PROGRAM STUDI PENDIDIKAN BAHASA INGGRIS
UNIVERSITAS KRISTEN DUTA WACANA
ENGLISH EDUCATION DEPARTMENT - DUTA WACANA CHRISTIAN UNIVERSITY

SURAT TUGAS

No. 011c/J.05/PBI/2019

Ketua Program Studi Pendidikan Bahasa Inggris Universitas Kristen Duta Wacana Yogyakarta dengan ini memberi tugas kepada,

Nama	: Ignatius Tri Endarto, M.A.
NIK	: 174 E 444
NIDN	: 0521039101
Jabatan	: Dosen

Untuk membuat buku ajar Mata Kuliah *Language Testing and Assessment* yang dipergunakan pada Semester Genap 2018/2019 dengan judul

Module of Language Testing and Assessment

Demikian hendaknya tugas ini dapat dilakukan dengan sebaik-baiknya dan hasilnya dilaporkan kepada pemberi tugas.

Yogyakarta, 21 Januari 2019

Paulus Widiatmoko, M.A.
Kaprodi PBI UKDW

Tembusan:
1. Arsip

PW/dst

Jl. dr. Wahidin Sudirohusodo 5-25, Yogyakarta 55224, Indonesia
Telp : +62 274 563929 ext 123/143
https : //www.ukdw.ac.id | e-mail : pbi@staff.ukdw.ac.id

MODULE OF

Language Testing and Assessment

Compiled by:
Ignatius Tri Endarto, M.A.

English Education Department
2nd Semester of Academic Year 2018/2019
UNIVERSITAS KRISTEN DUTA WACANA

Module of
Language Testing and Assessment

Compiled by:
Ignatius Tri Endarto, M.A.

This module is intended for
Language Testing and Assessment course
at the English Education Department
in the 2nd Semester of 2018/2019 Academic Year.
Universitas Kristen Duta Wacana

Halaman Pengesahan

Nama modul	:	*Module of Language Testing and Assessment*
Jumlah penyusun	:	1 (satu)
Identitas penyusun		
Nama	:	Ignatius Tri Endarto, M.A.
NIDN/NIK	:	0521039101/174E444
Jabatan/Golongan	:	Asisten Ahli 150/IIIB
Program studi	:	Pendidikan Bahasa Inggris
Bidang keahlian	:	Pendidikan Bahasa Inggris/Linguistik
Semester	:	Genap 2018/2019
Jumlah halaman	:	120 halaman
Biaya total	:	-
Sumber biaya	:	-
Waktu/durasipembuatan	:	21 Januari-18 Februari 2019

Yogyakarta, 18 Februari 2019

Mengesahkan,

Kaprodi PBI	Penyusun modul
Paulus Widiatmoko, M.A.	Ignatius Tri Endarto, M.A.
NIK: 064E320	NIK: 174E444

DAFTAR ISI

DESKRIPSI MATA KULIAH

<table>
<tr><td rowspan="6">Capaian Pembelajaran (CP)</td><td colspan="2">CPL – PRODI</td></tr>
<tr><td>LO7

LO14

LO18</td><td>Menguasai konsep teoritis mengenai perancangan alat ukur keberhasilan pembelajaran bahasa Inggris;

Mampu merancang dan membuat alat ukur keberhasilan pembelajaran dengan berbagai metode pengujian dan assessment salah satu atau beberapa bidang ESP;

Mampu menerapkan pemikiran logis, kritis, sistematis, dan inovatif dalam pengembangan kajian perancangan program, pengajaran, dan penelitian ESP dengan menerapkan pengetahuan dan/atau teknologi sesuai dengan bidang tersebut;</td></tr>
<tr><td colspan="2">CP – MK</td></tr>
<tr><td>M1</td><td>Mahasiswa mampu mendemonstrasikan pemahaman teoretis mengenai konsep dasar, metode, tujuan, serta jenis-jenis language testing dan assessment secara tepat melalui berbagai kegiatan seperti kuis, presentasi, project, dan penugasan (LO7, LO14, LO18);</td></tr>
<tr><td>M2</td><td>Mahasiswa mampu mengevaluasi berbagai jenis dan metode testing/assessment untuk mengukur kemampuan listening, reading, speaking, writing, maupun integrated skills dengan benar sesuai prinsip-prinsip dalam language testing/assessment (LO7, LO18);</td></tr>
<tr><td>M3</td><td>Mahasiswa mampu merancang, membuat, dan menerapkan contoh-contoh instrumen tes untuk berbagai tujuan (both General English and English for Specific Purposes) dengan mengikuti prosedur dan tahapan yang sesuai (LO7, LO14, LO18);</td></tr>
<tr><td>Deskripsi Singkat MK</td><td colspan="2">This course introduces students to the fundamental principles of language testing and language test evaluation. Moreover, it prepares students to develop skills in the design, trialing, moderation and validation of testing instruments for the purposes of practicing language skills.

At the beginning of this course, students will familiarize themselves with the basic concepts and principles of language testing and assessment. As they get through the course, they will be given theories and examples of language tests on various skills (listening, reading, speaking, writing, and integrated skills). Students will also evaluate a number of different test samples in order to identify pivotal features of tests to include in their own test making. At the end, students are expected to present their designed tests in front of the class and evaluate their peers' works.</td></tr>
</table>

Materi Pembelajaran/ Pokok Bahasan	1. Introduction to Language Testing & Assessment 2. Validity and Principles of Language Assessment 3. Assessing Listening 4. Assessing Reading 5. Assessing Speaking 6. Assessing Writing

INTRODUCTION TO LANGUAGE TESTING AND ASSESSMENT

WHAT IS A TEST?

A test, in simple terms, is a *method of measuring a person's ability, knowledge, or performance in a given domain.* Let's look at the components of this definition. A test is first a method. It is an instrument—a set of techniques, procedures, or items—that requires performance on the part of the test-taker. To qualify as a test, the method must be explicit and structured: multiple-choice questions with prescribed correct answers; a writing prompt with a scoring rubric; an oral interview based on a question script and a checklist of expected responses to be filled in by the administrator.

Second, a test must measure. Some tests measure general ability, while others focus on very specific competencies or objectives. A multi-skill proficiency test determines a general ability level; a quiz on recognizing correct use of definite articles measures specific knowledge. The way the results or measurements are communicated may vary. Some tests, such as a classroom-based short-answer essay test, may earn the test-taker a letter grade accompanied by the instructor's marginal comments. Others, particularly large-scale standardized tests, provide a total numerical score, a percentile rank, and perhaps some subscores. If an instrument does not specify a form of reporting measurement—a means for offering the test-taker some kind of result—then that technique cannot appropriately be defined as a test.

Next, a test measures an individual's ability, knowledge, or performance. Testers need to understand who the test-takers are. What is their previous experience and background? Is the test appropriately matched to their abilities? How should test-takers interpret their scores?

A test measures *performance,* but the results imply the test-taker's ability, or, to use a concept common in the field of linguistics, competence. Most language tests measure one's ability to perform language, that is, to speak, write, read, or listen to a subset of language. On the other hand, it is not uncommon to find tests designed to tap into a test-taker's knowledge about language: defining a vocabulary item, reciting a grammatical rule, or identifying a rhetorical feature in written discourse. Performance-based tests sample the test-taker's actual use of language, but from those samples the test administrator infers general competence. A test of reading comprehension, for example, may consist of several short reading passages each followed by a limited number of comprehension questions—a small sample of a second language learner's total reading behavior. But from the results of that test, the examiner may infer a certain level of general reading ability.

Finally, a test measures a given domain. In the case of a proficiency test, even though the actual performance on the test involves only a sampling of skills, that domain is overall proficiency in a language—general competence in all skills of a language. Other tests may have more specific criteria. A test of pronunciation might well be a test of only a limited set of phonemic minimal pairs. A vocabulary test may focus on only the set of words covered in a particular lesson or unit. One of the biggest obstacles to overcome in constructing adequate tests is to measure the desired criterion and not include other factors inadvertently, an issue that is addressed in Chapters 2 and 3.

A well-constructed test is an instrument that provides an accurate measure of the test-taker's ability within a particular domain. The definition sounds fairly simple, but in fact, constructing a good test is a complex task involving both science and art.

ASSESSMENT AND TEACHING

Assessment is a popular and sometimes misunderstood term in current educational practice. You might be tempted to think of testing and assessing as synonymous terms, but they are not. Tests are prepared administrative procedures that occur at identifiable times in a curriculum when learners muster all their faculties to offer peak performance, knowing that their responses are being measured and evaluated.

Assessment, on the other hand, is an ongoing process that encompasses a much wider domain. Whenever a student responds to a question, offers a comment, or tries out a new word or structure, the teacher subconsciously makes an assessment of the student's performance. Written work—from a jotted-down phrase to a formal essay—is performance that ultimately is assessed by self, teacher, and possibly other students. Reading and listening activities usually require some sort of productive performance that the teacher implicitly judges, however peripheral that judgment may be. A good teacher never ceases to assess students, whether those assessments are incidental or intended.

Tests, then, are a subset of assessment; they are certainly not the only form of assessment that a teacher can make. Tests can be useful devices, but they are only one among many procedures and tasks that teachers can ultimately use to assess students.

But now, you might be thinking, if you make assessments every time you teach something in the classroom, does all teaching involve assessment? Are teachers constantly assessing students with no interaction that is assessment-free?

The answer depends on your perspective. For optimal learning to take place, students in the classroom must have the freedom to experiment, to try out their own hypotheses about language without feeling that their overall competence is being judged in terms of those trials and errors. In the same way that tournament tennis players must, before a tournament, have the freedom to practice their skills with no implications for their final placement on that day of days, so also must learners have ample opportunities to "play" with language in a classroom without being formally

graded. Teaching sets up the practice games of language learning: the opportunities for learners to listen, think, take risks, set goals, and process feedback from the "coach" and then recycle through the skills that they are trying to master. (A diagram of the relationship among testing, teaching, and assessment is found in Figure 1.1.)

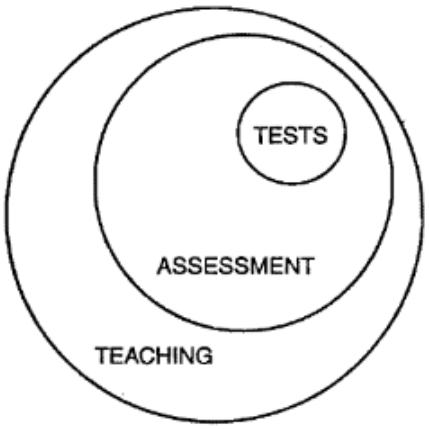


Figure 1.1. Tests, assessment, and teaching

At the same time, during these practice activities, teachers (and tennis coaches) are indeed observing students' performance and making various evaluations of each learner: How did the performance compare to previous performance? Which aspects of the performance were better than others? Is the learner performing up to an expected potential? How does the performance compare to that of others in the same learning community? In the ideal classroom, all these observations feed into the way the teacher provides instruction to each student.

Informal and Formal Assessment

One way to begin untangling the lexical conundrum created by distinguishing among tests, assessment, and teaching is to distinguish between informal and formal assessment. **Informal assessment** can take a number of forms, starting with incidental, unplanned comments and responses, along with coaching and other impromptu feedback to the student. Examples include saying "Nice job!" "Good work!" "Did you say *can* or *can't*?" "I think you meant to say you *broke* the glass, not you *break* the glass," or putting a ☺ on some homework.

Informal assessment does not stop there. A good deal of a teacher's informal assessment is embedded in classroom tasks designed to elicit performance without recording results and making fixed judgments about a student's competence. Examples at this end of the continuum are marginal comments on papers, responding to a draft of an essay, advice about how to better pronounce a word, a

suggestion for a strategy for compensating for a reading difficulty, and showing how to modify a student's note-taking to better remember the content of a lecture.

On the other hand, **formal assessments** are exercises or procedures specifically designed to tap into a storehouse of skills and knowledge. They are systematic, planned sampling techniques constructed to give teacher and student an appraisal of student achievement. To extend the tennis analogy, formal assessments are the tournament games that occur periodically in the course of a regimen of practice.

Is formal assessment the same as a test? We can say that all tests are formal assessments, but not all formal assessment is testing. For example, you might use a student's journal or portfolio of materials as a formal assessment of the attainment of certain course objectives, but it is problematic to call those two procedures "tests." A systematic set of observations of a student's frequency of oral participation in class is certainly a formal assessment, but it too is hardly what anyone would call a test. Tests are usually relatively time-constrained (usually spanning a class period or at most several hours) and draw on a limited sample of behavior.

Formative and Summative Assessment

Another useful distinction to bear in mind is the function of an assessment: How is the procedure to be used? Two functions are commonly identified in the literature: formative and summative assessment. Most of our classroom assessment is **formative assessment:** evaluating students in the process of "forming" their competencies and skills with the goal of helping them to continue that growth process. The key to such formation is the delivery (by the teacher) and internalization (by the student) of appropriate feedback on performance, with an eye toward the future continuation (or formation) of learning.

For all practical purposes, virtually all kinds of informal assessment are (or should be) formative. They have as their primary focus the ongoing development of the learner's language. So when you give a student a comment or a suggestion, or call attention to an error, that feedback is offered in order to improve the learner's language ability.

Summative assessment aims to measure, or summarize, what a student has grasped, and typically occurs at the end of a course or unit of instruction. A summation of what a student has learned implies looking back and taking stock of how well that student has accomplished objectives, but does not necessarily point the way to future progress. Final exams in a course and general proficiency exams are examples of summative assessment.

One of the problems with prevailing attitudes toward testing is the view that *all* tests (quizzes, periodic review tests, midterm exams, etc.) are summative. At various points in your past educational experiences, no doubt you've considered such tests as summative. You may have thought, "Whew! I'm glad that's over. Now I don't have to remember that stuff anymore!" A challenge to you as a teacher is to change that attitude among your students: Can you instill a more formative quality to what

your students might otherwise view as a summative test? Can you offer your students an opportunity to convert tests into "learning experiences"? We will take up that challenge in subsequent chapters in this book.

Norm-Referenced and Criterion-Referenced Tests

Another dichotomy that is important to clarify here and that aids in sorting out common terminology in assessment is the distinction between norm-referenced and criterion-referenced testing. In **norm-referenced tests,** each test-taker's score is interpreted in relation to a mean (average score), median (middle score), standard deviation (extent of variance in scores), and/or percentile rank. The purpose in such tests is to place test-takers along a mathematical continuum in rank order. Scores are usually reported back to the test-taker in the form of a numerical score (for example, 230 out of 300) and a percentile rank (such as 84 percent, which means that the test-taker's score was higher than 84 percent of the total number of test-takers, but lower than 16 percent in that administration). Typical of norm-referenced tests are standardized tests like the Scholastic Aptitude Test (SAT®) or the Test of English as a Foreign Language (TOEFL®), intended to be administered to large audiences, with results efficiently disseminated to test-takers. Such tests must have fixed, predetermined responses in a format that can be scored quickly at minimum expense. Money and efficiency are primary concerns in these tests.

Criterion-referenced tests, on the other hand, are designed to give test-takers feedback, usually in the form of grades, on specific course or lesson objectives. Classroom tests involving the students in only one class, and connected to a curriculum, are typical of criterion-referenced testing. Here, much time and effort on the part of the teacher (test administrator) are sometimes required in order to deliver useful, appropriate feedback to students, or what Oller (1979, p. 52) called "instructional value." In a criterion-referenced test, the distribution of students' scores across a continuum may be of little concern as long as the instrument assesses appropriate objectives. In *Language Assessment,* with an audience of classroom language teachers and teachers in training, and with its emphasis on classroom-based assessment (as opposed to standardized, large-scale testing), criterion-referenced testing is of more prominent interest than norm-referenced testing.

APPROACHES TO LANGUAGE TESTING: A BRIEF HISTORY

Now that you have a reasonably clear grasp of some common assessment terms, we now turn to one of the primary concerns of this book: the creation and use of tests, particularly classroom tests. A brief history of language testing over the past half-century will serve as a backdrop to an understanding of classroom-based testing.

Historically, language-testing trends and practices have followed the shifting sands of teaching methodology (for a description of these trends, see Brown,

Teaching by Principles [hereinafter *TBP*], Chapter 2).[1] For example, in the 1950s, an era of behaviorism and special attention to contrastive analysis, testing focused on specific language elements such as the phonological, grammatical, and lexical contrasts between two languages. In the 1970s and 1980s, communicative theories of language brought with them a more integrative view of testing in which specialists claimed that "the whole of the communicative event was considerably greater than the sum of its linguistic elements" (Clark, 1983, p. 432). Today, test designers are still challenged in their quest for more authentic, valid instruments that simulate real-world interaction.

Discrete-Point and Integrative Testing

This historical perspective underscores two major approaches to language testing that were debated in the 1970s and early 1980s. These approaches still prevail today, even if in mutated form: the choice between discrete-point and integrative testing methods (Oller, 1979). **Discrete-point tests** are constructed on the assumption that language can be broken down into its component parts and that those parts can be tested successfully. These components are the skills of listening, speaking, reading, and writing, and various units of language (discrete points) of phonology/graphology, morphology, lexicon, syntax, and discourse. It was claimed that an overall language proficiency test, then, should sample all four skills and as many linguistic discrete points as possible.

Such an approach demanded a decontextualization that often confused the test-taker. So, as the profession emerged into an era of emphasizing communication, authenticity, and context, new approaches were sought. Oller (1979) argued that language competence is a unified set of interacting abilities that cannot be tested separately. His claim was that communicative competence is so global and requires such integration (hence the term "integrative" testing) that it cannot be captured in additive tests of grammar, reading, vocabulary, and other discrete points of language. Others (among them Cziko, 1982, and Savignon, 1982) soon followed in their support for integrative testing.

What does an **integrative test** look like? Two types of tests have historically been claimed to be examples of integrative tests: cloze tests and dictations. A **cloze test** is a reading passage (perhaps 150 to 300 words) in which roughly every sixth or seventh word has been deleted; the test-taker is required to supply words that fit into those blanks. (See Chapter 8 for a full discussion of cloze testing.) Oller (1979)

claimed that cloze test results are good measures of overall proficiency. According to theoretical constructs underlying this claim, the ability to supply appropriate words in blanks requires a number of abilities that lie at the heart of competence in a language: knowledge of vocabulary, grammatical structure, discourse structure, reading skills and strategies, and an internalized "expectancy" grammar (enabling one to predict an item that will come next in a sequence). It was argued that successful completion of cloze items taps into all of those abilities, which were said to be the essence of global language proficiency.

Dictation is a familiar language-teaching technique that evolved into a testing technique. Essentially, learners listen to a passage of 100 to 150 words read aloud by an administrator (or audiotape) and write what they hear, using correct spelling. The listening portion usually has three stages: an oral reading without pauses; an oral reading with long pauses between every phrase (to give the learner time to write down what is heard); and a third reading at normal speed to give test-takers a chance to check what they wrote. (See Chapter 6 for more discussion of dictation as an assessment device.)

Supporters argue that dictation is an integrative test because it taps into grammatical and discourse competencies required for other modes of performance in a language. Success on a dictation requires careful listening, reproduction in writing of what is heard, efficient short-term memory, and, to an extent, some expectancy rules to aid the short-term memory. Further, dictation test results tend to correlate strongly with other tests of proficiency. Dictation testing is usually classroom-centered since large-scale administration of dictations is quite impractical from a scoring standpoint. Reliability of scoring criteria for dictation tests can be improved by designing multiple-choice or exact-word cloze test scoring.

Proponents of integrative test methods soon centered their arguments on what became known as the **unitary trait hypothesis,** which suggested an "indivisible" view of language proficiency: that vocabulary, grammar, phonology, the "four skills," and other discrete points of language could not be disentangled from each other in language performance. The unitary trait hypothesis contended that there is a general factor of language proficiency such that all the discrete points do *not* add up to that whole.

Others argued strongly against the unitary trait position. In a study of students in Brazil and the Philippines, Farhady (1982) found significant and widely varying differences in performance on an ESL proficiency test, depending on subjects' native country, major field of study, and graduate versus undergraduate status. For example, Brazilians scored very low in listening comprehension and relatively high in reading comprehension. Filipinos, whose scores on five of the six components of the test were considerably higher than Brazilians' scores, were actually lower than Brazilians in reading comprehension scores. Farhady's contentions were supported in other research that seriously questioned the unitary trait hypothesis. Finally, in the face of the evidence, Oller retreated from his earlier stand and admitted that "the unitary trait hypothesis was wrong" (1983, p. 352).

Communicative Language Testing

By the mid-1980s, the language-testing field had abandoned arguments about the unitary trait hypothesis and had begun to focus on designing communicative language-testing tasks. Bachman and Palmer (1996, p. 9) include among "fundamental" principles of language testing the need for a correspondence between language test-performance and language use: "In order for a particular language test to be useful for its intended purposes, test performance must correspond in demonstrable ways to language use in non-test situations." The problem that language assessment experts faced was that tasks tended to be artificial, contrived, and unlikely to mirror language use in real life. As Weir (1990, p. 6) noted, "Integrative tests such as cloze only tell us about a candidate's linguistic competence. They do not tell us anything directly about a student's performance ability."

And so a quest for authenticity was launched, as test designers centered on communicative performance. Following Canale and Swain's (1980) model of communicative competence, Bachman (1990) proposed a model of language competence consisting of organizational and pragmatic competence, respectively subdivided into grammatical and textual components, and into illocutionary and sociolinguistic components. (Further discussion of both Canale and Swain's and Bachman's models can be found in *PLLT,* Chapter 9.) Bachman and Palmer (1996, pp. 70f) also emphasized the importance of **strategic competence** (the ability to employ communicative strategies to compensate for breakdowns as well as to enhance the rhetorical effect of utterances) in the process of communication. All elements of the model, especially pragmatic and strategic abilities, needed to be included in the constructs of language testing and in the actual performance required of test-takers.

Communicative testing presented challenges to test designers, as we will see in subsequent chapters of this book. Test constructors began to identify the kinds of real-world tasks that language learners were called upon to perform. It was clear that the contexts for those tasks were extraordinarily widely varied and that the sampling of tasks for any one assessment procedure needed to be validated by what language users actually do with language. Weir (1990, p. 11) reminded his readers that "to measure language proficiency ... account must now be taken of: where, when, how, with whom, and why language is to be used, and on what topics, and with what effect." And the assessment field became more and more concerned with the authenticity of tasks and the genuineness of texts. (See Skehan, 1988, 1989, for a survey of communicative testing research.)

Performance-Based Assessment

In language courses and programs around the world, test designers are now tackling this new and more student-centered agenda (Alderson, 2001, 2002). Instead of just offering paper-and-pencil selective response tests of a plethora of separate items, performance-based assessment of language typically involves oral production,

written production, open-ended responses, integrated performance (across skill areas), group performance, and other interactive tasks. To be sure, such assessment is time-consuming and therefore expensive, but those extra efforts are paying off in the form of more direct testing because students are assessed as they perform actual or simulated real-world tasks. In technical terms, higher content validity (see Chapter 2 for an explanation) is achieved because learners are measured in the process of performing the targeted linguistic acts.

