentation provides one or more of three kinds of information about each level.

1. A description of the language behavior appropriate to the level. For example, "L2: Minimum social proficiency. Able to understand in routine social situations and limited work situations. Can get the gist of most conversations in everyday social situations though may sometimes misinterpret or need utterances to be repeated."

2. A series of examples of observed behavior or tasks that a learner at a particular level might carry out. These examples are intended to give interviewers some guidance on the sorts of tasks learners can be asked to perform to show their proficiency at each level. For example, "W1+: Survival proficiency. Can write a note to school explaining a child's absence."

3. A series of comments explaining the key features entailed in the transition from one level to the next. For example, "S3: Minimum vocational proficiency. The key factor now emerging is register flexibility." Sometimes, however, the examples of specific tasks (see 2) are very unspecific, and look more like general descriptions of language behavior (see 1). For example, "Frequently interprets questions as statements unless repeated and redundantly marked by sentence structure, Wh-word, strong intonation, or context."

The assignment of a learner to a point on the ASLPR scale is done on the basis of an interview, during which language behavior is elicited in a series of realistic tasks. The resultant behavior is matched to that point on the scale that most nearly describes it. The point of the interview is thus to derive a global impression of the learner's behavior, and to match it with the ASLPR level definition that seems to correspond most closely to the global impression. Therefore, the decision-making rules for assigning a rating to an interviewee represent another difference between the ASLPR and the ILR scales. The ILR scale is noncompensatory; the fact that an examinee is closer to the next higher rating does not mean that the higher rating should be assigned. On the ASLPR scale, on the other hand, the higher rating would be assigned.

According to the *Report on the Formal Trialling of the ASLPR*, reliability and validity of the scale were studied in two data-gathering projects. In the first project, 24 subjects were interviewd for S, L, and R, and these interviews were recorded on video cassettes. They were also given several tasks from which W could be judged. Tapes and writing samples on 16 of the subjects were then rated by 21 Australian and 15 Chinese teachers of ESL, and their ratings were correlated with the official ratings assigned by the test developers. The correlations were all very good, usually above .90. Less impressive correlations were obtained for the same interviewees with the CELT, and with cloze and dictation exercises.

Interview-based proficiency rating scales like the ASLPR have certain strengths that may make them seem an attractive alternative to formal test instruments. They also have considerable drawbacks. In fact, the strengths and weaknesses of the two approaches might almost be said to be in complementary distribution.

A major weakness of the ASLPR (and other interview-based approaches, such as the ILR Oral Proficiency Interview) is its built-in tendency to become a variable instrument. There is no ASLPR test instrument or item bank (apart from the general guidelines for the conduct of the interview and the sample interview kits), yet ASLPR claims to be a test of the four separate macroskills. It would be more appropriate to call it a scale and a set of procedures. The guidance on the sorts of tasks learners can be asked to perform to show their proficiency at each level is in the form of examples only, with no particular canonical or absolute status.

The rater is, in fact, encouraged to replace one task or exercise that is not deemed appropriate with another that is more appropriate, as long as the latter is of equal complexity. Unfortunately, there is no guidance as to what constitutes equal complexity or equal appropriateness in the behavioral description provided at each point of the scale. Is filling in a deposit form for a bank a task of the same level of complexity (and hence equally appropriate) as filling in a change of address notice at the motor registration bureau? How does one judge? What are the criteria that constitute equivalence of functional, linguistic, or communicative complexity?

The ASLPR provides very specific writing exercises and clear guidelines for scoring them. However, we are told in the trialling study (p. 7) that if the particular task set could not be performed, the supervisor was at liberty to propose another of equivalent complexity. As long as the individual rater is free to substitute any task for the example tasks given, the whole process comes close to being the use of a variable instrument, a possible consequence of which is variable measurements. Although all instruments contain measurement error, it is especially important with the interview-based approach to establish the validity of a rater's interview, and the accuracy and consistency of his or her ratings.

The other serious drawback of an interview-based approach is the influence of interviewing technique on the rating process and its outcome. The ability to relate to a learner and set up a nonthreatening situation where the learner's language proficiency can be explored and analysed in depth must surely require special skills and sensitivities that cannot be assumed to be universally available. It may also be significant that during trial testing of the ASLPR there was no attempt to investigate the effect of the interviewer or interview length on the reliability of ASLPR ratings. Even if one assumes adequate interviewing techniques there is still the problem of interpretation of the scale. Yet, the authors