In an English language-teaching context, performance-based assessment means that you may have a difficult time distinguishing between formal and informal assessment. If you rely a little less on formally structured tests and a little more on evaluation while students are performing various tasks, you will be taking some steps toward meeting the goals of performance-based testing. (See Chapter 10 for a further discussion of performance-based assessment.)

A characteristic of many (but not all) performance-based language assessments is the presence of interactive tasks. In such cases, the assessments involve learners in actually performing the behavior that we want to measure. In interactive tasks, test-takers are measured in the act of speaking, requesting, responding, or in combining listening and speaking, and in integrating reading and writing. Paper-and-pencil tests certainly do not elicit such communicative performance.

A prime example of an interactive language assessment procedure is an oral interview. The test-taker is required to listen accurately to someone else and to respond appropriately. If care is taken in the test design process, language elicited and volunteered by the student can be personalized and meaningful, and tasks can approach the authenticity of real-life language use (see Chapter 7).

CURRENT ISSUES IN CLASSROOM TESTING

The design of communicative, performance-based assessment rubrics continues to challenge both assessment experts and classroom teachers. Such efforts to improve various facets of classroom testing are accompanied by some stimulating issues, all of which are helping to shape our current understanding of effective assessment. Let's look at three such issues: the effect of new theories of intelligence on the testing industry; the advent of what has come to be called "alternative"assessment; and the increasing popularity of computer-based testing.

New Views on Intelligence

Intelligence was once viewed strictly as the ability to perform (a) linguistic and (b) logical-mathematical problem solving. This "IQ" (intelligence quotient) concept of intelligence has permeated the Western world and its way of testing for almost a century. Since "smartness" in general is measured by timed, discrete-point tests consisting of a hierarchy of separate items, why shouldn't every field of study be so measured? For many years, we have lived in a world of standardized, norm-referenced

tests that are timed in a multiple-choice format consisting of a multiplicity of logic-constrained items, many of which are inauthentic.

However, research on intelligence by psychologists like Howard Gardner, Robert Sternberg, and Daniel Goleman has begun to turn the psychometric world upside down. Gardner (1983, 1999), for example, extended the traditional view of intelligence to seven different components.[2] He accepted the traditional conceptualizations of linguistic intelligence and logical-mathematical intelligence on which standardized IQ tests are based, but he included five other "frames of mind" in his theory of multiple intelligences:

- spatial intelligence (the ability to find your way around an environment, to form mental images of reality)
- musical intelligence (the ability to perceive and create pitch and rhythmic patterns)
- bodily-kinesthetic intelligence (fine motor movement, athletic prowess)
- interpersonal intelligence (the ability to understand others and how they feel, and to interact effectively with them)
- intrapersonal intelligence (the ability to understand oneself and to develop a sense of self-identity)

Robert Sternberg (1988, 1997) also charted new territory in intelligence research in recognizing creative thinking and manipulative strategies as part of intelligence. All "smart" people aren't necessarily adept at fast, reactive thinking. They may be very innovative in being able to think beyond the normal limits imposed by existing tests, but they may need a good deal of processing time to enact this creativity. Other forms of smartness are found in those who know how to manipulate their environment, namely, other people. Debaters, politicians, successful salespersons, smooth talkers, and con artists are all smart in their manipulative ability to persuade others to think their way, vote for them, make a purchase, or do something they might not otherwise do.

More recently, Daniel Goleman's (1995) concept of "EQ" (emotional quotient) has spurred us to underscore the importance of the emotions in our cognitive processing. Those who manage their emotions—especially emotions that can be detrimental—tend to be more capable of fully intelligent processing. Anger, grief, resentment, self-doubt, and other feelings can easily impair peak performance in everyday tasks as well as higher-order problem solving.

These new conceptualizations of intelligence have not been universally accepted by the academic community (see White, 1998, for example). Nevertheless, their intuitive appeal infused the decade of the 1990s with a sense of both freedom and responsibility in our testing agenda. Coupled with parallel educational reforms at the time (Armstrong, 1994), they helped to free us from relying exclusively on

timed, discrete-point, analytical tests in measuring language. We were prodded to cautiously combat the potential tyranny of "objectivity" and its accompanying impersonal approach. But we also assumed the responsibility for tapping into whole language skills, learning processes, and the ability to negotiate meaning. Our challenge was to test interpersonal, creative, communicative, interactive skills, and in doing so to place some trust in our subjectivity and intuition.

Traditional and "Alternative" Assessment

Implied in some of the earlier description of performance-based classroom assessment is a trend to supplement traditional test designs with alternatives that are more authentic in their elicitation of meaningful communication. Table 1.1 highlights differences between the two approaches (adapted from Armstrong, 1994, and Bailey, 1998, p. 207).

Two caveats need to be stated here. First, the concepts in Table 1.1 represent some overgeneralizations and should therefore be considered with caution. It is difficult, in fact, to draw a clear line of distinction between what Armstrong (1994) and Bailey (1998) have called traditional and alternative assessment. Many forms of assessment fall in between the two, and some combine the best of both.

Second, it is obvious that the table shows a bias toward alternative assessment, and one should not be misled into thinking that everything on the left-hand side is tainted while the list on the right-hand side offers salvation to the field of language assessment! As Brown and Hudson (1998) aptly pointed out, the assessment traditions available to us should be valued and utilized for the functions that they provide. At the same time, we might all be stimulated to look at the right-hand list and ask ourselves if, among those concepts, there are alternatives to assessment that we can constructively use in our classrooms.

It should be noted here that considerably more time and higher institutional budgets are required to administer and score assessments that presuppose more

Table 1.1. Traditional and alternative assessment

Traditional Assessment	Alternative Assessment
One-shot, standardized exams	Continuous long-term assessment
Timed, multiple-choice format	Untimed, free-response format
Decontextualized test items	Contextualized communicative tasks
Scores suffice for feedback	Individualized feedback and washback
Norm-referenced scores	Criterion-referenced scores
Focus on the "right" answer	Open-ended, creative answers
Summative	Formative
Oriented to product	Oriented to process
Non-interactive performance	Interactive performance
Fosters extrinsic motivation	Fosters intrinsic motivation

subjective evaluation, more individualization, and more interaction in the process of offering feedback. The payoff for the latter, however, comes with more useful feedback to students, the potential for intrinsic motivation, and ultimately a more complete description of a student's ability. (See Chapter 10 for a complete treatment of alternatives in assessment.) More and more educators and advocates for educational reform are arguing for a de-emphasis on large-scale standardized tests in favor of building budgets that will offer the kind of contextualized, communicative performance-based assessment that will better facilitate learning in our schools. (In Chapter 4, issues surrounding standardized testing are addressed at length.)

Computer-Based Testing

Recent years have seen a burgeoning of assessment in which the test-taker performs responses on a computer. Some computer-based tests (also known as "computer-assisted" or "web-based" tests) are small-scale "home-grown" tests available on websites. Others are standardized, large-scale tests in which thousands or even tens of thousands of test-takers are involved. Students receive prompts (or probes, as they are sometimes referred to) in the form of spoken or written stimuli from the computerized test and are required to type (or in some cases, speak) their responses. Almost all computer-based test items have fixed, closed-ended responses; however, tests like the Test of English as a Foreign Language (TOEFL®) offer a written essay section that must be scored by humans (as opposed to automatic, electronic, or machine scoring). As this book goes to press, the designers of the TOEFL are on the verge of offering a spoken English section.

A specific type of computer-based test, a **computer-adaptive test,** has been available for many years but has recently gained momentum. In a computer-adaptive test (CAT), each test-taker receives a set of questions that meet the test specifications and that are generally appropriate for his or her performance level. The CAT starts with questions of moderate difficulty. As test-takers answer each question, the computer scores the question and uses that information, as well as the responses to previous questions, to determine which question will be presented next. As long as examinees respond correctly, the computer typically selects questions of greater or equal difficulty. Incorrect answers, however, typically bring questions of lesser or equal difficulty. The computer is programmed to fulfill the test design as it continuously adjusts to find questions of appropriate difficulty for test-takers at all performance levels. In CATs, the test-taker sees only one question at a time, and the computer scores each question before selecting the next one. As a result, test-takers cannot skip questions, and once they have entered and confirmed their answers, they cannot return to questions or to any earlier part of the test.

Computer-based testing, with or without CAT technology, offers these advantages:

- classroom-based testing
- self-directed testing on various aspects of a language (vocabulary, grammar, discourse, one or all of the four skills, etc.)

- practice for upcoming high-stakes standardized tests
- some individualization, in the case of CATs
- large-scale standardized tests that can be administered easily to thousands of test-takers at many different stations, then scored electronically for rapid reporting of results

Of course, some disadvantages are present in our current predilection for computerizing testing. Among them:

- Lack of security and the possibility of cheating are inherent in classroom-based, unsupervised computerized tests.
- Occasional "home-grown" quizzes that appear on unofficial websites may be mistaken for validated assessments.
- The multiple-choice format preferred for most computer-based tests contains the usual potential for flawed item design (see Chapter 3).
- Open-ended responses are less likely to appear because of the need for human scorers, with all the attendant issues of cost, reliability, and turn-around time.
- The human interactive element (especially in oral production) is absent.

More is said about computer-based testing in subsequent chapters, especially Chapter 4, in a discussion of large-scale standardized testing. In addition, the following websites provide further information and examples of computer-based tests:

Educational Testing Service	**www.ets.org**
Test of English as a Foreign Language	**www.toefl.org**
Test of English for International Communication	**www.toeic.com**
International English Language Testing System	**www.ielts.org**
Dave's ESL Café (computerized quizzes)	**www.eslcafe.com**

Some argue that computer-based testing, pushed to its ultimate level, might mitigate against recent efforts to return testing to its artful form of being tailored by teachers for their classrooms, of being designed to be performance-based, and of allowing a teacher–student dialogue to form the basis of assessment. This need not be the case. Computer technology can be a boon to communicative language testing. Teachers and test-makers of the future will have access to an ever-increasing range of tools to safeguard against impersonal, stamped-out formulas for assessment. By using technological innovations creatively, testers will be able to enhance authenticity, to increase interactive exchange, and to promote autonomy.

VALIDITY AND PRINCIPLES OF LANGUAGE ASSESSMENT

A1.2 THREE 'TYPES' OF VALIDITY IN EARLY THEORY

In the early days of validity investigation, validity was broken down into three 'types' that were typically seen as distinct. Each type of validity was related to the kind of evidence that would count towards demonstrating that a test was valid. Cronbach and Meehl (1955) described these as:

- Criterion-oriented validity
  - Predictive validity
  - Concurrent validity
- Content validity
- Construct validity

We will introduce each of these in turn, and then show how this early approach has changed.

A1.2.1 Criterion-oriented validity

When considering criterion-oriented validity, the tester is interested in the relationship between a particular test and a criterion to which we wish to make predictions. For example, I may wish to predict from scores on a test of second-language academic reading ability whether individuals can cope with first-semester undergraduate business studies texts in an English-medium university. What we are really interested in here is the criterion, whatever it is that we wish to know about, but for which we don't have any direct evidence. In the example above we cannot see whether future students can do the reading that will be expected of them before they actually arrive at the university and start their course.

In this case the validity evidence is the strength of the predictive relationship between the test score and that performance on the criterion. Of course, it is necessary to decide what would count as 'ability to cope with' – as it is something that must be measurable. Defining precisely what we mean by such words and phrases is a central part of investigating validity.

Predictive validity is the term used when the test scores are used to predict some future criterion, such as academic success. If the scores are used to predict a criterion at the same time the test is given, we are studying *concurrent validity*.

Returning to the example given above, let us assume that in this case 'ability to cope' is defined as a subject tutor's judgment of whether students can adequately read set texts to understand lectures and write assignments. We might be interested in discovering the relationship between students' scores on our test prior to starting academic studies and the judgments of the tutors once the students have started their programme. This would be a *predictive validity study*. We would hope that we could identify a score on the reading test above which tutors would judge readers to be competent, and below which they would judge some readers to lack the necessary reading skills for academic study. This would be the 'cut score' for making a predictive decision about the likelihood of future success on the criterion.

Suppose that my reading test is too long, and for practical purposes it needs to be made much shorter. As we know that shorter tests mean that we collect less evidence about reading ability, one of the questions we would wish to ask is to what extent the shorter test is capable of predicting the scores on the longer test. In other words, could the shorter test replace the larger test and still be useful? This would be an example of a *concurrent validity study* that uses the longer test as the criterion.

A1.2.2 Content validity

Content validity is defined as any attempt to show that the content of the test is a representative sample from the domain that is to be tested. In our example of the academic reading test it would be necessary to show that the texts selected for the test are typical of the types of texts that would be used in first-year undergraduate business courses. This is usually done using expert judges. These may be subject teachers, or language teachers who have many years' experience in teaching business English. The judges are asked to look at texts that have been selected for inclusion on the test and evaluate them for their representativeness within the content area. Secondly, the items used on the test should result in responses to the text from which we can make inferences about the test takers' ability to process the texts in ways expected of students on their academic courses. For example, we may discover that business students are primarily required to read texts to extract key factual information, take notes and use the notes in writing assignments. In our reading test we would then try to develop items that tap the ability to identify key facts.

Carroll (1980: 67) argued that achieving content validity in testing English for Academic Purposes (EAP) consisted of describing the test takers, analysing their 'communicative needs' and specifying test content on the basis of their needs. In early approaches to communicative language testing the central issue in establishing content validity was how best to 'sample' from needs and the target domain (Fulcher, 1999a: 222–223).

A1.2.3 Construct validity

The first problem with construct validity is defining what a 'construct' is. Perhaps the easiest way to understand the term 'construct' is to think of the many abstract nouns that we use on a daily basis, but for which it would be extremely hard to point to an example. Consider these, the first of which we have already touched on.

1 Love
2 Intelligence
3 Anxiety
4 Thoughtfulness
5 Fluency
6 Aptitude
7 Extroversion
8 Timidity
9 Persuasiveness
10 Empathy.

As we use these terms in everyday life we have no need to define them. We all assume that we know what they mean, and that the meaning is shared. So we can talk with our friends about how much empathy someone we know may have, or how fluent a speaker someone is. But this is to talk at the level of everyday concepts. For a general term to become a construct, it must have two further properties. Firstly, it must be defined in such a way that it becomes measurable. In order to measure 'fluency' we have to state what we could possibly observe in speech to make a decision about whether a speaker is fluent. It turns out that many people have different definitions of fluency, ranging from simple speed of speech, to lack of hesitation (or strictly 'pauses', because 'hesitation' is a construct itself), to specific observable features of speech (see Fulcher, 1996). Secondly, any construct should be defined in such a way that it can have relationships with other constructs that are different. For example, if I generate descriptions of 'fluency' and 'anxiety' I may hypothesize that, as anxiety increases, fluency will decrease, and vice versa. If this hypothesis is tested and can be supported, we have the very primitive beginnings of a theory of speaking that relates how we perform to emotional states.

To put this another way, concepts become constructs when they are so defined that they can become 'operational' – we can measure them in a test of some kind by linking the term to something observable (whether this is ticking a box or performing some communicative action), and we can establish the place of a construct in a theory that relates one construct to another (Kerlinger and Lee, 2000: 40), as in the case of fluency and anxiety above.

A1.2.4 Construct validity and truth

In the early history of validity theory there was an assumption that there is such a thing as a 'psychologically real construct' that has an independent existence in the test taker, and that the test scores represent the degree of presence or absence of this very real property. As Cronbach and Meehl (1955: 284) put it:

> Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim.

This brings us to our first philosophical observation. It has frequently been argued that early validity theorists were positivistic in their outlook. That is, they assumed that their constructs actually existed in the heads of the test takers. Again, Cronbach and Meehl (1955: 284) state: 'Scientifically speaking, to "make clear what something is" means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a nomological network.'

The idea of a nomological network is not difficult to grasp. Firstly, it contains a number of constructs, and their names are abstract, like those in the list above. In language teaching and testing, 'fluency' and 'accuracy' are two well-known constructs. Secondly, the nomological network contains the observable variables – those things that we can see and measure directly, whereas we cannot see 'fluency' and 'accuracy' directly.

What might these observable variables be? Whatever we choose makes up the definition of the constructs. For fluency we may wish to observe speed of delivery or the number of unfilled pauses, for example. For accuracy, we could look at the ratio of correct to incorrect tense use or word order. From what we can observe, we then make an inference about how 'fluent' or how 'accurate' a student's use of the second language is.

The network is created by asking what we expect the relationship between 'fluency' and 'accuracy' to be. One hypothesis could be that in speech, as fluency increases, accuracy decreases, because learners cannot pay attention to form when the demands of processing take up all the capacity of short-term memory. Another hypothesis could be that, as accuracy increases, the learner becomes more fluent, because language form has become automatic. Stating this kind of relationship between constructs therefore constitutes a theory, and theory is very powerful. Even in this simple example we could now set out a testable research hypothesis: fluency and accuracy are inversely related in students below X level of proficiency, and above it they are positively related.

A1.3.3 Pragmatic validity

What we learn from the different approaches and definitions of validity is that validity theory itself is changing and evolving. We also learn that the things we look at to investigate validity may change over time. Similarly, our understanding of the validity of test use for a particular purpose is dependent upon evidence that supports that use, but the evidence and arguments surrounding them may be challenged, undermined or developed, over time.

What we call pragmatic validity is therefore dependent upon a view that in language testing there is no such thing as an 'absolute' answer to the validity question. The role of the language tester is to collect evidence to support test use and interpretation that a larger community – the stakeholders (students, testers, teachers and society) – accept. But this truth may change as new evidence comes to light. As James (1907: 88) put it, 'truth *happens* to an idea' through a process, and 'its validity is the process of its valid-*ation*' (Italics in the original).

The language tester cannot point to facts and claim a test valid. There are many possible interpretations of facts. What he or she has to do is create an argument that best explains the facts available. It is interesting to note that we talk of validity 'arguments' – a topic that we return to in Unit 10. The word 'argument' implies that there will be disagreement, and that there will be other interpretations of the facts that challenge the validity argument. 'Disagreements are not settled by the facts, but are the means by which the facts are settled' (Fish, 1995: 253). This is entirely in keeping with, but an expansion of, Messick's (1989) view that at the heart of validity was investigating alternative hypotheses to explain evidence collected as part of the validation process.

In a pragmatic theory of validity, how would we decide whether an argument was *adequate* to support an intended use of a test? Peirce (undated: 4–5) has suggested that the kinds of arguments we construct in language testing may be evaluated through *abduction*, or what he later called *retroduction*. He explains that retroduction is:

> the process in which the mind goes over all the facts of the case, absorbs them, digests them, sleeps over them, assimilates them, dreams of them, and finally is prompted to deliver them in a form, which, if it adds something to them, does so not only because the addition serves to render intelligible what without it, is unintelligible. I have hitherto called this kind of reasoning which issues in explanatory hypotheses and the like, *abduction*, because I see reason to think that this is what Aristotle intended to denote by the corresponding Greek term 'apagoge' in the 25th chapter of the 2nd Book of his Analytics. But since this, after all, is only conjectural, I have on reflexion decided to give this kind of reasoning the name of *retroduction* to imply that it turns back and leads from the consequent of an admitted consequence, to its antecedent. Observe, if you please, the difference of

meaning between a *consequent*, the thing led to, and a *consequence*, the general fact by virtue of which a given antecedent leads to a certain *consequent*.

In short, we interpret facts to make them meaningful, working from the end to the explanation. In order to understand this more clearly, we will relate it to the stories of Sir Arthur Conan Doyle, for it is 'abduction' or 'retroduction' that is at the heart of every single Sherlock Holmes story ever written.

Other Principles of Language Assessment

PRACTICALITY

An effective test is **practical**. This means that it

- is not excessively expensive,
- stays within appropriate time constraints,
- is relatively easy to administer, and
- has a scoring/evaluation procedure that is specific and time-efficient.

A test that is prohibitively expensive is impractical. A test of language proficiency that takes a student five hours to complete is impractical—it consumes more time (and money) than necessary to accomplish its objective. A test that requires individual one-on-one proctoring is impractical for a group of several hundred test-takers and only a handful of examiners. A test that takes a few minutes for a student to take and several hours for an examiner to evaluate is impractical for most classroom situations. A test that can be scored only by computer is impractical if the test takes place a thousand miles away from the nearest computer. The value and quality of a test sometimes hinge on such nitty-gritty, practical considerations.

Here's a little horror story about practicality gone awry. An administrator of a six-week summertime short course needed to place the 50 or so students who had enrolled in the program. A quick search yielded a copy of an old English Placement Test from the University of Michigan. It had 20 listening items based on an audiotape and 80 items on grammar, vocabulary, and reading comprehension, all multiple-choice format. A scoring grid accompanied the test. On the day of the test, the required number of test booklets had been secured, a proctor had been assigned to monitor the process, and the administrator and proctor had planned to have the scoring completed by later that afternoon so students could begin classes the next day. Sounds simple, right? Wrong.

The students arrived, test booklets were distributed, and directions were given. The proctor started the tape. Soon students began to look puzzled. By the time the tenth item played, everyone looked bewildered. Finally, the proctor checked a test booklet and was horrified to discover that the wrong tape was playing; it was a tape for another form of the same test! Now what? She decided to randomly select a short passage from a textbook that was in the room and give the students a dictation. The students responded reasonably well. The next 80 non-tape-based items proceeded without incident, and the students handed in their score sheets and dictation papers.

When the red-faced administrator and the proctor got together later to score the tests, they faced the problem of how to score the dictation—a more subjective process than some other forms of assessment (see Chapter 6). After a lengthy exchange, the two established a point system, but after the first few papers had been scored, it was clear that the point system needed revision. That meant going back to the first papers to make sure the new system was followed.

The two faculty members had barely begun to score the 80 multiple-choice items when students began returning to the office to receive their placements. Students were told to come back the next morning for their results. Later that evening, having combined dictation scores and the 80-item multiple-choice scores, the two frustrated examiners finally arrived at placements for all students.

It's easy to see what went wrong here. While the listening comprehension section of the test was apparently highly practical, the administrator had failed to check the materials ahead of time (which, as you will see below, is a factor that touches on unreliability as well). Then, they established a scoring procedure that did not fit into the time constraints. In classroom-based testing, time is almost always a crucial practicality factor for busy teachers with too few hours in the day!

RELIABILITY

A **reliable** test is consistent and dependable. If you give the same test to the same student or matched students on two different occasions, the test should yield similar results. The issue of reliability of a test may best be addressed by considering a number of factors that may contribute to the unreliability of a test. Consider the

following possibilities (adapted from Mousavi, 2002, p. 804): fluctuations in the student, in scoring, in test administration, and in the test itself.

Student-Related Reliability

The most common learner-related issue in reliability is caused by temporary illness, fatigue, a "bad day," anxiety, and other physical or psychological factors, which may make an "observed" score deviate from one's "true" score. Also included in this category are such factors as a test-taker's "test-wiseness" or strategies for efficient test taking (Mousavi, 2002, p. 804).

Rater Reliability

Human error, subjectivity, and bias may enter into the scoring process. **Inter-rater reliability** occurs when two or more scorers yield inconsistent scores of the same test, possibly for lack of attention to scoring criteria, inexperience, inattention, or even preconceived biases. In the story above about the placement test, the initial scoring plan for the dictations was found to be unreliable—that is, the two scorers were not applying the same standards.

Rater-reliability issues are not limited to contexts where two or more scorers are involved. **Intra-rater reliability** is a common occurrence for classroom teachers because of unclear scoring criteria, fatigue, bias toward particular "good" and "bad" students, or simple carelessness. When I am faced with up to 40 tests to grade in only a week, I know that the standards I apply—however subliminally—to the first few tests will be different from those I apply to the last few. I may be "easier" or "harder" on those first few papers or I may get tired, and the result may be an inconsistent evaluation across all tests. One solution to such intra-rater unreliability is to read through about half of the tests before rendering any final scores or grades, then to recycle back through the whole set of tests to ensure an even-handed judgment. In tests of writing skills, rater reliability is particularly hard to achieve since writing proficiency involves numerous traits that are difficult to define. The careful specification of an analytical scoring instrument, however, can increase rater reliability (J. D. Brown, 1991).

Test Administration Reliability

Unreliability may also result from the conditions in which the test is administered. I once witnessed the administration of a test of aural comprehension in which a tape recorder played items for comprehension, but because of street noise outside the building, students sitting next to windows could not hear the tape accurately. This was a clear case of unreliability caused by the conditions of the test administration. Other sources of unreliability are found in photocopying variations, the amount of light in different parts of the room, variations in temperature, and even the condition of desks and chairs.

Test Reliability

Sometimes the nature of the test itself can cause measurement errors. If a test is too long, test-takers may become fatigued by the time they reach the later items and hastily respond incorrectly. Timed tests may discriminate against students who do not perform well on a test with a time limit. We all know people (and you may be included in this category!) who "know" the course material perfectly but who are adversely affected by the presence of a clock ticking away. Poorly written test items (that are ambiguous or that have more than one correct answer) may be a further source of test unreliability.

AUTHENTICITY

A fourth major principle of language testing is **authenticity,** a concept that is a little slippery to define, especially within the art and science of evaluating and designing tests. Bachman and Palmer (1996, p. 23) define authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a target language task," and then suggest an agenda for identifying those target language tasks and for transforming them into valid test items.

Essentially, when you make a claim for authenticity in a test task, you are saying that this task is likely to be enacted in the "real world." Many test item types fail to simulate real-world tasks. They may be contrived or artificial in their attempt to target a grammatical form or a lexical item. The sequencing of items that bear no relationship to one another lacks authenticity. One does not have to look very long to find reading comprehension passages in proficiency tests that do not reflect a real-world passage.

In a test, authenticity may be present in the following ways:

- The language in the test is as natural as possible.
- Items are contextualized rather than isolated.
- Topics are meaningful (relevant, interesting) for the learner.
- Some thematic organization to items is provided, such as through a story line or episode.
- Tasks represent, or closely approximate, real-world tasks.

The authenticity of test tasks in recent years has increased noticeably. Two or three decades ago, unconnected, boring, contrived items were accepted as a necessary component of testing. Things have changed. It was once assumed that large-scale testing could not include performance of the productive skills and stay within budgetary constraints, but now many such tests offer speaking and writing components. Reading passages are selected from real-world sources that test-takers are likely to have encountered or will encounter. Listening comprehension sections feature natural language with hesitations, white noise, and interruptions. More and more tests offer items that are "episodic" in that they are sequenced to form meaningful units, paragraphs, or stories.

You are invited to take up the challenge of authenticity in your classroom tests. As we explore many different types of task in this book, especially in Chapters 6 through 9, the principle of authenticity will be very much in the forefront.

WASHBACK

A facet of consequential validity, discussed above, is "the effect of testing on teaching and learning" (Hughes, 2003, p. 1), otherwise known among language-testing specialists as **washback**. In large-scale assessment, washback generally refers to the effects the tests have on instruction in terms of how students prepare for the test.

"Cram" courses and "teaching to the test" are examples of such washback. Another form of washback that occurs more in classroom assessment is the information that "washes back" to students in the form of useful diagnoses of strengths and weaknesses. Washback also includes the effects of an assessment on teaching and learning prior to the assessment itself, that is, on preparation for the assessment. Informal performance assessment is by nature more likely to have built-in washback effects because the teacher is usually providing interactive feedback. Formal tests can also have positive washback, but they provide no washback if the students receive a simple letter grade or a single overall numerical score.

The challenge to teachers is to create classroom tests that serve as learning devices through which washback is achieved. Students' incorrect responses can become windows of insight into further work. Their correct responses need to be praised, especially when they represent accomplishments in a student's interlanguage. Teachers can suggest strategies for success as part of their "coaching" role. Washback enhances a number of basic principles of language acquisition: intrinsic motivation, autonomy, self-confidence, language ego, interlanguage, and strategic investment, among others. (See *PLLT* and *TBP* for an explanation of these principles.)

One way to enhance washback is to comment generously and specifically on test performance. Many overworked (and underpaid!) teachers return tests to students with a single letter grade or numerical score and consider their job done. In reality, letter grades and numerical scores give absolutely no information of intrinsic interest to the student. Grades and scores reduce a mountain of linguistic and cognitive performance data to an absurd molehill. At best, they give a relative indication of a formulaic judgment of performance as compared to others in the class—which fosters competitive, not cooperative, learning.

With this in mind, when you return a written test or a data sheet from an oral production test, consider giving more than a number, grade, or phrase as your feedback. Even if your evaluation is not a neat little paragraph appended to the test, you can respond to as many details throughout the test as time will permit. Give praise for strengths—the "good stuff"—as well as constructive criticism of weaknesses. Give strategic hints on how a student might improve certain elements of performance. In other words, take some time to make the test performance an intrinsically motivating experience from which a student will gain a sense of accomplishment and challenge.

A little bit of washback may also help students through a specification of the numerical scores on the various subsections of the test. A subsection on verb tenses, for example, that yields a relatively low score may serve the diagnostic purpose of showing the student an area of challenge.

Another viewpoint on washback is achieved by a quick consideration of differences between **formative** and **summative** tests, mentioned in Chapter 1. Formative tests, by definition, provide washback in the form of information to the learner on progress toward goals. But teachers might be tempted to feel that summative tests, which provide assessment at the end of a course or program, do not need to offer much in the way of washback. Such an attitude is unfortunate because the end of

every language course or program is always the beginning of further pursuits, more learning, more goals, and more challenges to face. Even a final examination in a course should carry with it some means for giving washback to students.

In my courses I never give a final examination as the last scheduled classroom session. I always administer a final exam during the penultimate session, then complete the evaluation of the exams in order to return them to students during the last class. At this time, the students receive scores, grades, and comments on their work, and I spend some of the class session addressing material on which the students were not completely clear. My summative assessment is thereby enhanced by some beneficial washback that is usually not expected of final examinations.

Finally, washback also implies that students have ready access to you to discuss the feedback and evaluation you have given. While you almost certainly have known teachers with whom you wouldn't dare argue about a grade, an interactive, cooperative, collaborative classroom nevertheless can promote an atmosphere of dialogue between students and teachers regarding evaluative judgments. For learning to continue, students need to have a chance to feed back on your feedback, to seek clarification of any issues that are fuzzy, and to set new and appropriate goals for themselves for the days and weeks ahead.

ASSESSING LISTENING

OBSERVING THE PERFORMANCE OF THE FOUR SKILLS

Before focusing on listening itself, think about the two interacting concepts of **performance** and **observation**. All language users perform the acts of listening, speaking, reading, and writing. They of course rely on their underlying competence in order to accomplish these performances. When you propose to assess someone's ability in one or a combination of the four skills, you assess that person's *competence,* but you observe the person's *performance.* Sometimes the performance does not indicate true competence: a bad night's rest, illness, an emotional distraction, test anxiety, a memory block, or other student-related reliability factors could affect performance, thereby providing an unreliable measure of actual competence.

So, one important principle for assessing a learner's competence is to consider the fallibility of the results of a single performance, such as that produced in a test. As with any attempt at measurement, it is your obligation as a teacher to **triangulate** your measurements: consider at least two (or more) performances and/or contexts before drawing a conclusion. That could take the form of one or more of the following designs:

- several tests that are combined to form an assessment
- a single test with multiple test tasks to account for learning styles and performance variables
- in-class and extra-class graded work
- alternative forms of assessment (e.g., journal, portfolio, conference, observation, self-assessment, peer-assessment).

Multiple measures will always give you a more reliable and valid assessment than a single measure.

A second principle is one that we teachers often forget. We must rely as much as possible on *observable* performance in our assessments of students. Observable means being able to see or hear the performance of the learner (the senses of touch, taste, and smell don't apply very often to language testing!). What, then, is observable among the four skills of listening, speaking, reading, and writing? Table 6.1 offers an answer.

Isn't it interesting that in the case of the receptive skills, we can observe neither the process of performing nor a product? I can hear your argument already: "But I can *see* that she's listening because she's nodding her head and frowning and smiling and asking relevant questions." Well, you're not observing the listening performance; you're observing the *result* of the listening. You can no more observe listening (or reading) than you can see the wind blowing. The process of

Table 6.1. Observable performance of the four skills

	Can the teacher directly observe . . .	
	the process?	the product?
Listening	No	No
Speaking	Yes	No*
Reading	No	No
Writing	Yes	Yes

*Except in the case of an audio or video recording that preserves the output.

the listening performance itself is the *invisible, inaudible* process of internalizing meaning from the auditory signals being transmitted to the ear and brain. Or you may argue that the product of listening is a spoken or written response from the student that indicates correct (or incorrect) auditory processing. Again, the product of listening and reading is not the spoken or written response. The product is within the structure of the brain, and until teachers carry with them little portable MRI scanners to detect meaningful intake, it is impossible to observe the product. You observe only the result of the meaningful input in the form of spoken or written output, just as you observe the result of the wind by noticing trees waving back and forth.

The productive skills of speaking and writing allow us to hear and see the process as it is performed. Writing gives a permanent product in the form of a written piece. But unless you have recorded speech, there is no *permanent* observable product for speaking performance because all those words you just heard have vanished from your perception and (you hope) have been transformed into meaningful intake somewhere in your brain.

Receptive skills, then, are clearly the more enigmatic of the two modes of performance. You cannot observe the actual act of listening or reading, nor can you see or hear an actual product! You can observe learners only *while* they are listening or reading. The upshot is that all assessment of listening and reading must be made on the basis of observing the test-taker's speaking or writing (or nonverbal response), and not on the listening or reading itself. So, all assessment of receptive performance must be made by inference!

How discouraging, right? Well, not necessarily. We have developed reasonably good assessment tasks to make the necessary jump, through the process of inference, from unobservable reception to a conclusion about comprehension competence. And all this is a good reminder of the importance not just of triangulation but of the potential fragility of the assessment of comprehension ability. The actual performance is made "behind the scenes," and those of us who propose to make reliable assessments of receptive performance need to be on our guard.

THE IMPORTANCE OF LISTENING

Listening has often played second fiddle to its counterpart, speaking. In the standardized testing industry, a number of separate oral production tests are available (Test of Spoken English, Oral Proficiency Inventory, and PhonePass®, to name several that are described Chapter 7 of this book), but it is rare to find just a listening test. One reason for this emphasis is that listening is often implied as a component of speaking. How could you speak a language without also listening? In addition, the overtly observable nature of speaking renders it more empirically measurable then listening. But perhaps a deeper cause lies in universal biases toward speaking. A good speaker is often (unwisely) valued more highly than a good listener. To determine if someone is a proficient user of a language, people customarily ask, "Do you speak Spanish?" People rarely ask, "Do you *understand* and speak Spanish?"

Every teacher of language knows that one's oral production ability—other than monologues, speeches, reading aloud, and the like—is only as good as one's listening comprehension ability. But of even further impact is the likelihood that *input* in the aural-oral mode accounts for a large proportion of successful language acquisition. In a typical day, we do measurably more listening than speaking (with the exception of one or two of your friends who may be nonstop chatterboxes!). Whether in the workplace, educational, or home contexts, aural comprehension far outstrips oral production in quantifiable terms of time, number of words, effort, and attention.

We therefore need to pay close attention to listening as a mode of performance for assessment in the classroom. In this chapter, we will begin with basic principles and types of listening, then move to a survey of tasks that can be used to assess listening. (For a review of issues in teaching listening, you may want to read Chapter 16 of *TBP*.)

BASIC TYPES OF LISTENING

As with all effective tests, designing appropriate assessment tasks in listening begins with the specification of objectives, or criteria. Those objectives may be classified in terms of several types of listening performance. Think about what you do when you listen. Literally in nanoseconds, the following processes flash through your brain:

1. You recognize speech sounds and hold a temporary "imprint" of them in short-term memory.
2. You simultaneously determine the type of speech event (monologue, interpersonal dialogue, transactional dialogue) that is being processed and attend to its context (who the speaker is, location, purpose) and the content of the message.
3. You use (bottom-up) linguistic decoding skills and/or (top-down) background schemata to bring a plausible interpretation to the message, and assign a *literal* and *intended meaning* to the utterance.

4. In most cases (except for repetition tasks, which involve short-term memory only), you delete the exact linguistic form in which the message was originally received in favor of conceptually retaining important or relevant information in long-term memory.

Each of these stages represents a potential assessment objective:

- comprehending of surface structure elements such as phonemes, words, intonation, or a grammatical category
- understanding of pragmatic context
- determining meaning of auditory input
- developing the gist, a global or comprehensive understanding

From these stages we can derive four commonly identified types of listening performance, each of which comprises a category within which to consider assessment tasks and procedures.

1. *Intensive.* Listening for perception of the components (phonemes, words, intonation, discourse markers, etc.) of a larger stretch of language.
2. *Responsive.* Listening to a relatively short stretch of language (a greeting, question, command, comprehension check, etc.) in order to make an equally short response.
3. *Selective.* Processing stretches of discourse such as short monologues for several minutes in order to "scan" for certain information. The purpose of such performance is not necessarily to look for global or general meanings, but to be able to comprehend designated information in a context of longer stretches of spoken language (such as classroom directions from a teacher, TV or radio news items, or stories). Assessment tasks in selective listening could ask students, for example, to listen for names, numbers, a grammatical category, directions (in a map exercise), or certain facts and events.
4. *Extensive.* Listening to develop a top-down, global understanding of spoken language. Extensive performance ranges from listening to lengthy lectures to listening to a conversation and deriving a comprehensive message or purpose. Listening for the gist, for the main idea, and making inferences are all part of extensive listening.

For full comprehension, test-takers may at the extensive level need to invoke **interactive** skills (perhaps note-taking, questioning, discussion): listening that includes all four of the above types as test-takers actively participate in discussions, debates, conversations, role plays, and pair and group work. Their listening performance must be intricately integrated with speaking (and perhaps other skills) in the authentic give-and-take of communicative interchange. (Assessment of interactive skills will be embedded in Chapter 7.)

MICRO- AND MACROSKILLS OF LISTENING

A useful way of synthesizing the above two lists is to consider a finite number of micro- and macroskills implied in the performance of listening comprehension. Richards' (1983) list of microskills has proven useful in the domain of specifying objectives for learning and may be even more useful in forcing test makers to carefully identify specific assessment objectives. In the following box, the skills are subdivided into what I prefer to think of as microskills (attending to the smaller bits and chunks of language, in more of a bottom-up process) and macroskills (focusing on the larger elements involved in a top-down approach to a listening task).The micro- and macroskills provide 17 different objectives to assess in listening.

Micro- and macroskills of listening (adapted from Richards, 1983)

Microskills

1. Discriminate among the distinctive sounds of English.
2. Retain chunks of language of different lengths in short-term memory.
3. Recognize English stress patterns, words in stressed and unstressed positions, rhythmic structure, intonation contours, and their role in signaling information.
4. Recognize reduced forms of words.
5. Distinguish word boundaries, recognize a core of words, and interpret word order patterns and their significance.
6. Process speech at different rates of delivery.
7. Process speech containing pauses, errors, corrections, and other performance variables.
8. Recognize grammatical word classes (nouns, verbs, etc.), systems (e.g., tense, agreement, pluralization), patterns, rules, and elliptical forms.
9. Detect sentence constituents and distinguish between major and minor constituents.
10. Recognize that a particular meaning may be expressed in different grammatical forms.
11. Recognize cohesive devices in spoken discourse.

Macroskills

12. Recognize the communicative functions of utterances, according to situations, participants, goals.
13. Infer situations, participants, goals using real-world knowledge.
14. From events, ideas, and so on, described, predict outcomes, infer links and connections between events, deduce causes and effects, and detect such relations as main idea, supporting idea, new information, given information, generalization, and exemplification.

15. Distinguish between literal and implied meanings.
16. Use facial, kinesic, body language, and other nonverbal clues to decipher meanings.
17. Develop and use a battery of listening strategies, such as detecting key words, guessing the meaning of words from context, appealing for help, and signaling comprehension or lack thereof.

Implied in the taxonomy above is a notion of what makes many aspects of listening difficult, or why listening is not simply a linear process of recording strings of language as they are transmitted into our brains. Developing a sense of which aspects of listening performance are predictably difficult will help you to challenge your students appropriately and to assign weights to items. Consider the following list of what makes listening difficult (adapted from Richards, 1983; Ur, 1984; Dunkel, 1991):

1. *Clustering:* attending to appropriate "chunks" of language—phrases, clauses, constituents
2. *Redundancy:* recognizing the kinds of repetitions, rephrasing, elaborations, and insertions that unrehearsed spoken language often contains, and benefiting from that recognition
3. *Reduced forms:* understanding the reduced forms that may not have been a part of an English learner's past learning experiences in classes where only formal "textbook" language has been presented
4. *Performance variables:* being able to "weed out" hesitations, false starts, pauses, and corrections in natural speech
5. *Colloquial language:* comprehending idioms, slang, reduced forms, shared cultural knowledge
6. *Rate of delivery:* keeping up with the speed of delivery, processing automatically as the speaker continues
7. *Stress, rhythm, and intonation:* correctly understanding prosodic elements of spoken language, which is almost always much more difficult than understanding the smaller phonological bits and pieces
8. *Interaction:* managing the interactive flow of language from listening to speaking to listening, etc.

DESIGNING ASSESSMENT TASKS: INTENSIVE LISTENING

Once you have determined objectives, your next step is to design the tasks, including making decisions about how you will elicit performance and how you will expect the test-taker to respond. We will look at tasks that range from intensive listening performance, such as minimal phonemic pair recognition, to extensive comprehension of language in communicative contexts. The focus in this section is on the microskills of intensive listening.

Recognizing Phonological and Morphological Elements

A typical form of intensive listening at this level is the assessment of recognition of phonological and morphological elements of language. A classic test task gives a spoken stimulus and asks test-takers to identify the stimulus from two or more choices, as in the following two examples:

Phonemic pair, consonants

Test-takers hear:	He's from California.
Test-takers read:	(a) He's from California. (b) She's from California.

Phonemic pair, vowels

Test-takers hear:	Is he living?
Test-takers read:	(a) Is he leaving? (b) Is he living?

In both cases above, minimal phonemic distinctions are the target. If you are testing recognition of morphology, you can use the same format:

Morphological pair, -ed ending

Test-takers hear:	I missed you very much.
Test-takers read:	(a) I missed you very much. (b) I miss you very much.

Hearing the past tense morpheme in this sentence challenges even advanced learners, especially if no context is provided. Stressed and unstressed words may also be tested with the same rubric. In the following example, the reduced form (contraction) of *can not* is tested:

Stress pattern in can't

Test-takers hear:	My girlfriend can't go to the party.
Test-takers read:	(a) My girlfriend can't go to the party. (b) My girlfriend can go to the party.

Because they are decontextualized, these kinds of tasks leave something to be desired in their authenticity. But they are a step better than items that simply provide a one-word stimulus:

One-word stimulus

Test-takers hear:	vine
Test-takers read:	(a) vine (b) wine

Paraphrase Recognition

The next step up on the scale of listening comprehension microskills is words, phrases, and sentences, which are frequently assessed by providing a stimulus sentence and asking the test-taker to choose the correct paraphrase from a number of choices.

Sentence paraphrase

Test-takers hear:	Hellow, my name's Keiko. I come from Japan.
Test-takers read:	(a) Keiko is comfortable in Japan. (b) Keiko wants to come to Japan. (c) Keiko is Japanese. (d) Keiko likes Japan.

In the above item, the idiomatic *come from* is the phrase being tested. To add a little context, a conversation can be the stimulus task to which test-takers must respond with the correct paraphrase:

Dialogue paraphrase

Test-takers hear:	Man:	Hi, Maria, my name's George.
	Woman:	Nice to meet you, George. Are you American?
	Man:	No, I'm Canadian.
Test-takers read:		(a) George lives in the United States. (b) George is American. (c) George comes from Canada. (d) Maria is Canadian.

Here, the criterion is recognition of the adjective form used to indicate country of origin: Canadian, American, Brazilian, Italian, etc.

DESIGNING ASSESSMENT TASKS: RESPONSIVE LISTENING

A question-and-answer format can provide some interactivity in these lower-end listening tasks. The test-taker's response is the appropriate answer to a question.

Appropriate response to a question

Test-takers hear:	How much time did you take to do your homework?
Test-takers read:	(a) In about an hour. (b) About an hour. (c) About $10. (d) Yes, I did.

The objective of this item is recognition of the *wh*-question *how much* and its appropriate response. Distractors are chosen to represent common learner errors: (a) responding to *how much* vs. *how much longer;* (c) confusing *how much* in reference to time vs. the more frequent reference to money; (d) confusing a *wh*-question with a *yes/no* question.

None of the tasks so far discussed have to be framed in a multiple-choice format. They can be offered in a more open-ended framework in which test-takers write or speak the response. The above item would then look like this:

Open-ended response to a question

Test-takers hear:	How much time did you take to do your homework?
Test-takers write or speak:	_____.

If open-ended response formats gain a small amount of authenticity and creativity, they of course suffer some in their practicality, as teachers must then read students' responses and judge their appropriateness, which takes time.

DESIGNING ASSESSMENT TASKS: SELECTIVE LISTENING

A third type of listening performance is **selective** listening, in which the test-taker listens to a limited quantity of aural input and must discern within it some specific information. A number of techniques have been used that require selective listening.

Listening Cloze

Listening cloze tasks (sometimes called **cloze dictations** or **partial dictations**) require the test-taker to listen to a story, monologue, or conversation and simultaneously

read the written text in which selected words or phrases have been deleted. **Cloze procedure** is most commonly associated with reading only (see Chapter 9). In its generic form, the test consists of a passage in which every *n*th word (typically every seventh word) is deleted and the test-taker is asked to supply an appropriate word. In a listening cloze task, test-takers see a transcript of the passage that they are listening to and fill in the blanks with the words or phrases that they hear.

One potential weakness of listening cloze techniques is that they may simply become reading comprehension tasks. Test-takers who are asked to listen to a story with periodic deletions in the written version may not need to listen at all, yet may still be able to respond with the appropriate word or phrase. You can guard against this eventuality if the blanks are items with high information load that cannot be easily predicted simply by reading the passage. In the example below (adapted from Bailey, 1998, p. 16), such a shortcoming was avoided by focusing only on the criterion of numbers. Test-takers hear an announcement from an airline agent and see the transcript with the underlined words deleted:

Listening cloze

Test-takers hear:

Ladies and gentlemen, I now have some connecting gate information for those of you making connections to other flights out of San Francisco.

Flight *seven-oh-six* to Portland will depart from gate *seventy-three* at *nine-thirty* P.M.
Flight *ten-forty-five* to Reno will depart at *nine-fifty* P.M. from gate *seventeen*.
Flight *four-forty* to Monterey will depart at *nine-thirty-five* P.M. from gate *sixty*.
And flight *sixteen-oh-three* to Sacramento will depart from gate *nineteen* at *ten-fifteen* P.M.

Test-takers write the missing words or phrases in the blanks.

Other listening cloze tasks may focus on a grammatical category such as verb tenses, articles, two-word verbs, prepositions, or transition words/phrases. Notice two important structural differences between listening cloze tasks and standard reading cloze. In a listening cloze, deletions are governed by the objective of the test, not by mathematical deletion of every *n*th word; and more than one word may be deleted, as in the above example.

Listening cloze tasks should normally use an **exact word** method of scoring, in which you accept as a correct response only the actual word or phrase that was spoken and consider other **appropriate words** as incorrect. (See Chapter 8 for further discussion of these two methods.) Such stringency is warranted; your objective is, after all, to test listening comprehension, not grammatical or lexical expectancies.

Information Transfer

Selective listening can also be assessed through an **information transfer** technique in which aurally processed information must be transferred to a visual representation, such as labeling a diagram, identifying an element in a picture, completing a form, or showing routes on a map.

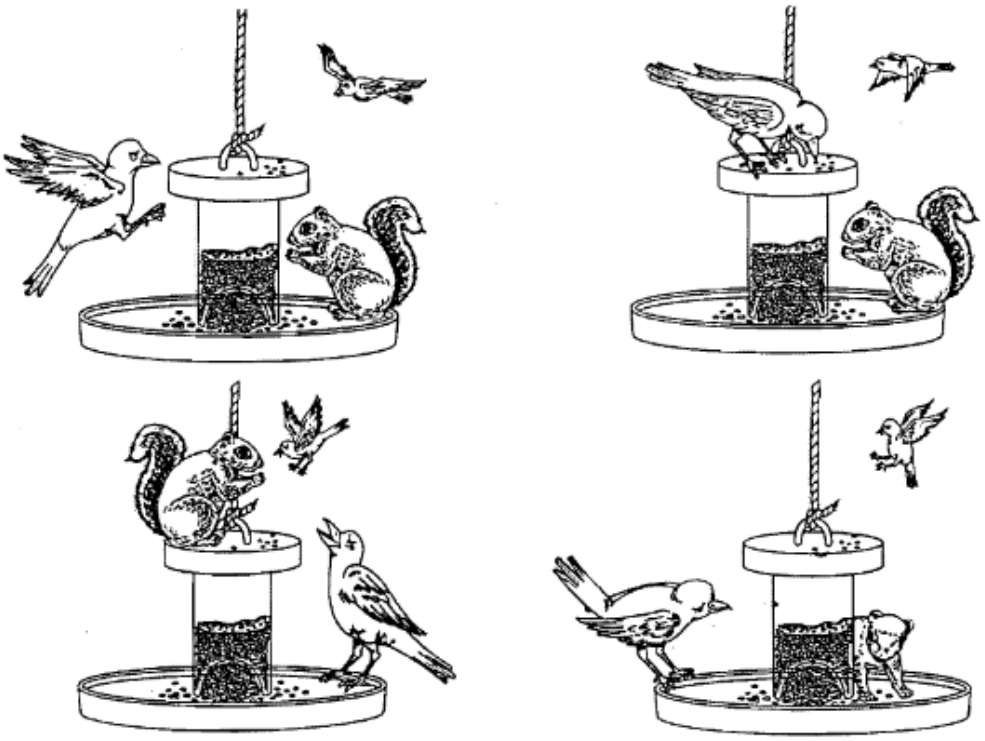
At the lower end of the scale of linguistic complexity, simple **picture-cued** items are sometimes efficient rubrics for assessing certain selected information. Consider the following item:

Information transfer: multiple-picture–cued selection

Test-takers hear:

Choose the correct picture. In my back yard I have a bird feeder. Yesterday, there were two birds and a squirrel fighting for the last few seeds in the bird feeder. The squirrel was on top of the bird feeder while the larger bird sat at the bottom of the feeder screeching at the squirrel. The smaller bird was flying around the squirrel, trying to scare it away.

Test-takers see:



The preceding example illustrates the need for test-takers to focus on just the relevant information. The objective of this task is to test prepositions and prepositional phrases of location (*at the bottom, on top of, around,* along with *larger, smaller*), so other words and phrases such as *back yard, yesterday, last few seeds,* and *scare away* are supplied only as context and need not be tested. (The task also presupposes, of course, that test-takers are able to identify the difference between a bird and a squirrel!)

In another genre of picture-cued tasks, a number of people and/or actions are presented in one picture, such as a group of people at a party. Assuming that all the items, people, and actions are clearly depicted and understood by the test-taker, assessment may take the form of

- questions: "Is the tall man near the door talking to a short woman?"
- true/false: "The woman wearing a red skirt is watching TV."
- identification: "Point to the person who is standing behind the lamp." "Draw a circle around the person to the left of the couch."

In a third picture-cued option used by the Test of English for International Communication (TOEIC®), one single photograph is presented to the test-taker, who then hears four different statements and must choose one of the four to describe the photograph. Here is an example.

Information transfer: single-picture–cued verbal multiple-choice

Test-takers see:	a photograph of a woman in a laboratory setting, with no glasses on, squinting through a microscope with her right eye, and with her left eye closed.
Test-takers hear:	(a) She's speaking into a microphone. (b) She's putting on her glasses. (c) She has both eyes open. (d) She's using a microscope.

Information transfer tasks may reflect greater authenticity by using charts, maps, grids, timetables, and other artifacts of daily life. In the example below, test-takers hear a student's daily schedule, and the task is to fill in the partially completed weekly calendar.

Information transfer: chart-filling

Test-takers hear:

Now you will hear information about Lucy's daily schedule. The information will be given twice. The first time just listen carefully. The second time, there will be a pause after each sentence. Fill in Lucy's blank daily schedule with the correct information. The example has already been filled in.

You will hear: Lucy gets up at eight o'clock every morning except on weekends.

You will fill in the schedule to provide the information.

Now listen to the information about Lucy's schedule. Remember, you will first hear all the sentences; then you will hear each sentence separately with time to fill in your chart.

Lucy gets up at 8:00 every morning except on weekends. She has English on Monday, Wednesday, and Friday at ten o'clock. She has History on Tuesdays and Thursdays at two o'clock. She takes Chemistry on Monday from two o'clock to six o'clock. She plays tennis on weekends at four o'clock. She eats lunch at twelve o'clock every day except Saturday and Sunday.

Now listen a second time. There will be a pause after each sentence to give you time to fill in the chart. (Lucy's schedule is repeated with a pause after each sentence).

Test-takers see the following weekly calendar grid:

	Monday	Tuesday	Wednesday	Thursday	Friday	Weekends
8:00	get up	get up	get up	get up	get up	
10:00						
12:00						
2:00						
4:00						
6:00						

Such chart-filling tasks are good examples of aural **scanning** strategies. A listener must discern from a number of pieces of information which pieces are relevant. In the above example, virtually all of the stimuli are relevant, and very few words can be ignored. In other tasks, however, much more information might be presented than is needed (as in the birdfeeder item on page 127), forcing the test-taker to select the correct bits and pieces necessary to complete a task.

Chart-filling tasks increase in difficulty as the linguistic stimulus material becomes more complex. In one task described by Ur (1984, pp. 108–112), test-takers listen to a very long description of animals in various cages in a zoo. While they listen, they can look at a map of the layout of the zoo with unlabeled cages. Their task is to fill in the correct animal in each cage, but the complexity of the language used to describe the positions of cages and their inhabitants is very challenging. Similarly, Hughes (1989, p. 138) described a map-marking task in which test-takers must process around 250 words of colloquial language in order to complete the tasks of identifying names, positions, and directions in a car accident scenario on a city street.

Sentence Repetition

The task of simply repeating a sentence or a partial sentence, or **sentence repetition,** is also used as an assessment of listening comprehension. As in a dictation (discussed below), the test-taker must retain a stretch of language long enough to reproduce it, and then must respond with an oral repetition of that stimulus. Incorrect listening comprehension, whether at the phonemic or discourse level, may be manifested in the correctness of the repetition. A miscue in repetition is scored as a miscue in listening. In the case of somewhat longer sentences, one could argue that the ability to recognize and retain chunks of language as well as threads of meaning might be assessed through repetition. In Chapter 7, we will look closely at PhonePass, a commercially produced test that relies largely on sentence repetition to assess both oral production and listening comprehension.

Sentence repetition is far from a flawless listening assessment task. Buck (2001, p. 79) noted that such tasks "are not just tests of listening, but tests of general oral skills." Further, this task may test only recognition of sounds, and it can easily be contaminated by lack of short-term memory ability, thus invalidating it as an assessment of comprehension alone. And the teacher may never be able to distinguish a listening comprehension error from an oral production error. Therefore, sentence repetition tasks should be used with caution.

DESIGNING ASSESSMENT TASKS: EXTENSIVE LISTENING

Drawing a clear distinction between any two of the categories of listening referred to here is problematic, but perhaps the fuzziest division is between selective and extensive listening. As we gradually move along the continuum from smaller to larger stretches of language, and from micro- to macroskills of listening, the probability of using more extensive listening tasks increases. Some important questions about designing assessments at this level emerge.

1. Can listening performance be distinguished from cognitive processing factors such as memory, associations, storage, and recall?
2. As assessment procedures become more communicative, does the task take into account test-takers' ability to use grammatical expectancies, lexical collocations, semantic interpretations, and pragmatic competence?
3. Are test tasks themselves correspondingly content valid and authentic—that is, do they mirror real-world language and context?
4. As assessment tasks become more and more open-ended, they more closely resemble pedagogical tasks, which leads one to ask what the difference is between assessment and teaching tasks. The answer is *scoring:* the former imply specified scoring procedures, while the latter do not.

We will try to address these questions as we look at a number of extensive or quasi-extensive listening comprehension tasks.

Dictation

Dictation is a widely researched genre of assessing listening comprehension. In a dictation, test-takers hear a passage, typically of 50 to 100 words, recited three times: first, at normal speed; then, with long pauses between phrases or natural word groups, during which time test-takers write down what they have just heard; and finally, at normal speed once more so they can check their work and proofread. Here is a sample dictation at the intermediate level of English.

Dictation

> *First reading (natural speed, no pauses, test-takers listen for gist):*
>
> The state of California has many geographical areas. On the western side is the Pacific Ocean with its beaches and sea life. The central part of the state is a large fertile valley. The southeast has a hot desert, and north and west have beautiful mountains and forests. Southern California is a large urban area populated by millions of people.
>
> *Second reading (slowed speed, pause at each // break, test-takers write):*
>
> The state of California // has many geographical areas. // On the western side // is the Pacific Ocean // with its beaches and sea life. // The central part of the state // is a large fertile valley. // The southeast has a hot desert, // and north and west // have beautiful mountains and forests. // Southern California // is a large urban area // populated by millions of people.
>
> *Third reading (natural speed, test-takers check their work).*

Dictations have been used as assessment tools for decades. Some readers still cringe at the thought of having to render a correctly spelled, verbatim version of a paragraph or story recited by the teacher. Until research on integrative testing was published (see Oller, 1971), dictations were thought to be not much more than glorified spelling tests. However, the required integration of listening and writing in a dictation, along with its presupposed knowledge of grammatical and discourse expectancies, brought this technique back into vogue. Hughes (1989), Cohen (1994), Bailey (1998), and Buck (2001) all defend the plausibility of dictation as an integrative test that requires some sophistication in the language in order to process and write down all segments correctly. Thus, I include dictation here under the rubric of extensive tasks, although I am more comfortable with labeling it quasi-extensive.

The difficulty of a dictation task can be easily manipulated by the length of the word groups (or **bursts,** as they are technically called), the length of the pauses, the speed at which the text is read, and the complexity of the discourse, grammar, and vocabulary used in the passage.

Scoring is another matter. Depending on your context and purpose in administering a dictation, you will need to decide on scoring criteria for several possible kinds of errors:

- spelling error only, but the word appears to have been heard correctly
- spelling and/or obvious misrepresentation of a word, illegible word
- grammatical error (For example, test-taker hears *I can't do it,* writes *I can do it.*)
- skipped word or phrase
- permutation of words
- additional words not in the original
- replacement of a word with an appropriate synonym

Determining the weight of each of these errors is a highly idiosyncratic choice; specialists disagree almost more than they agree on the importance of the above categories. They do agree (Buck, 2001) that a dictation is not a spelling test, and that the first item in the list above should not be considered an error. They also suggest that point systems be kept simple (for maintaining practicality and reliability) and that a deductible scoring method, in which points are subtracted from a hypothetical total, is usually effective.

Dictation seems to provide a reasonably valid method for integrating listening and writing skills and for tapping into the cohesive elements of language implied in short passages. However, a word of caution lest you assume that dictation provides a quick and easy method of assessing extensive listening comprehension. If the bursts in a dictation are relatively long (more than five-word segments), this method places a certain amount of load on memory and processing of meaning (Buck, 2001, p. 78). But only a moderate degree of cognitive processing is required, and claiming that dictation fully assesses the ability to comprehend pragmatic or illocutionary elements of language, context, inference, or semantics may be going too far. Finally, one can easily question the authenticity of dictation: it is rare in the real world for people to write down more than a few chunks of information (addresses, phone numbers, grocery lists, directions, for example) at a time.

Despite these disadvantages, the practicality of the administration of dictations, a moderate degree of reliability in a well-established scoring system, and a strong correspondence to other language abilities speaks well for the inclusion of dictation among the possibilities for assessing extensive (or quasi-extensive) listening comprehension.

Communicative Stimulus-Response Tasks

Another—and more authentic—example of extensive listening is found in a popular genre of assessment task in which the test-taker is presented with a stimulus monologue or conversation and then is asked to respond to a set of comprehension questions. Such tasks (as you saw in Chapter 4 in the discussion of standardized testing) are commonly used in commercially produced proficiency tests. The monologues, lectures, and brief conversations used in such tasks are sometimes a little contrived,

and certainly the subsequent multiple-choice questions don't mirror communicative, real-life situations. But with some care and creativity, one can create reasonably authentic stimuli, and in some rare cases the response mode (as shown in one example below) actually approaches complete authenticity. Here is a typical example of such a task.

Dialogue and multiple-choice comprehension items

Test-takers hear:

Directions: Now you will hear a conversation between Lynn and her doctor. You will hear the conversation two times. After you hear the conversation the second time, choose the correct answer for questions 11–15 below. Mark your answers on the answer sheet provided.

Doctor:	Good morning, Lynn. What's the problem?
Lynn:	Well, you see, I have a terrible headache, my nose is running, and I'm really dizzy.
Doctor:	Okay. Anything else?
Lynn:	I've been coughing, I think I have a fever, and my stomach aches.
Doctor:	I see. When did this start?
Lynn:	Well, let's see, I went to the lake last weekend, and after I returned home I started sneezing.
Doctor:	Hmm. You must have the flu. You should get lots of rest, drink hot beverages, and stay warm. Do you follow me?
Lynn:	Well, uh, yeah, but . . . shouldn't I take some medicine?
Doctor:	Sleep and rest are as good as medicine when you have the flu.
Lynn:	Okay, thanks, Dr. Brown.

Test-takers read:

11. What is Lynn's problem?
(A) She feels horrible.
(B) She ran too fast at the lake.
(C) She's been drinking too many hot beverages.

12. When did Lynn's problem start?
(A) When she saw her doctor.
(B) Before she went to the lake.
(C) After she came home from the lake.

13. The doctor said that Lynn _____.
(A) flew to the lake last weekend
(B) must not get the flu
(C) probably has the flu

14. The doctor told Lynn _____.
(A) to rest
(B) to follow him
(C) to take some medicine

15. According to Dr. Brown, sleep and rest are _____ medicine when you have the flu.
(A) more effective than
(B) as effective as
(C) less effective than

Does this meet the criterion of authenticity? If you want to be painfully fussy, you might object that it is rare in the real world to eavesdrop on someone else's doctor-patient conversation. Nevertheless, the conversation itself is relatively authentic; we all have doctor-patient exchanges like this. Equally authentic, if you add a grain of salt, are monologues, lecturettes, and news stories, all of which are commonly utilized as listening stimuli to be followed by comprehension questions aimed at assessing certain objectives that are built into the stimulus.

Is the task itself (of responding to multiple-choice questions) authentic? It's plausible to assert that *any task* of this kind following a one-way listening to a conversation is artificial: we simply don't often encounter little quizzes about conversations we've heard (unless it's your parent, spouse, or best friend who wants to get in on the latest gossip!). The questions posed above, with the possible exception of **#14**, are unlikely to appear in a lifetime of doctor visits. Yet the ability to respond correctly to such items can be construct validated as an appropriate measure of field-independent listening skills: the ability to remember certain details from a conversation. (As an aside here, many highly proficient native speakers of English might miss some of the above questions if they heard the conversation only once and if they had no visual access to the items until after the conversation was done!)

To compensate for the potential inauthenticity of post-stimulus comprehension questions, you might, with a little creativity, be able to find contexts where questions that probe understanding are more appropriate. Consider the following situation:

Dialogue and authentic questions on details

Test-takers hear:
You will hear a conversation between a detective and a man. The tape will play the conversation twice. After you hear the conversation a second time, choose the correct answers on your test sheet.

Detective:	Where were you last night at eleven P.M., the time of the murder?
Man:	Uh, let's see, well, I was just starting to see a movie.
Detective:	Did you go alone?
Man:	No, uh, well, I was with my friend, uh, Bill. Yeah, I was with Bill.

Detective:	What did you do after that?
Man:	We went out to dinner, then I dropped her off at her place.
Detective:	Then you went home?
Man:	Yeah.
Detective:	When did you get home?
Man:	A little before midnight.

Test-takers read:

7. Where was the man at 11:00 P.M.?
   (A) In a restaurant.
   (B) In a theater.
   (C) At home.
8. Was he with someone?
   (A) He was alone.
   (B) He was with his wife.
   (C) He was with a friend.
9. Then what did he do?
   (A) He ate out.
   (B) He made dinner.
   (C) He went home.
10. When did he get home?
    (A) About 11:00.
    (B) Almost 12:00.
    (C) Right after the movie.
11. The man is probably lying because (name two clues):
    1. _____
    2. _____

In this case, test-takers are brought into a little scene in a crime story. The questions following are plausible questions that might be asked to review fact and fiction in the conversation. Question **#11**, of course, provides an extra shot of reality: the test-taker must name the probable lies told by the man (he referred to Bill as "her"; he saw a movie and ate dinner in the space of one hour), which requires the process of inference.

Authentic Listening Tasks

Ideally, the language assessment field would have a stockpile of listening test types that are cognitively demanding, communicative, and authentic, not to mention interactive by means of an integration with speaking. However, the nature of a test as a *sample* of performance and a set of tasks with limited time frames implies an equally limited capacity to mirror all the real-world contexts of listening performance. "There

is no such thing as a communicative test," stated Buck (2001, p. 92). "Every test requires some components of communicative language ability, and no test covers them all. Similarly, with the notion of authenticity, every task shares some characteristics with target-language tasks, and no test is completely authentic."

Beyond the rubrics of intensive, responsive, selective, and quasi-extensive communicative contexts described above, can we assess aural comprehension in a truly communicative context? Can we, at this end of the range of listening tasks, ascertain from test-takers that they have processed the main idea(s) of a lecture, the gist of a story, the pragmatics of a conversation, or the unspoken inferential data present in most authentic aural input? Can we assess a test-taker's comprehension of humor, idiom, and metaphor? The answer is a cautious yes, but not without some concessions to practicality. And the answer is a more certain yes if we take the liberty of stretching the concept of assessment to extend beyond tests and into a broader framework of alternatives. Here are some possibilities.

1. Note-taking. In the academic world, classroom lectures by professors are common features of a non-native English-user's experience. One form of a midterm examination at the American Language Institute at San Francisco State University (Kahn, 2002) uses a 15-minute lecture as a stimulus. One among several response formats includes note-taking by the test-takers. These notes are evaluated by the teacher on a 30-point system, as follows:

Scoring system for lecture notes

0–15 points
Visual representation: Are your notes clear and easy to read? Can you easily find and retrieve information from them? Do you use the space on the paper to visually represent ideas? Do you use indentation, headers, numbers, etc.?

0–10 points
Accuracy: Do you accurately indicate main ideas from lectures? Do you note important details and supporting information and examples? Do you leave out unimportant information and tangents?

0–5 points
Symbols and abbreviations: Do you use symbols and abbreviations as much as possible to save time? Do you avoid writing out whole words, and do you avoid writing down every single word the lecturer says?

The process of scoring is time consuming (a loss of practicality), and because of the subjectivity of the point system, it lacks some reliability. But the gain is in offering students an authentic task that mirrors exactly what they have been focusing on in the classroom. The notes become an indirect but arguably valid form of assessing global listening comprehension. The task fulfills the criteria of cognitive demand, communicative language, and authenticity.

2. *Editing.* Another authentic task provides both a written and a spoken stimulus, and requires the test-taker to listen for discrepancies. Scoring achieves relatively high reliability as there are usually a small number of specific differences that must be identified. Here is the way the task proceeds.

Editing a written version of an aural stimulus

> *Test-takers read:* the written stimulus material (a news report, an email from a friend, notes from a lecture, or an editorial in a newspaper).
>
> *Test-takers hear:* a spoken version of the stimulus that deviates, in a finite number of facts or opinions, from the original written form.
>
> *Test-takers mark:* the written stimulus by circling any words, phrases, facts, or opinions that show a discrepancy between the two versions.

One potentially interesting set of stimuli for such a task is the description of a political scandal first from a newspaper with a political bias, and then from a radio broadcast from an "alternative" news station. Test-takers are not only forced to listen carefully to differences but are subtly informed about biases in the news.

3. *Interpretive tasks.* One of the intensive listening tasks described above was paraphrasing a story or conversation. An interpretive task extends the stimulus material to a longer stretch of discourse and forces the test-taker to infer a response. Potential stimuli include

- song lyrics,
- [recited] poetry,
- radio/television news reports, and
- an oral account of an experience.

Test-takers are then directed to interpret the stimulus by answering a few questions (in open-ended form). Questions might be:

- "Why was the singer feeling sad?"
- "What events might have led up to the reciting of this poem?"
- "What do you think the political activists might do next, and why?"
- "What do you think the storyteller felt about the mysterious disappearance of her necklace?"

This kind of task moves us away from what might traditionally be considered a test toward an informal assessment, or possibly even a pedagogical technique or activity. But the task conforms to certain time limitations, and the questions can be quite specific, even though they ask the test-taker to use inference. While reliable scoring may be an issue (there may be more than one correct interpretation), the authenticity of

the interaction in this task and potential washback to the student surely give it some prominence among communicative assessment procedures.

4. *Retelling.* In a related task, test-takers listen to a story or news event and simply retell it, or summarize it, either orally (on an audiotape) or in writing. In so doing, test-takers must identify the gist, main idea, purpose, supporting points, and/or conclusion to show full comprehension. Scoring is partially predetermined by specifying a minimum number of elements that must appear in the retelling. Again reliability may suffer, and the time and effort needed to read and evaluate the response lowers practicality. Validity, cognitive processing, communicative ability, and authenticity are all well incorporated into the task.

§ § § § §

A fifth category of listening comprehension was hinted at earlier in the chapter: **interactive** listening. Because such interaction presupposes a process of *speaking* in concert with listening, the interactive nature of listening will be addressed in the next chapter. Don't forget that a significant proportion of real-world listening performance is interactive. With the exception of media input, speeches, lectures, and eavesdropping, many of our listening efforts are directed toward a two-way process of speaking and listening in face-to-face conversations.

ASSESSING READING

Even as we are bombarded with an unending supply of visual and auditory media, the written word continues in its function to convey information, to amuse and entertain us, to codify our social, economic, and legal conventions, and to fulfill a host of other functions. In literate societies, most "normal" children learn to read by the age of five or six, and some even earlier. With the exception of a small number of people with learning disabilities, reading is a skill that is taken for granted.

In foreign language learning, reading is likewise a skill that teachers simply expect learners to acquire. Basic, beginning-level textbooks in a foreign language presuppose a student's reading ability if only because it's a *book* that is the medium. Most formal tests use the written word as a stimulus for test-taker response; even oral interviews may require reading performance for certain tasks. Reading, arguably the most essential skill for success in all educational contexts, remains a skill of paramount importance as we create assessments of general language ability.

Is reading so natural and normal that learners should simply be exposed to written texts with no particular instruction? Will they just absorb the skills necessary to convert their perception of a handful of letters into meaningful chunks of information? Not necessarily. For learners of English, two primary hurdles must be cleared in order to become efficient readers. First, they need to be able to master fundamental **bottom-up** strategies for processing separate letters, words, and phrases, as well as **top-down,** conceptually driven strategies for comprehension. Second, as part of that top-down approach, second language readers must develop appropriate content and formal **schemata**—background information and cultural experience—to carry out those interpretations effectively.

The assessment of reading ability does not end with the measurement of comprehension. Strategic pathways to full understanding are often important factors to include in assessing learners, especially in the case of most classroom assessments that are **formative** in nature. An inability to comprehend may thus be traced to a need to enhance a test-taker's strategies for achieving ultimate comprehension. For example, an academic technical report may be comprehensible to a student at the sentence level, but if the learner has not exercised certain strategies for noting the discourse conventions of that genre, misunderstanding may occur.

As we consider a number of different types or genres of written texts, the components of reading ability, and specific tasks that are commonly used in the assessment of reading, let's not forget the unobservable nature of reading. Like listening, one cannot see the **process** of reading, nor can one observe a specific **product** of reading. Other than observing a reader's eye movements and page turning, there is no technology that enables us to "see" sequences of graphic symbols traveling from the pages of a book into compartments of the brain (in a possible bottom-up process). Even more outlandish is the notion that one might be able to watch information from the brain make its way down onto the page (in typical top-down strategies). Further, once something is read—information from the written text is stored—no technology allows us to empirically measure exactly what is lodged in the brain. All assessment of reading must be carried out by inference.

TYPES (GENRES) OF READING

Each type or **genre** of written text has its own set of governing rules and conventions. A reader must be able to anticipate those conventions in order to process meaning efficiently. With an extraordinary number of genres present in any literate culture, the reader's ability to process texts must be very sophisticated. Consider the following abridged list of common genres, which ultimately form part of the specifications for assessments of reading ability.

Genres of reading

1. **Academic reading**
   - general interest articles (in magazines, newspapers, etc.)
   - technical reports (e.g., lab reports), professional journal articles
   - reference material (dictionaries, etc.)
   - textbooks, theses
   - essays, papers
   - test directions
   - editorials and opinion writing
2. **Job-related reading**
   - messages (e.g., phone messages)
   - letters/emails
   - memos (e.g., interoffice)
   - reports (e.g., job evaluations, project reports)
   - schedules, labels, signs, announcements
   - forms, applications, questionnaires
   - financial documents (bills, invoices, etc.)
   - directories (telephone, office, etc.)
   - manuals, directions

3. **Personal reading**

newspapers and magazines
letters, emails, greeting cards, invitations
messages, notes, lists
schedules (train, bus, plane, etc.)
recipes, menus, maps, calendars
advertisements (commercials, want ads)
novels, short stories, jokes, drama, poetry
financial documents (e.g., checks, tax forms, loan applications)
forms, questionnaires, medical reports, immigration documents
comic strips, cartoons

When we realize that this list is only the beginning, it is easy to see how overwhelming it is to learn to read in a foreign language! The genre of a text enables readers to apply certain **schemata** that will assist them in extracting appropriate meaning. If, for example, readers know that a text is a recipe, they will expect a certain arrangement of information (ingredients) and will know to search for a sequential order of directions. Efficient readers also have to know what their purpose is in reading a text, the strategies for accomplishing that purpose, and how to retain the information.

The content validity of an assessment procedure is largely established through the genre of a text. For example, if learners in a program of English for tourism have been learning how to deal with customers needing to arrange bus tours, then assessments of their ability should include guidebooks, maps, transportation schedules, calendars, and other relevant texts.

MICROSKILLS, MACROSKILLS, AND STRATEGIES FOR READING

Aside from attending to genres of text, the skills and strategies for accomplishing reading emerge as a crucial consideration in the assessment of reading ability. The micro- and macroskills below represent the spectrum of possibilities for objectives in the assessment of reading comprehension.

Micro- and macroskills for reading comprehension

Microskills

1. Discriminate among the distinctive graphemes and orthographic patterns of English.
2. Retain chunks of language of different lengths in short-term memory.
3. Process writing at an efficient rate of speed to suit the purpose.

4. Recognize a core of words, and interpret word order patterns and their significance.
5. Recognize grammatical word classes (nouns, verbs, etc.), systems (e.g., tense, agreement, pluralization), patterns, rules, and elliptical forms.
6. Recognize that a particular meaning may be expressed in different grammatical forms.
7. Recognize cohesive devices in written discourse and their role in signaling the relationship between and among clauses.

Macroskills

8. Recognize the rhetorical forms of written discourse and their significance for interpretation.
9. Recognize the communicative functions of written texts, according to form and purpose.
10. Infer context that is not explicit by using background knowledge.
11. From described events, ideas, etc., infer links and connections between events, deduce causes and effects, and detect such relations as main idea, supporting idea, new information, given information, generalization, and exemplification.
12. Distinguish between literal and implied meanings.
13. Detect culturally specific references and interpret them in a context of the appropriate cultural schemata.
14. Develop and use a battery of reading strategies, such as scanning and skimming, detecting discourse markers, guessing the meaning of words from context, and activating schemata for the interpretation of texts.

The assessment of reading can imply the assessment of a storehouse of reading strategies, as indicated in item #14. Aside from simply testing the ultimate achievement of comprehension of a written text, it may be important in some contexts to assess one or more of a storehouse of classic reading strategies. The brief taxonomy of strategies below is a list of possible assessment criteria.

Some principal strategies for reading comprehension

1. Identify your purpose in reading a text.
2. Apply spelling rules and conventions for bottom-up decoding.
3. Use lexical analysis (prefixes, roots, suffixes, etc.) to determine meaning.
4. Guess at meaning (of words, idioms, etc.) when you aren't certain.
5. Skim the text for the gist and for main ideas.
6. Scan the text for specific information (names, dates, key words).
7. Use silent reading techniques for rapid processing.

8. Use marginal notes, outlines, charts, or semantic maps for understanding and retaining information.
9. Distinguish between literal and implied meanings.
10. Capitalize on discourse markers to process relationships.

TYPES OF READING

In the previous chapters we saw that both listening and speaking could be subdivided into at least five different types of listening and speaking performance. In the case of reading, variety of performance is derived more from the multiplicity of types of texts (the genres listed above) than from the variety of overt types of performance. Nevertheless, for considering assessment procedures, several types of reading performance are typically identified, and these will serve as organizers of various assessment tasks.

1. Perceptive. In keeping with the set of categories specified for listening comprehension, similar specifications are offered here, except with some differing terminology to capture the uniqueness of reading. Perceptive reading tasks involve attending to the *components* of larger stretches of discourse: letters, words, punctuation, and other graphemic symbols. Bottom-up processing is implied.

2. Selective. This category is largely an artifact of assessment formats. In order to ascertain one's reading recognition of lexical, grammatical, or discourse features of language within a very short stretch of language, certain typical tasks are used: picture-cued tasks, matching, true/false, multiple-choice, etc. Stimuli include sentences, brief paragraphs, and simple charts and graphs. Brief responses are intended as well. A combination of bottom-up and top-down processing may be used.

3. Interactive. Included among interactive reading types are stretches of language of several paragraphs to one page or more in which the reader must, in a psycholinguistic sense, *interact* with the text. That is, reading is a process of negotiating meaning; the reader brings to the text a set of schemata for understanding it, and intake is the product of that interaction. Typical genres that lend themselves to interactive reading are anecdotes, short narratives and descriptions, excerpts from longer texts, questionnaires, memos, announcements, directions, recipes, and the like. The focus of an interactive task is to identify relevant features (lexical, symbolic, grammatical, and discourse) within texts of moderately short length with the objective of retaining the information that is processed. Top-down processing is typical of such tasks, although some instances of bottom-up performance may be necessary.

4. Extensive. Extensive reading, as discussed in this book, applies to texts of more than a page, up to and including professional articles, essays, technical reports, short stories, and books. (It should be noted that reading research commonly refers to "extensive reading" as longer stretches of discourse, such as long articles and books that are usually read outside a classroom hour. Here that definition is

massaged a little in order to encompass any text longer than a page.) The purposes of assessment usually are to tap into a learner's global understanding of a text, as opposed to asking test-takers to "zoom in" on small details. Top-down processing is assumed for most extensive tasks.

The four types of reading are demonstrated in Figure 8.1, which shows the relationships of length, focus, and processing mode among the four types.

	Length			Focus		Process	
	Short	Medium	Long	Form	Meaning	Bottom-Up	Top-Down
Perceptive	••			••		••	
Selective	•	•		••	•	•	•
Interactive		••		•	••	•	••
Extensive			••		••		••

•• strong emphasis
• moderate emphasis

Figure 8.1. Types of reading by length, focus, and process

DESIGNING ASSESSMENT TASKS: PERCEPTIVE READING

At the beginning level of reading a second language lies a set of tasks that are fundamental and basic: recognition of alphabetic symbols, capitalized and lowercase letters, punctuation, words, and grapheme-phoneme correspondences. Such tasks of perception are often referred to as **literacy** tasks, implying that the learner is in the early stages of becoming "literate." Some learners are already literate in their own native language, but in other cases the second language may be the first language that they have ever learned to read. This latter context poses cognitive and sometimes age-related issues that need to be considered carefully. Assessment of literacy is no easy assignment, and if you are interested in this particular challenging area, further reading beyond this book is advised (Harp, 1991; Farr & Tone, 1994; Genesee, 1994; Cooper, 1997). Assessment of basic reading skills may be carried out in a number of different ways.

Reading Aloud

The test-taker sees separate letters, words, and/or short sentences and reads them aloud, one by one, in the presence of an administrator. Since the assessment is of *reading* comprehension, any recognizable oral approximation of the target response is considered correct.

Written Response

The same stimuli are presented, and the test-taker's task is to reproduce the probe in writing. Because of the transfer across different skills here, evaluation of the test-taker's response must be carefully treated. If an error occurs, make sure you determine its source; what might be assumed to be a writing error, for example, may actually be a reading error, and vice versa.

Multiple-Choice

Multiple-choice responses are not only a matter of choosing one of four or five possible answers. Other formats, some of which are especially useful at the low levels of reading, include same/different, circle the answer, true/false, choose the letter, and matching. Here are some possibilities.

Minimal pair distinction

*Test-takers read:** Circle "S" for same or "D" for different.

1.	led	let	S	D
2.	bit	bit	S	D
3.	seat	set	S	D
4.	too	to	S	D

**In the case of very low level learners, the teacher/administrator reads directions.*

Grapheme recognition task

*Test-takers read:** Circle the "odd" item, the one that doesn't "belong."

1.	piece	peace	piece
2.	book	book	boot

**In the case of very low level learners, the teacher/administrator reads directions.*

Picture-Cued Items

Test-takers are shown a picture, such as the one on the next page, along with a written text and are given one of a number of possible tasks to perform.

Picture-cued word identification (Brown & Sahni, 1994, p. 124)

Test-takers hear: Point to the word that you read here.

cat	clock	chair

With the same picture, the test-taker might read sentences and then point to the correct part of the picture:

Picture-cued sentence identification

Test-takers hear: Point to the part of the picture that you read about here.

Test-takers see the picture and read each sentence written on a separate card.

The man is reading a book.

The cat is under the table.

Or a true/false procedure might be presented with the same picture cue:

Picture-cued true/false sentence identification

Test-takers read:

1. The pencils are under the table. T F
2. The cat is on the table. T F
3. The picture is over the couch. T F

Matching can be an effective method of assessing reading at this level. With objects labeled A, B, C, D, E in the picture, the test-taker reads words and writes the appropriate letter beside the word:

Picture-cued matching word identification

Test-takers read:

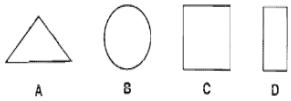
1. clock ____
2. chair ____
3. books ____
4. cat ____
5. table ____

Finally, test-takers might see a word or phrase and then be directed to choose one of four pictures that is being described, thus requiring the test-taker to transfer from a verbal to a nonverbal mode. In the following item, test-takers choose the correct letter:

Multiple-choice picture-cued word identification

Test-takers read: Rectangle

Test-takers see, and choose the correct item:



DESIGNING ASSESSMENT TASKS: SELECTIVE READING

Just above the rudimentary skill level of perception of letters and words is a category in which the test designer focuses on formal aspects of language (lexical, grammatical, and a few discourse features). This category includes what many incorrectly think of as testing "vocabulary and grammar." How many textbooks provide little tests and quizzes labeled "vocabulary and grammar" and never feature any other skill besides reading? Lexical and grammatical aspects of language are simply the forms we use to perform all four of the skills of listening, speaking, reading, and writing. (Notice that in all of these chapters on the four skills, formal features of language have become a potential focus for assessment.)

Here are some of the possible tasks you can use to assess lexical and grammatical aspects of *reading* ability.

Multiple-Choice (for Form-Focused Criteria)

By far the most popular method of testing a reading knowledge of vocabulary and grammar is the multiple-choice format, mainly for reasons of practicality: it is easy to administer and can be scored quickly. The most straightforward multiple-choice items may have little context, but might serve as a vocabulary or grammar check.

Multiple-choice vocabulary/grammar tasks

1. He's not married. He's _____.
   A. young
   B. single
   C. first
   D. a husband
2. If there's no doorbell, please _____ on the door.
   A. kneel
   B. type
   C. knock
   D. shout
3. The mouse is _____ the bed.
   A. under
   B. around
   C. between
4. The bank robbery occurred _____ I was in the restroom.
   A. that
   B. during
   C. while
   D. which

5. Yeast is an organic catalyst _____ known to prehistoric humanity.
   A. was
   B. which was
   C. which it
   D. which

This kind of darting from one context to another to another in a test has become so commonplace that learners almost expect the disjointedness. Some improvement of these items is possible by providing some context within each item:

Contextualized multiple-choice vocabulary/grammar tasks

1. Oscar: Do you like champagne?
   Lucy: No, I can't _____ it!
   A. stand
   B. prefer
   C. hate
2. Manager: Do you like to work by yourself?
   Employee: Yes, I like to work _____.
   A. independently
   B. definitely
   C. impatiently
3. Jack: Do you have a coat like this?
   John: Yes, mine is _____ yours.
   A. so same as
   B. the same like
   C. as same as
   D. the same as
4. Boss: Where did I put the Johnson file?
   Sectretary: I think _____ is on your desk.
   A. you were the file looking at
   B. the you were looking at file
   C. the file you were looking at
   D. you were looking at the file

A better contextualized format is to offer a modified cloze test (see page 201 for a treatment of cloze testing) adjusted to fit the objectives being assessed. In the example below, a few lines of English add to overall context.

Multiple-choice cloze vocabulary/grammar task

I've lived in the United States (21) _____ three years. I (22) _____ live in Costa Rica. I (23) _____ speak any English. I used to (24) _____ homesick, but now I enjoy (25) _____ here. I have never (26) _____ back home (27) _____ I came to the United States, but I might (28) _____ to visit my family soon.

21. A. since
B. for
C. during

22. A. used to
B. use to
C. was

23. A. couldn't
B. could
C. can

24. A. been
B. be
C. being

25. A. live
B. to live
C. living

26. A. be
B. been
C. was

27. A. when
B. while
C. since

28. A. go
B. will go
C. going

The context of the story in this example may not specifically help the test-taker to respond to the items more easily, but it allows the learner to attend to one set of related sentences for eight items that assess vocabulary and grammar. Other contexts might involve some content dependencies, such that earlier sentences predict the correct response for a later item. Thus, a pair of sentences in a short narrative might read:

He showed his suitcase (29) _____ me, but it wasn't big (30) _____ to fit all his clothes. So I gave him my suitcase, which was (31) _____.

29. A. for
B. from
C. to

30. A. so
B. too
C. enough

31. A. larger
B. smaller
C. largest

To respond to item #31 correctly, the test-taker needs to be able to comprehend the context of needing a *larger,* but not an equally grammatically correct *smaller,* suitcase. While such dependencies offer greater authenticity to an assessment, they also add the potential problem of a test-taker's missing several later items because of an earlier comprehension error.

Matching Tasks

At this selective level of reading, the test-taker's task is simply to respond correctly, which makes matching an appropriate format. The most frequently appearing criterion in matching procedures is vocabulary. Following is a typical format:

Vocabulary matching task

Write in the letter of the definition on the right that matches the word on the left.

_____ **1.** exhausted	a. unhappy
_____ **2.** disappointed	b. understanding of others
_____ **3.** enthusiastic	c. tired
_____ **4.** empathetic	d. excited

To add a communicative quality to matching, the first numbered list is sometimes a set of sentences with blanks in them, with a list of words to choose from:

Selected response fill-in vocabulary task

1. At the end of the long race, the runners were totally _____.
2. My parents were _____ with my bad performance on the final exam.
3. Everyone in the office was _____ about the new salary raises.
4. The _____ listening of the counselor made Christina feel well understood.

Choose from among the following:

disappointed
empathetic
exhausted
enthusiastic

Alderson (2000, p. 218) suggested matching procedures at an even more sophisticated level, where test-takers have to discern pragmatic interpretations of certain signs or labels such as "Freshly made sandwiches" and "Use before 10/23/02." Matches for those two are "We sell food" and "This is too old," which are selected from a number of other options.

Matching tasks have the advantage of offering an alternative to traditional multiple-choice or fill-in-the-blank formats and are sometimes easier to construct than multiple-choice items, as long as the test designer has chosen the matches carefully. Some disadvantages do come with this framework, however. They can become more of a puzzle-solving process than a genuine test of comprehension as test-takers struggle with the search for a match, possibly among 10 or 20 different items. Like other tasks in this section, they also are contrived exercises that are endemic to academia that will seldom be found in the real world.

Editing Tasks

Editing for grammatical or rhetorical errors is a widely used test method for assessing linguistic competence in reading. The TOEFL® and many other tests employ this technique with the argument that it not only focuses on grammar but also introduces a simulation of the authentic task of editing, or discerning errors in written passages. Its authenticity may be supported if you consider proofreading as a real-world skill that is being tested. Here is a typical set of examples of editing.

Multiple-choice grammar editing task (Phillips, 2001, p. 219)

Test-takers read: Choose the letter of the underlined word that is not correct.

1. The abrasively action of the wind wears away softer layers of rock.
   A B C D

2. There are two way of making a gas condense: cooling it or putting it under
   A B C D
   pressure.

3. Researchers have discovered that the application of bright light can sometimes
   A B
   be uses to overcome jet lag.
   C D

The above examples, with their disparate subject-matter content, are not as authentic as asking test-takers to edit a whole essay (see discussion below, pages 207–208). Of course, if learners have never practiced error detection tasks, the task itself is of some difficulty. Nevertheless, error detection has been

shown to be positively correlated with both listening comprehension and reading comprehension results on the TOEFL, at r = .58 and .76, respectively (*TOEFL Score User Guide*, 2001). Despite some authenticity quibbles, this task maintains a construct validity that justifies its use.

Picture-Cued Tasks

In the previous section we looked at picture-cued tasks for perceptive recognition of symbols and words. Pictures and photographs may be equally well utilized for examining ability at the selective level. Several types of picture-cued methods are commonly used.

1. Test-takers read a sentence or passage and choose one of four pictures that is being described. The sentence (or sentences) at this level is more complex. A computer-based example follows:

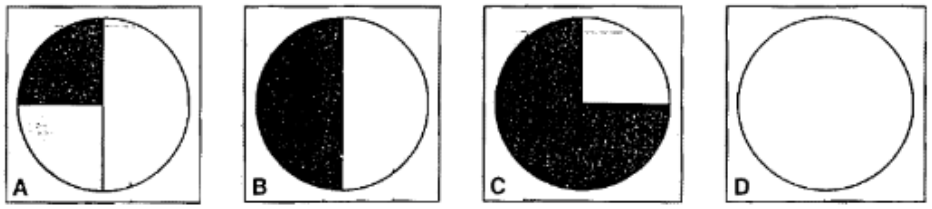
Multiple-choice picture-cued response (Phillips, 2001, p. 276)

Test-takers read a three-paragraph passage, one sentence of which is:

During at least three quarters of the year, the Arctic is frozen.

Click on the chart that shows the relative amount of time each year that water is available to plants in the Arctic.

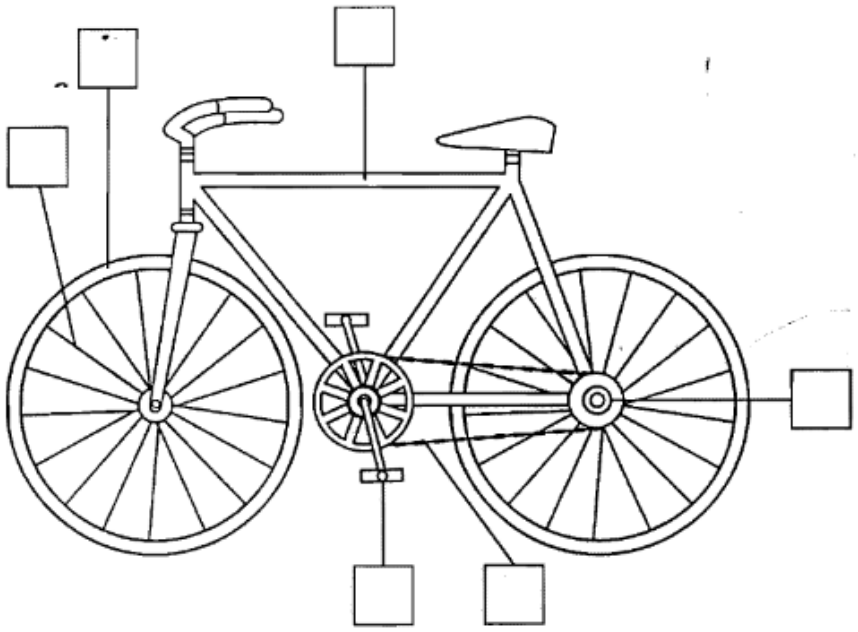
Test-takers see the following four pictures:



2. Test-takers read a series of sentences or definitions, each describing a labeled part of a picture or diagram. Their task is to identify each labeled item. In the following diagram, test-takers do not necessarily know each term, but by reading the definition they are able to make an identification. For example:

Diagram-labeling task

Test-takers see:



Test-takers read:

Label the picture with the number of the corresponding item described below.

1. wire supports extending from the hub of a wheel to its perimeter
2. a long, narrow support pole between the seat and the handlebars
3. a small, geared wheel concentric with the rear wheel
4. a long, linked, flexible metal device that propels the vehicle
5. a small rectangular lever operated by the foot to propel the vehicle
6. a tough but somewhat flexible rubber item that circles each wheel

The essential difference between the picture-cued tasks here and those that were outlined in the previous section is the complexity of the language.

Gap-Filling Tasks

Many of the multiple-choice tasks described above can be converted into gap-filling, or "fill-in-the-blank," items in which the test-taker's response is to write a word or phrase. An extension of simple gap-filling tasks is to create sentence completion items where test-takers read part of a sentence and then complete it by writing a phrase.

Sentence completion tasks

Oscar:	Doctor, what should I do if I get sick?
Doctor:	It is best to stay home and _____.
	If you have a fever, _____.
	You should drink as much _____.
	The worst thing you can do is _____.
	You should also _____.

The obvious disadvantage of this type of task is its questionable assessment of reading ability. The task requires both reading and writing performance, thereby rendering it of low validity in isolating reading as the sole criterion. Another drawback is scoring the variety of creative responses that are likely to appear. You will have to make a number of judgment calls on what comprises a correct response. In a test of reading comprehension only, you must accept as correct any responses that demonstrate comprehension of the first part of the sentence. This alone indicates that such tasks are better categorized as integrative procedures.

DESIGNING ASSESSMENT TASKS: INTERACTIVE READING

Tasks at this level, like selective tasks, have a combination of form-focused and meaning-focused objectives but with more emphasis on meaning. Interactive tasks may therefore imply a little more focus on top-down processing than on bottom-up. Texts are a little longer, from a paragraph to as much as a page or so in the case of ordinary prose. Charts, graphs, and other graphics may be somewhat complex in their format.

Cloze Tasks

One of the most popular types of reading assessment task is the **cloze** procedure. The word *cloze* was coined by educational psychologists to capture the Gestalt psychological concept of "closure," that is, the ability to fill in gaps in an incomplete image (visual, auditory, or cognitive) and supply (from background schemata) omitted details.

In written language, a sentence with a word left out should have enough context that a reader can close that gap with a calculated guess, using linguistic expectancies (formal schemata), background experience (content schemata), and some strategic competence. Based on this assumption, cloze tests were developed for native language readers and defended as an appropriate gauge of reading ability. Some research (Oller, 1973, 1976, 1979) on second language acquisition vigorously defends cloze testing as an integrative measure not only of reading ability but also

of other language abilities. It was argued that the ability to make coherent guesses in cloze gaps also taps into the ability to listen, speak, and write. With the decline of zeal for the search for the ideal integrative test in recent years, cloze testing has returned to a more appropriate status as one of a number of assessment procedures available for testing reading ability.

Cloze tests are usually a minimum of two paragraphs in length in order to account for discourse expectancies. They can be constructed relatively easily as long as the specifications for choosing deletions and for scoring are clearly defined. Typically every seventh word (plus or minus two) is deleted (known as **fixed-ratio deletion**), but many cloze test designers instead use a **rational deletion** procedure of choosing deletions according to the grammatical or discourse functions of the words. Rational deletion also allows the designer to avoid deleting words that would be difficult to predict from the context. For example, in the sentence "Everyone in the crowd enjoyed the gorgeous sunset," the seventh word is *gorgeous,* but learners could easily substitute other appropriate adjectives. Traditionally, cloze passages have between 30 and 50 blanks to fill, but a passage with as few as half a dozen blanks can legitimately be labeled a cloze test.

Two approaches to the scoring of cloze tests are commonly used. The **exact word** method gives credit to test-takers only if they insert the exact word that was originally deleted. The second method, **appropriate word** scoring, credits the test-taker for supplying any word that is grammatically correct and that makes good sense in the context. In the sentence above about the "gorgeous sunset," the test-takers would get credit for supplying *beautiful, amazing,* and *spectacular.* The choice between the two methods of scoring is one of practicality/reliability vs. face validity. In the exact word approach, scoring can be done quickly (especially if the procedure uses a multiple-choice technique) and reliably. The second approach takes more time because the teacher must determine whether each response is indeed appropriate, but students will perceive the test as being fairer: they won't get "marked off" for appropriate, grammatically correct responses.

The following excerpts from a longer essay illustrate the difference between rational and fixed-ratio deletion, and between exact word and appropriate word scoring.

Cloze procedure, fixed-ratio deletion (every seventh word)

> The recognition that one's feelings of (1) _____ and unhappiness can coexist much like (2) _____ and hate in a close relationship (3) _____ offer valuable clues on how to (4) _____ a happier life. It suggests, for (5) _____, that changing or avoiding things that (6) _____ you miserable may well make you (7) _____ miserable but probably no happier.

Cloze procedure, rational deletion (prepositions and conjunctions)

> The recognition that one's feelings (1) _____ happiness (2) _____ unhappiness can coexist much like love and hate (3) _____ a close relationship may offer valuable clues (4) _____ how to lead a happier life. It suggests, (5) _____ example, that changing (6) _____ avoiding things that make you miserable may well make you less miserable (7) _____ probably no happier.

In both versions there are seven deletions, but the second version allows the test designer to tap into prediction of prepositions and conjunctions in particular. And the second version provides more washback as students focus on targeted grammatical features.

Both of the scoring methods named above could present problems, with the first version presenting a little more ambiguity. Possible responses might include:

Fixed-ratio version, blank #3: *may, might, could, can*
#4: *lead, live, have, seek*
#5: *example, instance*

Rational deletion version, blank #4: *on, about*
#6: *or, and*
#7: *but, and*

Arranging a cloze test in a multiple-choice format allows even more rapid scoring: hand scoring with an answer key or hole-punched grid, or computer scoring using scannable answer sheets. Multiple-choice cloze tests must of course adhere to all the other guidelines for effective multiple-choice items that were covered in Chapter 4, especially the choice of appropriate distractors; therefore they can take much longer to construct—possibly too long to pay off in a classroom setting.

Some variations on standard cloze testing have appeared over the years; two of the better known are the C-test and the cloze-elide procedure. In the **C-test** (Klein-Braley & Raatz, 1984; Klein-Braley, 1985; Dörnyei & Katona, 1992), the second half (according to the number of letters) of every other word is obliterated and the test-taker must restore each word. While Klein-Braley and others vouched for its validity and reliability, many consider this technique to be "even more irritating to complete than cloze tests" (Alderson, 2000, p. 225). Look at the following example and judge for yourself:

C-test procedure

> The recognition th_ _ one's feel_ _ _ _ of happ_ _ _ _ _ and unhap_ _ _ _ _ _ can coe_ _ _ _ much li_ _ love a_ _ hate i_ a cl_ _ _ relati_ _ _ _ _ _ may of_ _ _ valuable cl_ _ _ on h_ _ to le_ _ a hap_ _ _ _ life. I_ suggests, f_ _ example, th_ _ changing o_ avoiding thi_ _ _ that ma_ _ you mise_ _ _ _ _ may we_ _ make y_ _ less mise_ _ _ _ _ but prob_ _ _ _ no hap_ _ _ _.

The second variation, the **cloze-elide** procedure, inserts words into a text that don't belong. The test-taker's task is to detect and cross out the "intrusive" words. Look at the same familiar passage:

Cloze-elide procedure

> The recognition that one's now feelings of happiness and unhappiness can under coexist much like love and hate in a close then relationship may offer valuable clues on how to lead a happier with life. It suggests, for example, that changing or avoiding my things that make you miserable may well make you less miserable ever but probably no happier.

Critics of this procedure (Davies, 1975) claimed that the cloze-elide procedure is actually a test of reading speed and not of proofreading skill, as its proponents asserted. Two disadvantages are nevertheless immediately apparent: (1) Neither the words to insert nor the frequency of insertion appears to have any rationale. (2) Fast and efficient readers are not adept at detecting the intrusive words. Good readers naturally weed out such potential interruptions.

Impromptu Reading Plus Comprehension Questions

If cloze testing is the most-researched procedure for assessing reading, the traditional "Read a passage and answer some questions" technique is undoubtedly the oldest and the most common. Virtually every proficiency test uses the format, and one would rarely consider assessing reading without some component of the assessment involving impromptu reading and responding to questions.

In Chapter 4, in the discussion on proficiency testing, we looked at a typical reading comprehension passage and a set of questions from the TOEFL. Here's another such passage.

Reading comprehension passage (Phillips, 2001, pp. 421–422)

Questions 1–10

The Hollywood sign in the hills that line the northern border of Los Angeles is a famous landmark recognized the world over. The white-painted, 50-foot-high, sheet metal letters can be seen from great distances across the Los Angeles basin.

Line (5) The sign was not constructed, as one might suppose, by the movie business as a means of celebrating the importance of Hollywood to this industry; instead, it was first constructed in 1923 as a means of advertising homes for sale in a 500-acre housing subdivision in a part of Los Angeles called "Hollywoodland." The sign that was constructed at the time, of course, said "Hollywoodland." Over the years, people began referring to the area by the shortened version "Hollywood," and after the sign and its site were donated to the city in 1945, the last four letters were removed.

(10) The sign suffered from years of disrepair, and in 1973 it needed to be completely replaced, at a cost of $27,700 per letter. Various celebrities were instrumental in helping to raise needed funds. Rock star Alice Cooper, for example, bought an O in memory of Groucho Marx, and Hugh Hefner of *Playboy* fame held a benefit party to raise the money for the Y. The construction of the new sign was finally completed in 1978.

1. What is the topic of this passage?
   (A) A famous sign
   (B) A famous city
   (C) World landmarks
   (D) Hollywood versus Hollywoodland
2. The expression "the world over" in line 2 could best be replaced by
   (A) in the northern parts of the world
   (B) on top of the world
   (C) in the entire world
   (D) in the skies
3. It can be inferred from the passage that most people think that the Hollywood sign was first constructed by
   (A) an advertising company
   (B) the movie industry
   (C) a construction company
   (D) the city of Los Angeles
4. The pronoun "it" in line 5 refers to
   (A) the sign
   (B) the movie business
   (C) the importance of Hollywood
   (D) this industry
5. According to the passage, the Hollywood sign was first built in
   (A) 1923
   (B) 1949
   (C) 1973
   (D) 1978
6. Which of the following is NOT mentioned about Hollywoodland?
   (A) It used to be the name of an area of Los Angeles.
   (B) It was formerly the name on the sign in the hills.
   (C) There were houses for sale there.
   (D) It was the most expensive area of Los Angeles.
7. The passage indicates that the sign suffered because
   (A) people damaged it
   (B) it was not fixed
   (C) the weather was bad
   (D) it was poorly constructed
8. It can be inferred from the passage that the Hollywood sign was how old when it was necessary to replace it completely?
   (A) Ten years old
   (B) Twenty-six years old
   (C) Fifty years old
   (D) Fifty-five years old
9. The word "replaced" in line 10 is closest in meaning to which of the following?
   (A) Moved to a new location
   (B) Destroyed
   (C) Found again
   (D) Exchanged for a newer one
10. According to the passage, how did celebrities help with the new sign?
    (A) They played instruments.
    (B) They raised the sign.
    (C) They helped get the money.
    (D) They took part in work parties to build the sign.

Notice that this set of questions, based on a 250-word passage, covers the comprehension of these features:

- main idea (topic)
- expressions/idioms/phrases in context
- inference (implied detail)
- grammatical features
- detail (scanning for a specifically stated detail)
- excluding facts not written (unstated details)
- supporting idea(s)
- vocabulary in context

These specifications, and the questions that exemplify them, are *not* just a string of "straight" comprehension questions that follow the thread of the passage. The questions represent a sample of the test specifications for TOEFL reading passages, which are derived from research on a variety of abilities good readers exhibit. Notice that many of them are consistent with strategies of effective reading: skimming for main idea, scanning for details, guessing word meanings from context, inferencing, using discourse markers, etc. To construct your own assessments that involve short reading passages followed by questions, you can begin with TOEFL-like specs as a basis. Your focus in your own classroom will determine which of these—and possibly other specifications—you will include in your assessment procedure, how you will frame questions, and how much weight you will give each item in scoring.

The technology of computer-based reading comprehension tests of this kind enables some additional types of items. Items such as the following are typical:

Computer-based TOEFL® reading comprehension item

- Click on the word in paragraph 1 that means "subsequent work."
- Look at the word *they* in paragraph 2. Click on the word that *they* refers to.
- The following sentence could be added to paragraph 2:

  Instead, he used the pseudonym Mrs. Silence Dogood.

  Where would it best fit into the paragraph? Click on the square ☐ to add the sentence to the paragraph.

- Click on the drawing that most closely resembles the prehistoric coelacanth. [*Four drawings are depicted on the screen.*]

Short-Answer Tasks

Multiple-choice items are difficult to construct and validate, and classroom teachers rarely have time in their busy schedules to design such a test. A popular alternative

to multiple-choice questions following reading passages is the age-old short-answer format. A reading passage is presented, and the test-taker reads questions that must be answered in a sentence or two. Questions might cover the same specifications indicated above for the TOEFL reading, but be worded in question form. For example, in a passage on the future of airline travel, the following questions might appear:

Open-ended reading comprehension questions

1. What do you think the main idea of this passage is?
2. What would you infer from the passage about the future of air travel?
3. In line 6 the word *sensation* is used. From the context, what do you think this word means?
4. What two ideas did the writer suggest for increasing airline business?
5. Why do you think the airlines have recently experienced a decline?

Do not take lightly the design of questions. It can be difficult to make sure that they reach their intended criterion. You will also need to develop consistent specifications for acceptable student responses and be prepared to take the time necessary to accomplish their evaluation. But these rather predictable disadvantages may be outweighed by the face validity of offering students a chance to construct their own answers, and by the washback effect of potential follow-up discussion.

Editing (Longer Texts)

The previous section of this chapter (on selective reading) described editing tasks, but there the discussion was limited to a list of unrelated sentences, each presented with an error to be detected by the test-taker. The same technique has been applied successfully to longer passages of 200 to 300 words. Several advantages are gained in the longer format.

First, *authenticity* is increased. The likelihood that students in English classrooms will read connected prose of a page or two is greater than the likelihood of their encountering the contrived format of unconnected sentences. Second, the task *simulates proofreading* one's own essay, where it is imperative to find and correct errors. And third, if the test is connected to a specific curriculum (such as placement into one of several writing courses), the test designer can draw up specifications for a number of grammatical and rhetorical *categories that match the content* of the courses. Content validity is thereby supported, and along with it the face validity of a task in which students are willing to invest.

Imao's (2001) test introduced one error in each numbered sentence. Test-takers followed the same procedure for marking errors as described in the previous section. Instructions to the student included a sample of the kind of connected prose that test-takers would encounter:

Contextualized grammar editing tasks (Imao, 2001)

(1) Ever since supermarkets first appeared, they have been take over the world.
A B C D

(2) Supermarkets have changed people's life styles, yet and at the same time,
A B C

changes in people's life styles have encouraged the opening of supermarkets. (3) As
D

a result this, many small stores have been forced out of business. (4) Moreover, some
A B C D B

small stores will be able to survive this unfavorable situation.
A C D

This can all be achieved in a multiple-choice format with computer scannable scoring for a rapid return of results. Moreover, not only does an overall score provide a holistic assessment, but for the placement purposes that Imao's research addressed, teachers were able to be given a diagnostic chart of each student's results within all of the specified categories of the test. For a total of 32 to 56 items in his editing test, Imao (2001, p. 185) was able to offer teachers a computer-generated breakdown of performance in the following categories:

Sentence structure
Verb tense
Noun/article features
Modal auxiliaries
Verb complements
Noun clauses
Adverb clauses
Conditionals
Logical connectors
Adjective clauses (including relative clauses)
Passives

These categories were selected for inclusion from a survey of instructors' syllabuses in writing courses and proofreading workshops. This is an excellent example of the washback effect of a relatively large-scale, standardized multiple-choice test. While one would not want to use such data as absolutely predictive of students' future

work, they can provide guidelines to a teacher on areas of potential focus as the writing course unfolds.

Scanning

Scanning is a strategy used by all readers to find relevant information in a text. Assessment of scanning is carried out by presenting test-takers with a text (prose or something in a chart or graph format) and requiring rapid identification of relevant bits of information. Possible stimuli include

- a one- to two-page news article,
- an essay,
- a chapter in a textbook,
- a technical report,
- a table or chart depicting some research findings,
- a menu, and
- an application form.

Among the variety of scanning objectives (for each of the genres named above), the test-taker must locate

- a date, name, or place in an article;
- the setting for a narrative or story;
- the principal divisions of a chapter;
- the principal research finding in a technical report;
- a result reported in a specified cell in a table;
- the cost of an item on a menu; and
- specified data needed to fill out an application.

Scoring of such scanning tasks is amenable to specificity if the initial directions are specific ("How much does the dark chocolate torte cost?"). Since one of the purposes of scanning is to *quickly* identify important elements, timing may also be calculated into a scoring procedure.

Ordering Tasks

Students always enjoy the activity of receiving little strips of paper, each with a sentence on it, and assembling them into a story, sometimes called the "strip story" technique. Variations on this can serve as an assessment of overall global understanding of a story and of the cohesive devices that signal the order of events or ideas. Alderson et al. (1995, p. 53) warn, however, against assuming that there is only one logical order. They presented these sentences for forming a little story.

Sentence-ordering task

Put the following sentences in the correct order:

A	it was called "The Last Waltz"
B	the street was in total darkness
C	because it was one he and Richard had learnt at school
D	Peter looked outside
E	he recognised the tune
F	and it seemed deserted
G	he thought he heard someone whistling

"D" was the first sentence, and test-takers were asked to order the sentences. It turned out that two orders were acceptable (DGECABF and DBFGECA), creating difficulties in assigning scores and leading the authors to discourage the use of this technique as an assessment device. But if you are willing to place this procedure in the category of informal and/or formative assessment, you might consider the technique useful. Different acceptable sentence orders become an instructive point for subsequent discussion in class, and you thereby offer washback into students' understanding of how to connect sentences and ideas in a story or essay.

Information Transfer: Reading Charts, Maps, Graphs, Diagrams

Every educated person must be able to comprehend charts, maps, graphs, calendars, diagrams, and the like. Converting such nonverbal input into comprehensible intake requires not only an understanding of the graphic and verbal conventions of the medium but also a linguistic ability to interpret that information to someone else. Reading a map implies understanding the conventions of map graphics, but it is often accompanied by telling someone where to turn, how far to go, etc. Scanning a menu requires an ability to understand the structure of most menus as well as the capacity to give an order when the time comes. Interpreting the numbers on a stock market report involves the interaction of understanding the numbers and of conveying that understanding to others.

All of these media presuppose the reader's appropriate schemata for interpreting them and often are accompanied by oral or written discourse in order to convey, clarify, question, argue, and debate, among other linguistic functions. Virtually every language curriculum, from rock-bottom beginning levels to high-advanced, utilizes this nonverbal, visual/symbolic dimension. It is therefore imperative that assessment procedures include measures of comprehension of nonverbal media.

To comprehend information in this medium (hereafter referred to simply as "graphics"), learners must be able to

- comprehend specific conventions of the various types of graphics;
- comprehend labels, headings, numbers, and symbols;
- comprehend the possible relationships among elements of the graphic; and
- make inferences that are not presented overtly.

The act of comprehending graphics includes the linguistic performance of oral or written interpretations, comments, questions, etc. This implies a process of **information transfer** from one skill to another: in this case, from reading verbal and/or nonverbal information to speaking/writing. Assessment of these abilities covers a broad spectrum of tasks. Here is a start of the many possibilities.

Tasks for assessing interpretation of graphic information

1. Read a graphic; answer simple, direct information questions. For example:
   - map: "Where is the post office?"
   - family tree: "Who is Tony's great grandmother?"
   - statistical table: "What does $p < .05$ mean?"
   - diagram of a steam engine: "Label the following parts."
2. Read a graphic; describe or elaborate on information.
   - map: "Compare the distance between San Francisco and Sacramento to the distance between San Francisco and Monterey."
   - store advertisements: "Who has the better deal on grapes, Safeway or Albertsons?"
   - menu: "What comes with the grilled salmon entrée?"
3. Read a graphic; infer/predict information.
   - stock market report: "Based on past performance, how do you think Macrotech Industries will do in the future?"
   - directions for assembling a bookshelf: "How long do you think it will take to put this thing together?"
4. Read a passage; choose the correct graphic for it.
   - article about the size of the ozone hole in the Antarctic: "Which chart represents the size of the ozone hole?"
   - passage about the history of bicycles: "Click on the drawing that shows a penny-farthing bicycle."
5. Read a passage with an accompanying graphic; interpret both.
   - article about hunger and population, with a bar graph: "Which countries have the most hungry people and why?"
   - article on number of automobiles produced and their price over a 10-year period, with a table: "What is the best generalization you can make about production and the cost of automobiles?"

6. Read a passage; create or use a graphic to illustrate.
   - directions from the bank to the post office: "On the map provided, trace the route from the bank to the post office."
   - article about deforestation and carbon dioxide levels: "Make a bar graph to illustrate the information in the article."
   - story including members of a family: "Draw Jeff and Christina's family tree."
   - description of a class schedule: "Fill in Mary's weekly class schedule."

All these tasks involve retrieving information from either written or graphic media and transferring that information to productive performance. It is sometimes too easy to simply conclude that *reading* must involve only 26 alphabetic letters, with spaces and punctuation, thus omitting a huge number of resources that we consult every day.

DESIGNING ASSESSMENT TASKS: EXTENSIVE READING

Extensive reading involves somewhat longer texts than we have been dealing with up to this point. Journal articles, technical reports, longer essays, short stories, and books fall into this category. The reason for placing such reading into a separate category is that reading of this type of discourse almost always involves a focus on meaning using mostly top-down processing, with only occasional use of a targeted bottom-up strategy. Also, because of the extent of such reading, formal assessment is unlikely to be contained within the time constraints of a typical formal testing framework, which presents a unique challenge for assessment purposes.

Another complication in assessing extensive reading is that the expected response from the reader is likely to involve as much written (or sometimes oral) performance as reading. For example, in asking test-takers to respond to an article or story, one could argue that a greater emphasis is placed on writing than on reading. This is no reason to sweep extensive reading assessment under the rug; teachers should not shrink from the assessment of this highly sophisticated skill.

Before examining a few tasks that have proved to be useful in assessing extensive reading, it is essential to note that a number of the tasks described in previous categories can apply here. Among them are

- impromptu reading plus comprehension questions,
- short-answer tasks,
- editing,
- scanning,
- ordering,
- information transfer, and
- interpretation (discussed under graphics).

In addition to those applications are tasks that are unique to extensive reading: skimming, summarizing, responding to reading, and note-taking.

Skimming Tasks

Skimming is the process of rapid coverage of reading matter to determine its gist or main idea. It is a prediction strategy used to give a reader a sense of the topic and purpose of a text, the organization of the text, the perspective or point of view of the writer, its ease or difficulty, and/or its usefulness to the reader. Of course skimming can apply to texts of less than one page, so it would be wise not to confine this type of task just to extensive texts.

Assessment of skimming strategies is usually straightforward: the test-taker skims a text and answers questions such as the following:

Skimming tasks

> What is the main idea of this text?
> What is the author's purpose in writing the text?
> What kind of writing is this [newspaper article, manual, novel, etc.]?
> What type of writing is this [expository, technical, narrative, etc.]?
> How easy or difficult do you think this text will be?
> What do you think you will learn from the text?
> How useful will the text be for your [profession, academic needs, interests]?

Responses are oral or written, depending on the context. Most assessments in the domain of skimming are informal and formative: they are grist for an imminent discussion, a more careful reading to follow, or an in-class discussion, and therefore their washback potential is good. Insofar as the subject matter and tasks are useful to a student's goals, authenticity is preserved. Scoring is less of an issue than providing appropriate feedback to students on their strategies of prediction.

Summarizing and Responding

One of the most common means of assessing extensive reading is to ask the test-taker to write a summary of the text. The task that is given to students can be very simply worded:

Directions for summarizing

> Write a summary of the text. Your summary should be about one paragraph in length (100–150 words) and should include your understanding of the main idea and supporting ideas.

Evaluating summaries is difficult: Do you give test-takers a certain number of points for targeting the main idea and its supporting ideas? Do you use a full/partial/no-credit point system? Do you give a holistic score? Imao (2001) used four criteria for the evaluation of a summary:

Criteria for assessing a summary (Imao, 2001, p. 184)

1. Expresses accurately the main idea and supporting ideas.
2. Is written in the student's own words; occasional vocabulary from the original text is acceptable.
3. Is logically organized.
4. Displays facility in the use of language to clearly express ideas in the text.

As you can readily see, a strict adherence to the criterion of assessing reading, and reading only, implies consideration of only the first factor; the other three pertain to writing performance. The first criterion is nevertheless a crucial factor; otherwise the reader-writer could pass all three of the other criteria with virtually no understanding of the text itself. Evaluation of the reading comprehension criterion will of necessity remain somewhat subjective because the teacher will need to determine degrees of fulfillment of the objective (see below for more about scoring this task).

Of further interest in assessing extensive reading is the technique of asking a student to **respond** to a text. The two tasks should not be confused with each other: summarizing requires a synopsis or overview of the text, while responding asks the reader to provide his or her own opinion on the text as a whole or on some statement or issue within it. Responding may be prompted by such directions as this:

Directions for responding to reading

In the article "Poisoning the Air We Breathe," the author suggests that a global dependence on fossil fuels will eventually make air in large cities toxic. Write an essay in which you agree or disagree with the author's thesis. Support your opinion with information from the article and from your own experience.

One criterion for a good response here is the extent to which the test-taker accurately reflects the content of the article and some of the arguments therein. Scoring is also difficult here because of the subjectivity of determining an accurate reflection of the article itself. For the reading component of this task, as well as the summary task described above, a holistic scoring system may be feasible:

Holistic scoring scale for summarizing and responding to reading

3	Demonstrates clear, unambiguous comprehension of the main and supporting ideas.
2	Demonstrates comprehension of the main idea but lacks comprehension of some supporting ideas.
1	Demonstrates only a partial comprehension of the main and supporting ideas.
0	Demonstrates no comprehension of the main and supporting ideas.

The teacher or test administrator must still determine shades of gray between the point categories, but the descriptions help to bridge the gap between an empirically determined evaluation (which is impossible) and wild, impressionistic guesses.

An attempt has been made here to underscore the *reading* component of summarizing and responding to reading, but it is crucial to consider the interactive relationship between reading and writing that is highlighted in these two tasks. As you direct students to engage in such integrative performance, it is advisable not to treat them as tasks for assessing reading alone.

Note-Taking and Outlining

Finally, a reader's comprehension of extensive texts may be assessed through an evaluation of a process of note-taking and/or outlining. Because of the difficulty of controlling the conditions and time frame for both these techniques, they rest firmly in the category of informal assessment. Their utility is in the strategic training that learners gain in retaining information through marginal notes that highlight key information or organizational outlines that put supporting ideas into a visually manageable framework. A teacher, perhaps in one-on-one conferences with students, can use student notes/outlines as indicators of the presence or absence of effective reading strategies, and thereby point the learners in positive directions.

ASSESSING SPEAKING

BASIC TYPES OF SPEAKING

In Chapter 6, we cited four categories of listening performance assessment tasks. A similar taxonomy emerges for oral production.

1. Imitative. At one end of a continuum of types of speaking performance is the ability to simply parrot back (**imitate**) a word or phrase or possibly a sentence. While this is a purely phonetic level of oral production, a number of prosodic, lexical, and grammatical properties of language may be included in the criterion performance. We are interested only in what is traditionally labeled "pronunciation"; no inferences are made about the test-taker's ability to understand or convey meaning or to participate in an interactive conversation. The only role of listening here is in the short-term storage of a prompt, just long enough to allow the speaker to retain the short stretch of language that must be imitated.

2. Intensive. A second type of speaking frequently employed in assessment contexts is the production of short stretches of oral language designed to demonstrate competence in a narrow band of grammatical, phrasal, lexical, or phonological relationships (such as prosodic elements—intonation, stress, rhythm, juncture). The speaker must be aware of semantic properties in order to be able to respond, but interaction with an interlocutor or test administrator is minimal at best. Examples of **intensive** assessment tasks include directed response tasks, reading aloud, sentence and dialogue completion; limited picture-cued tasks including simple sequences; and translation up to the simple sentence level.

3. Responsive. **Responsive** assessment tasks include interaction and test comprehension but at the somewhat limited level of very short conversations, standard greetings and small talk, simple requests and comments, and the like. The stimulus is almost always a spoken prompt (in order to preserve authenticity), with perhaps only one or two follow-up questions or retorts:

A. Mary: Excuse me, do you have the time?
Doug: Yeah. Nine-fifteen.

B. T: What is the most urgent environmental problem today?
S: I would say massive deforestation.

C. Jeff: Hey, Stef, how's it going?
Stef: Not bad, and yourself?
Jeff: I'm good.
Stef: Cool. Okay, gotta go.

4. *Interactive.* The difference between responsive and interactive speaking is in the length and complexity of the interaction, which sometimes includes multiple exchanges and/or multiple participants. Interaction can take the two forms of **transactional** language, which has the purpose of exchanging specific information, or **interpersonal** exchanges, which have the purpose of maintaining social relationships. (In the three dialogues cited above, **A** and **B** were transactional, and **C** was interpersonal.) In interpersonal exchanges, oral production can become pragmatically complex with the need to speak in a casual register and use colloquial language, ellipsis, slang, humor, and other sociolinguistic conventions.

5. *Extensive (monologue).* Extensive oral production tasks include speeches, oral presentations, and story-telling, during which the opportunity for oral interaction from listeners is either highly limited (perhaps to nonverbal responses) or ruled out altogether. Language style is frequently more deliberative (planning is involved) and formal for extensive tasks, but we cannot rule out certain informal monologues such as casually delivered speech (for example, my vacation in the mountains, a recipe for outstanding pasta primavera, recounting the plot of a novel or movie).

MICRO- AND MACROSKILLS OF SPEAKING

In Chapter 6, a list of listening micro- and macroskills enumerated the various components of listening that make up criteria for assessment. A similar list of speaking skills can be drawn up for the same purpose: to serve as a taxonomy of skills from which you will select one or several that will become the objective(s) of an assessment task. The microskills refer to producing the smaller chunks of language such as phonemes, morphemes, words, collocations, and phrasal units. The macroskills imply the speaker's focus on the larger elements: fluency, discourse, function, style, cohesion, nonverbal communication, and strategic options. The micro- and macroskills total roughly 16 different objectives to assess in speaking.

Micro- and macroskills of oral production

Microskills

1. Produce differences among English phonemes and allophonic variants.
2. Produce chunks of language of different lengths.
3. Produce English stress patterns, words in stressed and unstressed positions, rhythmic structure, and intonation contours.
4. Produce reduced forms of words and phrases.
5. Use an adequate number of lexical units (words) to accomplish pragmatic purposes.
6. Produce fluent speech at different rates of delivery.

7. Monitor one's own oral production and use various strategic devices—pauses, fillers, self-corrections, backtracking—to enhance the clarity of the message.
8. Use grammatical word classes (nouns, verbs, etc.), systems (e.g., tense, agreement, pluralization), word order, patterns, rules, and elliptical forms.
9. Produce speech in natural constituents: in appropriate phrases, pause groups, breath groups, and sentence constituents.
10. Express a particular meaning in different grammatical forms.
11. Use cohesive devices in spoken discourse.

Macroskills

12. Appropriately accomplish communicative functions according to situations, participants, and goals.
13. Use appropriate styles, registers, implicature, redundancies, pragmatic conventions, conversation rules, floor-keeping and -yielding, interrupting, and other sociolinguistic features in face-to-face conversations.
14. Convey links and connections between events and communicate such relations as focal and peripheral ideas, events and feelings, new information and given information, generalization and exemplification.
15. Convey facial features, kinesics, body language, and other nonverbal cues along with verbal language.
16. Develop and use a battery of speaking strategies, such as emphasizing key words, rephrasing, providing a context for interpreting the meaning of words, appealing for help, and accurately assessing how well your interlocutor is understanding you.

As you consider designing tasks for assessing spoken language, these skills can act as a checklist of objectives. While the macroskills have the appearance of being more complex than the microskills, both contain ingredients of difficulty, depending on the stage and context of the test-taker.

There is such an array of oral production tasks that a complete treatment is almost impossible within the confines of one chapter in this book. Below is a consideration of the most common techniques with brief allusions to related tasks. As already noted in the introduction to this chapter, consider three important issues as you set out to design tasks:

1. No speaking task is capable of isolating the single skill of oral production. Concurrent involvement of the additional performance of aural comprehension, and possibly reading, is usually necessary.

2. Eliciting the specific criterion you have designated for a task can be tricky because beyond the word level, spoken language offers a number of productive options to test-takers. Make sure your elicitation prompt achieves its aims as closely as possible.

3. Because of the above two characteristics of oral production assessment, it is important to carefully specify scoring procedures for a response so that ultimately you achieve as high a reliability index as possible.

DESIGNING ASSESSMENT TASKS: IMITATIVE SPEAKING

You may be surprised to see the inclusion of simple phonological imitation in a consideration of assessment of oral production. After all, endless repeating of words, phrases, and sentences was the province of the long-since-discarded Audiolingual Method, and in an era of communicative language teaching, many believe that nonmeaningful imitation of sounds is fruitless. Such opinions have faded in recent years as we discovered that an overemphasis on fluency can sometimes lead to the decline of accuracy in speech. And so we have been paying more attention to pronunciation, especially suprasegmentals, in an attempt to help learners be more comprehensible.

An occasional phonologically focused repetition task is warranted as long as repetition tasks are not allowed to occupy a dominant role in an overall oral production assessment, and as long as you artfully avoid a negative washback effect. Such tasks range from word level to sentence level, usually with each item focusing on a specific phonological criterion. In a simple repetition task, test-takers repeat the stimulus, whether it is a pair of words, a sentence, or perhaps a question (to test for intonation production).

Word repetition task

Test-takers hear:	Repeat after me:	
	beat [*pause*] bit [*pause*]	
	bat [*pause*] vat [*pause*]	etc.
	I bought a boat yesterday.	
	The glow of the candle is growing.	etc.
	When did they go on vacation?	
	Do you like coffee?	etc.
Test-takers repeat the stimulus.		

A variation on such a task prompts test-takers with a brief written stimulus which they are to read aloud. (In the section below on intensive speaking, some tasks are described in which test-takers read aloud longer texts.) Scoring specifications must be clear in order to avoid reliability breakdowns. A common form of scoring simply indicates a two- or three-point system for each response.

Scoring scale for repetition tasks

2	acceptable pronunciation
1	comprehensible, partially correct pronunciation
0	silence, seriously incorrect pronunciation

The longer the stretch of language, the more possibility for error and therefore the more difficult it becomes to assign a point system to the text. In such a case, it may be imperative to score only the criterion of the task. For example, in the sentence "When did they go on vacation?" since the criterion is falling intonation for *wh*-questions, points should be awarded regardless of any mispronunciation.

DESIGNING ASSESSMENT TASKS: INTENSIVE SPEAKING

At the intensive level, test-takers are prompted to produce short stretches of discourse (no more than a sentence) through which they demonstrate linguistic ability at a specified level of language. Many tasks are "cued" tasks in that they lead the test-taker into a narrow band of possibilities.

Parts C and D of the PhonePass test fulfill the criteria of intensive tasks as they elicit certain expected forms of language. Antonyms like *high* and *low, happy* and *sad* are prompted so that the automated scoring mechanism anticipates only one word. The either/or task of Part D fulfills the same criterion. Intensive tasks may also be described as **limited** response tasks (Madsen, 1983), or **mechanical** tasks (Underhill, 1987), or what classroom pedagogy would label as **controlled** responses.

Directed Response Tasks

In this type of task, the test administrator elicits a particular grammatical form or a transformation of a sentence. Such tasks are clearly mechanical and not communicative, but they do require minimal processing of meaning in order to produce the correct grammatical output.

Directed response

Test-takers hear:	Tell me he went home.
	Tell me that you like rock music.
	Tell me that you aren't interested in tennis.
	Tell him to come to my office at noon.
	Remind him what time it is.

Read-Aloud Tasks

Intensive reading-aloud tasks include reading beyond the sentence level up to a paragraph or two. This technique is easily administered by selecting a passage that incorporates test specs and by recording the test-taker's output; the scoring is relatively easy because all of the test-taker's oral production is controlled. Because of the

results of research on the PhonePass test, reading aloud may actually be a surprisingly strong indicator of overall oral production ability.

For many decades, foreign language programs have used reading passages to analyze oral production. Prator's (1972) *Manual of American English Pronunciation* included a "diagnostic passage" of about 150 words that students could read aloud into a tape recorder. Teachers listening to the recording would then rate students on a number of phonological factors (vowels, diphthongs, consonants, consonant clusters, stress, and intonation) by completing a two-page diagnostic checklist on which all errors or questionable items were noted. These checklists ostensibly offered direction to the teacher for emphases in the course to come.

An earlier form of the Test of Spoken English (TSE®, see below) incorporated one read-aloud passage of about 120 to 130 words with a rating scale for pronunciation and fluency. The following passage is typical:

Read-aloud stimulus, paragraph length

> Despite the decrease in size—and, some would say, quality—of our cultural world, there still remain strong differences between the usual British and American writing styles. The question is, how do you get your message across? English prose conveys its most novel ideas as if they were timeless truths, while American writing exaggerates; if you believe half of what is said, that's enough. The former uses understatement; the latter, overstatement. There are also disadvantages to each characteristic approach. Readers who are used to being screamed at may not listen when someone chooses to whisper politely. At the same time, the individual who is used to a quiet manner may reject a series of loud imperatives.

The scoring scale for this passage provided a four-point scale for pronunciation and for fluency, as shown in the box below.

Test of Spoken English scoring scale (1987, p. 10)

Pronunciation:

Points:	
0.0–0.4	Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be unintelligible.
0.5–1.4	Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be occasionally unintelligible.
1.5–2.4	Some consistent phonemic errors and foreign stress and intonation patterns, but the speaker is intelligible.
2.5–3.0	Occasional non-native pronunciation errors, but the speaker is always intelligible.

Fluency:	
Points:	
0.0–0.4	Speech is so halting and fragmentary or has such a non-native flow that intelligibility is virtually impossible.
0.5–1.4	Numerous non-native pauses and/or a non-native flow that interferes with intelligibility.
1.5–2.4	Some non-native pauses but with a more nearly native flow so that the pauses do not interfere with intelligibility.
2.5–3.0	Speech is smooth and effortless, closely approximating that of a native speaker.

Such a rating list does not indicate how to gauge *intelligibility*, which is mentioned in both lists. Such slippery terms remind us that oral production scoring, even with the controls that reading aloud offers, is still an inexact science.

Underhill (1987, pp. 77–78) suggested some variations on the task of simply reading a short passage:

- reading a scripted dialogue, with someone else reading the other part
- reading sentences containing minimal pairs, for example:
  Try not to heat/hit the pan too much.
  The doctor gave me a bill/pill.
- reading information from a table or chart

If reading aloud shows certain practical advantages (predictable output, practicality, reliability in scoring), there are several drawbacks to using this technique for assessing oral production. Reading aloud is somewhat inauthentic in that we seldom read anything aloud to someone else in the real world, with the exception of a parent reading to a child, occasionally sharing a written story with someone, or giving a scripted oral presentation. Also, reading aloud calls on certain specialized oral abilities that may not indicate one's pragmatic ability to communicate orally in face-to-face contexts. You should therefore employ this technique with some caution, and certainly supplement it as an assessment task with other, more communicative procedures.

Sentence/Dialogue Completion Tasks and Oral Questionnaires

Another technique for targeting intensive aspects of language requires test-takers to read dialogue in which one speaker's lines have been omitted. Test-takers are first given time to read through the dialogue to get its gist and to think about appropriate lines to fill in. Then as the tape, teacher, or test administrator produces one part orally, the test-taker responds. Here's an example.

Dialogue completion task

Test-takers read (and then hear):

In a department store:

Salesperson: May I help you?
Customer: _____.

Salesperson: Okay, what size do you wear?
Customer: _____.

Salesperson: Hmmm. How about this green sweater here?
Customer: _____.

Salesperson: Oh. Well, if you don't like green, what color would you like?
Customer: _____.

Salesperson: How about this one?
Customer: _____.

Salesperson: Great!
Customer: _____.

Salesperson: It's on sale today for $39.95.
Customer: _____.

Salesperson: Sure, we take Visa, MasterCard, and American Express.
Customer: _____.

Test-takers respond with appropriate lines.

An advantage of this technique lies in its moderate control of the output of the test-taker. While individual variations in responses are accepted, the technique taps into a learner's ability to discern expectancies in a conversation and to produce sociolinguistically correct language. One disadvantage of this technique is its reliance on literacy and an ability to transfer easily from written to spoken English. Another disadvantage is the contrived, inauthentic nature of this task: Couldn't the same criterion performance be elicited in a live interview in which an impromptu role-play technique is used?

Perhaps more useful is a whole host of shorter dialogues of two or three lines, each of which aims to elicit a specified target. In the following examples, somewhat unrelated items attempt to elicit the past tense, future tense, *yes/no* question formation, and asking for the time. Again, test-takers see the stimulus in written form.

Directed response tasks

Test-takers see:

Interviewer: What did you do last weekend?
Test-taker: _____ .

Interviewer: What will you do after you graduate from this program?
Test-taker: _____ .

Test-taker: _____ ?
Interviewer: I was in Japan for two weeks.

Test-taker: _____ ?
Interviewer: It's ten-thirty.

Test-takers respond with appropriate lines.

One could contend that performance on these items is *responsive,* rather than ***intensive.*** True, the discourse involves responses, but there is a degree of control here that predisposes the test-taker to respond with certain expected forms. Such arguments underscore the fine lines of distinction between and among the selected five categories.

It could also be argued that such techniques are nothing more than a written form of questions that might otherwise (and more appropriately) be part of a standard oral interview. True, but the advantage that the written form offers is to provide a little more time for the test-taker to anticipate an answer, and it begins to remove the potential ambiguity created by aural misunderstanding. It helps to unlock the almost ubiquitous link between listening and speaking performance.

Underhill (1987) describes yet another technique that is useful for controlling the test-taker's output: form-filling, or what I might rename "oral questionnaire." Here the test-taker sees a questionnaire that asks for certain categories of information (personal data, academic information, job experience, etc.) and supplies the information orally.

Picture-Cued Tasks

One of the more popular ways to elicit oral language performance at both intensive and extensive levels is a picture-cued stimulus that requires a description from the test-taker. Pictures may be very simple, designed to elicit a word or a phrase; somewhat more elaborate and "busy"; or composed of a series that tells a story or incident. Here is an example of a picture-cued elicitation of the production of a simple minimal pair.

Picture-cued elicitation of minimal pairs

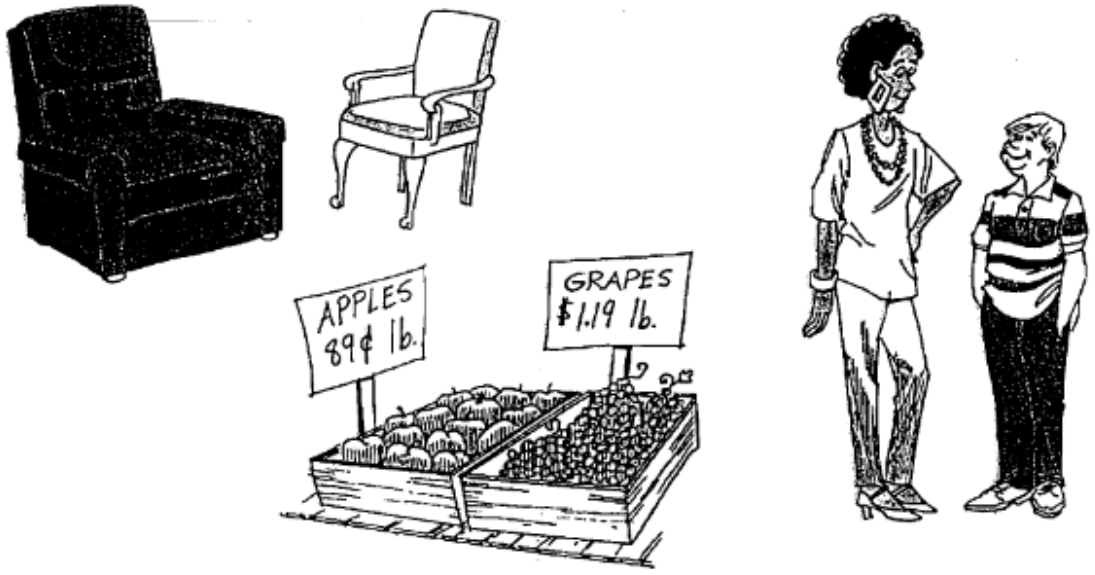
Test-takers see:

Test-takers hear: [*test administrator points to each picture in succession*]
What's this?

Grammatical categories may be cued by pictures. In the following sequences, comparatives are elicited:

Picture-cued elicitation of comparatives (Brown & Sahni, 1994, p. 135)

Test-takers see:



Test-takers hear: Use a comparative form to compare these objects.

Scoring responses on picture-cued intensive speaking tasks varies, depending on the expected performance criteria. The tasks above that asked just for one-word or simple-sentence responses can be evaluated simply as "correct" or "incorrect." The three-point rubric (2, 1, and 0) suggested earlier may apply as well, with these modifications:

Scoring scale for intensive tasks

2	comprehensible; acceptable target form
1	comprehensible; partially correct target form
0	silence, or seriously incorrect target form

Opinions about paintings, persuasive monologue, and directions on a map create a more complicated problem for scoring. More demand is placed on the test administrator to make calculated judgments, in which case a modified form of a scale such as the one suggested for evaluating interviews (below) could be used:

- grammar
- vocabulary
- comprehension
- fluency
- pronunciation
- task (accomplishing the objective of the elicited task)

Each category may be scored separately, with an additional composite score that attempts to synthesize overall performance. To attend to so many factors, you will probably need to have an audiotaped recording for multiple listening.

One moderately successful picture-cued technique involves a pairing of two test-takers. They are supplied with a set of four identical sets of numbered pictures, each minimally distinct from the others by one or two factors. One test-taker is directed by a cue card to describe *one* of the four pictures in as few words as possible. The second test-taker must then identify the picture. On the next page is an example of four pictures.

Translation (of Limited Stretches of Discourse)

Translation is a part of our tradition in language teaching that we tend to discount or disdain, if only because our current pedagogical stance plays down its importance. Translation methods of teaching are certainly passé in an era of direct approaches to creating communicative classrooms. But we should remember that in countries where English is not the native or prevailing language, translation is a meaningful communicative device in contexts where the English user is called on to be an interpreter. Also, translation is a well-proven communication strategy for learners of a second language.

Under certain constraints, then, it is not far-fetched to suggest translation as a device to check oral production. Instead of offering pictures or written stimuli, the test-taker is given a native language word, phrase, or sentence and is asked to translate it. Conditions may vary from expecting an instant translation of an orally elicited linguistic target to allowing more thinking time before producing a translation of somewhat longer texts, which may optionally be offered to the test-taker in written form. (Translation of extensive texts is discussed at the end of this chapter.) As an assessment procedure, the advantages of translation lie in its control of the output of the test-taker, which of course means that scoring is more easily specified.

DESIGNING ASSESSMENT TASKS: RESPONSIVE SPEAKING

Assessment of responsive tasks involves brief interactions with an interlocutor, differing from intensive tasks in the increased creativity given to the test-taker and from interactive tasks by the somewhat limited length of utterances.

Question and Answer

Question-and-answer tasks can consist of one or two questions from an interviewer, or they can make up a portion of a whole battery of questions and prompts in an oral interview. They can vary from simple questions like "What is this called in English?" to complex questions like "What are the steps governments should take, if any, to stem the rate of deforestation in tropical countries?" The first question is intensive in its purpose; it is a **display question** intended to elicit a predetermined correct response. We have already looked at some of these types of questions in the previous section. Questions at the responsive level tend to be genuine **referential questions** in which the test-taker is given more opportunity to produce meaningful language in response.

In designing such questions for test-takers, it's important to make sure that you know *why* you are asking the question. Are you simply trying to elicit strings of language output to gain a general sense of the test-taker's discourse competence? Are you combining discourse and grammatical competence in the same question? Is each question just one in a whole set of related questions? Responsive questions may take the following forms:

Questions eliciting open-ended responses

Test-takers hear:

1. What do you think about the weather today?
2. What do you like about the English language?
3. Why did you choose your academic major?
4. What kind of strategies have you used to help you learn English?
5. a. Have you ever been to the United States before?
   b. What other countries have you visited?
   c. Why did you go there? What did you like best about it?
   d. If you could go back, what would you like to do or see?
   e. What country would you like to visit next, and why?

Test-takers respond with a few sentences at most.

Notice that question #5 has five situationally linked questions that may vary slightly depending on the test-taker's response to a previous question.

Oral interaction with a test administrator often involves the latter forming all the questions. The flip side of this usual concept of question-and-answer tasks is to elicit questions from the test-taker. To assess the test-taker's ability to produce questions, prompts such as this can be used:

Elicitation of questions from the test-taker

Test-takers hear:

- Do you have any questions for me?
- Ask me about my family or job or interests.
- If you could interview the president or prime minister of your country, what would you ask that person?

Test-takers respond with questions.

A potentially tricky form of oral production assessment involves more than one test-taker with an interviewer, which is discussed later in this chapter. With two students in an interview context, both test-takers can ask questions of each other.

Giving Instructions and Directions

We are all called on in our daily routines to read instructions on how to operate an appliance, how to put a bookshelf together, or how to create a delicious clam chowder. Somewhat less frequent is the mandate to provide such instructions orally, but this speech act is still relatively common. Using such a stimulus in an assessment context provides an opportunity for the test-taker to engage in a relatively extended stretch of discourse, to be very clear and specific, and to use appropriate discourse markers and connectors. The technique is simple: the administrator poses the problem, and the test-taker responds. Scoring is based primarily on comprehensibility and secondarily on other specified grammatical or discourse categories. Here are some possibilities.

Eliciting instructions or directions

Test-takers hear:

- Describe how to make a typical dish from your country.
- What's a good recipe for making _____?
- How do you access email on a PC computer?
- How would I make a typical costume for a _____ celebration in your country?
- How do you program telephone numbers into a cell (mobile) phone?
- How do I get from _____ to _____ in your city?

Test-takers respond with appropriate instructions/directions.

Some pointers for creating such tasks: The test administrator needs to guard against test-takers knowing and preparing for such items in advance lest they simply parrot back a memorized set of sentences. An impromptu delivery of instructions is warranted here, or at most a minute or so of preparation time. Also, the choice of topics needs to be familiar enough so that you are testing not general knowledge but linguistic competence; therefore, topics beyond the content schemata of the test-taker are inadvisable. Finally, the task should require the test-taker to produce at least five or six sentences (of connected discourse) to adequately fulfill the objective.

This task can be designed to be more complex, thus placing it in the category of extensive speaking. If your objective is to keep the response short and simple, then make sure your directive does not take the test-taker down a path of complexity that he or she is not ready to face.

Paraphrasing

Another type of assessment task that can be categorized as responsive asks the test-taker to read or hear a limited number of sentences (perhaps two to five) and produce a paraphrase of the sentence. For example:

Paraphrasing a story

Test-takers hear: Paraphrase the following little story in your own words.

My weekend in the mountains was fabulous. The first day we backpacked into the mountains and climbed about 2,000 feet. The hike was strenuous but exhilarating. By sunset we found these beautiful alpine lakes and made camp there. The sunset was amazingly beautiful. The next two days we just kicked back and did little day hikes, some rock climbing, bird watching, swimming, and fishing. The hike out on the next day was really easy—all downhill—and the scenery was incredible.

Test-takers respond with two or three sentences.

A more authentic context for paraphrase is aurally receiving and orally relaying a message. In the example below, the test-taker must relay information from a telephone call to an office colleague named Jeff.

Paraphrasing a phone message

Test-takers hear:

Please tell Jeff that I'm tied up in traffic so I'm going to be about a half hour late for the nine o'clock meeting. And ask him to bring up our question about the employee benefits plan. If he wants to check in with me on my cell phone, have him call 415-338-3095. Thanks.

Test-takers respond with two or three sentences.

The advantages of such tasks are that they elicit short stretches of output and perhaps tap into test-takers' ability to practice the conversational art of conciseness by reducing the output/input ratio. Yet you have to question the criterion being assessed. Is it a listening task more than production? Does it test short-term memory rather than linguistic ability? And how does the teacher determine scoring of responses? If you use short paraphrasing tasks as an assessment procedure, it's important to pinpoint the objective of the task clearly. In this case, the integration of listening and speaking is probably more at stake than simple oral production alone.

DESIGNING ASSESSMENT TASKS: INTERACTIVE SPEAKING

The final two categories of oral production assessment (interactive and extensive speaking) include tasks that involve relatively long stretches of interactive discourse (interviews, role plays, discussions, games) and tasks of equally long duration but that involve less interaction (speeches, telling longer stories, and extended explanations and translations). The obvious difference between the two sets of tasks is the degree of interaction with an interlocutor. Also, interactive tasks are what some would describe as **interpersonal,** while the final category includes more **transactional** speech events.

Interview

When "oral production assessment" is mentioned, the first thing that comes to mind is an oral interview: a test administrator and a test-taker sit down in a direct face-to-face exchange and proceed through a protocol of questions and directives. The interview, which may be tape-recorded for re-listening, is then scored on one or more parameters such as accuracy in pronunciation and/or grammar, vocabulary usage, fluency, sociolinguistic/pragmatic appropriateness, task accomplishment, and even comprehension.

Interviews can vary in length from perhaps five to forty-five minutes, depending on their purpose and context. Placement interviews, designed to get a quick spoken sample from a student in order to verify placement into a course, may need only five minutes if the interviewer is trained to evaluate the output accurately. Longer comprehensive interviews such as the OPI (see the next section) are designed to cover predetermined oral production contexts and may require the better part of an hour.

Every effective interview contains a number of mandatory stages. Two decades ago, Michael Canale (1984) proposed a framework for oral proficiency testing that has withstood the test of time. He suggested that test-takers will perform at their best if they are led through four stages:

1. *Warm-up.* In a minute or so of preliminary small talk, the interviewer directs mutual introductions, helps the test-taker become comfortable with the situation, apprises the test-taker of the format, and allays anxieties. No scoring of this phase takes place.
2. *Level check.* Through a series of preplanned questions, the interviewer stimulates the test-taker to respond using expected or predicted forms and functions. If, for example, from previous test information, grades, or other data, the test-taker has been judged to be a "Level 2" (see below) speaker, the interviewer's prompts will attempt to confirm this assumption. The responses may take very simple or very complex form, depending on the entry level of the learner. Questions are usually designed to elicit grammatical categories (such as past tense or subject–verb agreement), discourse structure (a sequence of events), vocabulary usage, and/or sociolinguistic factors (politeness conventions, formal/informal language). This stage could also give the interviewer a picture of the test-taker's extroversion, readiness to speak, and confidence, all of which may be of significant consequence in the interview's results. Linguistic target criteria are scored in this phase. If this stage is lengthy, a tape-recording of the interview is important.
3. *Probe.* Probe questions and prompts challenge test-takers to go to the heights of their ability, to extend beyond the limits of the interviewer's expectation through increasingly difficult questions. Probe questions may be complex in their framing and/or complex in their cognitive and linguistic demand. Through probe items, the interviewer discovers the ceiling or limitation of the test-taker's proficiency. This need not be a separate stage entirely, but might be a set of questions that are interspersed into the previous stage. At the lower levels of proficiency, probe items may simply demand a higher range of vocabulary or grammar from the test-taker than predicted. At the higher levels, probe items will typically ask the test-taker to give an opinion or a value judgment, to discuss his or her field of specialization, to recount a narrative, or to respond to questions that are worded in complex form. Responses to probe questions may be scored, or they may be ignored if the test-taker displays an inability to handle such complexity.
4. *Wind-down.* This final phase of the interview is simply a short period of time during which the interviewer encourages the test-taker to relax with some easy questions, sets the test-taker's mind at ease, and provides information about when and where to obtain the results of the interview. This part is not scored.

Table 7.2. Oral proficiency scoring categories (Brown, 2001, pp. 406–407)

	Grammar	Vocabulary	Comprehension
I	Errors in grammar are frequent, but speaker can be understood by a native speaker used to dealing with foreigners attempting to speak his language.	Speaking vocabulary inadequate to express anything but the most elementary needs.	Within the scope of his very limited language experience, can understand simple questions and statements if delivered with slowed speech, repetition, or paraphrase.
II	Can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.	Has speaking vocabulary sufficient to express himself simply with some circumlocutions.	Can get the gist of most conversations of non-technical subjects (i.e., topics that require no specialized knowledge).
III	Control of grammar is good. Able to speak the language with sufficient structural accuracy to participate effectively in most formal and informal conversations on practical, social, and professional topics.	Able to speak the language with sufficient vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Vocabulary is broad enough that he rarely has to grope for a word.	Comprehension is quite complete at a normal rate of speech.
IV	Able to use the language accurately on all levels normally pertinent to professional needs. Errors in grammar are quite rare.	Can understand and participate in any conversation within the range of his experience with a high degree of precision of vocabulary.	Can understand any conversation within the range of his experience.
V	Equivalent to that of an educated native speaker.	Speech on all levels is fully accepted by educated native speakers in all its features including breadth of vocabulary and idioms, colloquialisms, and pertinent cultural references.	Equivalent to that of an educated native speaker.

Fluency	Pronunciation	Task
(No specific fluency description. Refer to other four language areas for implied level of fluency.)	Errors in pronunciation are frequent but can be understood by a native speaker used to dealing with foreigners attempting to speak his language.	Can ask and answer questions on topics very familiar to him. Able to satisfy routine travel needs and minimum courtesy requirements. (Should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases, and tell time.)
Can handle with confidence but not with facility most social situations, including introductions and casual conversations about current events, as well as work, family, and autobiographical information.	Accent is intelligible though often quite faulty.	Able to satisfy routine social demands and work requirements; needs help in handling any complication or difficulties.
Can discuss particular interests of competence with reasonable ease. Rarely has to grope for words.	Errors never interfere with understanding and rarely disturb the native speaker. Accent may be obviously foreign.	Can participate effectively in most formal and informal conversations on practical, social, and professional topics.
Able to use the language fluently on all levels normally pertinent to professional needs. Can participate in any conversation within the range of this experience with a high degree of fluency.	Errors in pronunciation are quite rare.	Would rarely be taken for a native speaker but can respond appropriately even in unfamiliar situations. Can handle informal interpreting from and into language.
Has complete fluency in the language such that his speech is fully accepted by educated native speakers.	Equivalent to and fully accepted by educated native speakers.	Speaking proficiency equivalent to that of an educated native speaker.

Table 7.3. Subcategories of oral proficiency scores

Level	Description
0	Unable to function in the spoken language
0+	Able to satisfy immediate needs using rehearsed utterances
1	Able to satisfy minimum courtesy requirements and maintain very simple face-to-face conversations on familiar topics
1+	Can initiate and maintain predictable face-to-face conversations and satisfy limited social demands
2	Able to satisfy routine social demands and limited work requirements
2+	Able to satisfy most work requirements with language usage that is often, but not always, acceptable and effective
3	Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics
3+	Often able to use the language to satisfy professional needs in a wide range of sophisticated and demanding tasks
4	Able to use the language fluently and accurately on all levels normally pertinent to professional needs
4+	Speaking proficiency is regularly superior in all respects, usually equivalent to that of a well-educated, highly articulate native speaker
5	Speaking proficiency is functionally equivalent to that of a highly articulate, well-educated native speaker and reflects the cultural standards of the country where the language is spoken

Role Play

Role playing is a popular pedagogical activity in communicative language-teaching classes. Within constraints set forth by the guidelines, it frees students to be somewhat creative in their linguistic output. In some versions, role play allows some rehearsal time so that students can map out what they are going to say. And it has the effect of lowering anxieties as students can, even for a few moments, take on the persona of someone other than themselves.

As an assessment device, role play opens some windows of opportunity for test-takers to use discourse that might otherwise be difficult to elicit. With prompts such as "Pretend that you're a tourist asking me for directions" or "You're buying a necklace from me in a flea market, and you want to get a lower price," certain personal, strategic, and linguistic factors come into the foreground of the test-taker's oral abilities. While role play can be controlled or "guided" by the interviewer, this technique takes test-takers beyond simple intensive and responsive levels to a level of creativity and complexity that approaches real-world pragmatics. Scoring presents the usual issues in any task that elicits somewhat unpredictable responses from test-takers. The test administrator must determine the assessment objectives of the role play, then devise a scoring technique that appropriately pinpoints those objectives.

Discussions and Conversations

As formal assessment devices, discussions and conversations with and among students are difficult to specify and even more difficult to score. But as *informal* techniques to assess learners, they offer a level of authenticity and spontaneity that other assessment techniques may not provide. Discussions may be especially appropriate tasks through which to elicit and observe such abilities as

- topic nomination, maintenance, and termination;
- attention getting, interrupting, floor holding, control;
- clarifying, questioning, paraphrasing;
- comprehension signals (nodding, "uh-huh," "hmm," etc.);
- negotiating meaning;
- intonation patterns for pragmatic effect;
- kinesics, eye contact, proxemics, body language; and
- politeness, formality, and other sociolinguistic factors.

Assessing the performance of participants through scores or checklists (in which appropriate or inappropriate manifestations of any category are noted) should be carefully designed to suit the objectives of the observed discussion. Of course, discussion is an integrative task, and so it is also advisable to give some cognizance to comprehension performance in evaluating learners.

Games

Among informal assessment devices are a variety of games that directly involve language production. Consider the following types:

Assessment games

1. "Tinkertoy" game: A Tinkertoy (or Lego block) structure is built behind a screen. One or two learners are allowed to view the structure. In successive stages of construction, the learners tell "runners" (who can't observe the structure) how to re-create the structure. The runners then tell "builders" behind another screen how to build the structure. The builders may question or confirm as they proceed, but only through the two degrees of separation. Object: re-create the structure as accurately as possible.
2. Crossword puzzles are created in which the names of all members of a class are clued by obscure information about them. Each class member must ask questions of others to determine who matches the clues in the puzzle.

3. Information gap grids are created such that class members must conduct mini-interviews of other classmates to fill in boxes, e.g., "born in July," "plays the violin," "has a two-year-old child," etc.
4. City maps are distributed to class members. Predetermined map directions are given to one student who, with a city map in front of him or her, describes the route to a partner, who must then trace the route and get to the correct final destination.

Clearly, such tasks have wandered away from the traditional notion of an oral production test and may even be well beyond *assessments*, but if you remember the discussion of these terms in Chapter 1 of this book, you can put the tasks into perspective. As assessments, the key is to specify a set of criteria and a reasonably practical and reliable scoring method. The benefit of such an informal assessment may not be as much in a summative evaluation as in its formative nature, with washback for the students.

DESIGNING ASSESSMENTS: EXTENSIVE SPEAKING

Extensive speaking tasks involve complex, relatively lengthy stretches of discourse. They are frequently variations on monologues, usually with minimal verbal interaction.

Oral Presentations

In the academic and professional arenas, it would not be uncommon to be called on to present a report, a paper, a marketing plan, a sales idea, a design of a new product, or a method. A summary of oral assessment techniques would therefore be incomplete without some consideration of extensive speaking tasks. Once again the rules for effective assessment must be invoked: (a) specify the criterion, (b) set appropriate tasks, (c) elicit optimal output, and (d) establish practical, reliable scoring procedures. And once again scoring is the key assessment challenge.

For oral presentations, a checklist or grid is a common means of scoring or evaluation. Holistic scores are tempting to use for their apparent practicality, but they may obscure the variability of performance across several subcategories, especially the two major components of content and delivery. Following is an example of a checklist for a prepared oral presentation at the intermediate or advanced level of English.

DESIGNING ASSESSMENTS: EXTENSIVE SPEAKING

Extensive speaking tasks involve complex, relatively lengthy stretches of discourse. They are frequently variations on monologues, usually with minimal verbal interaction.

Oral Presentations

In the academic and professional arenas, it would not be uncommon to be called on to present a report, a paper, a marketing plan, a sales idea, a design of a new product, or a method. A summary of oral assessment techniques would therefore be incomplete without some consideration of extensive speaking tasks. Once again the rules for effective assessment must be invoked: (a) specify the criterion, (b) set appropriate tasks, (c) elicit optimal output, and (d) establish practical, reliable scoring procedures. And once again scoring is the key assessment challenge.

For oral presentations, a checklist or grid is a common means of scoring or evaluation. Holistic scores are tempting to use for their apparent practicality, but they may obscure the variability of performance across several subcategories, especially the two major components of content and delivery. Following is an example of a checklist for a prepared oral presentation at the intermediate or advanced level of English.

Retelling a Story, News Event

In this type of task, test-takers hear or read a story or news event that they are asked to retell. This differs from the paraphrasing task discussed above (pages 161–162) in that it is a longer stretch of discourse and a different genre. The objectives in assigning such a task vary from listening comprehension of the original to production of a number of oral discourse features (communicating sequences and relationships of events, stress and emphasis patterns, "expression" in the case of a dramatic story), fluency, and interaction with the hearer. Scoring should of course meet the intended criteria.

Translation (of Extended Prose)

Translation of words, phrases, or short sentences was mentioned under the category of intensive speaking. Here, longer texts are presented for the test-taker to read in the native language and then translate into English. Those texts could come in many forms: dialogue, directions for assembly of a product, a synopsis of a story or play or movie, directions on how to find something on a map, and other genres. The advantage of translation is in the control of the content, vocabulary, and, to some extent, the grammatical and discourse features. The disadvantage is that translation of longer texts is a highly specialized skill for which some individuals obtain post-baccalaureate degrees! To judge a nonspecialist's oral language ability on such a skill may be completely invalid, especially if the test-taker has not engaged in translation at this level. Criteria for scoring should therefore take into account not only the purpose in stimulating a translation but the possibility of errors that are unrelated to oral production ability.

ASSESSING WRITING

Before looking at specific tasks, we must scrutinize the different genres of written language (so that context and purpose are clear), types of writing (so that stages of the development of writing ability are accounted for), and micro- and macroskills of writing (so that objectives can be pinpointed precisely).

GENRES OF WRITTEN LANGUAGE

Chapter 8's discussion of assessment of reading listed more than 50 written language genres. The same classification scheme is reformulated here to include the most common genres that a second language *writer* might produce, within and beyond the requirements of a curriculum. Even though this list is slightly shorter, you should be aware of the surprising multiplicity of options of written genres that second language learners need to acquire.

Genres of writing

1. Academic writing

- papers and general subject reports
- essays, compositions
- academically focused journals
- short-answer test responses
- technical reports (e.g., lab reports)
- theses, dissertations

2. Job-related writing

- messages (e.g., phone messages)
- letters/emails
- memos (e.g., interoffice)
- reports (e.g., job evaluations, project reports)
- schedules, labels, signs
- advertisements, announcements
- manuals

3. Personal writing

- letters, emails, greeting cards, invitations
- messages, notes
- calendar entries, shopping lists, reminders
- financial documents (e.g., checks, tax forms, loan applications)
- forms, questionnaires, medical reports, immigration documents
- diaries, personal journals
- fiction (e.g., short stories, poetry)

TYPES OF WRITING PERFORMANCE

Four categories of written performance that capture the range of written production are considered here. Each category resembles the categories defined for the other three skills, but these categories, as always, reflect the uniqueness of the skill area.

1. Imitative. To produce written language, the learner must attain skills in the fundamental, basic tasks of writing letters, words, punctuation, and very brief sentences. This category includes the ability to spell correctly and to perceive phoneme-grapheme correspondences in the English spelling system. It is a level at which learners are trying to master the mechanics of writing. At this stage, form is the primary if not exclusive focus, while context and meaning are of secondary concern.

2. Intensive (controlled). Beyond the fundamentals of imitative writing are skills in producing appropriate vocabulary within a context, collocations and idioms, and correct grammatical features up to the length of a sentence. Meaning and context are of some importance in determining correctness and appropriateness, but most assessment tasks are more concerned with a focus on form, and are rather strictly controlled by the test design.

3. Responsive. Here, assessment tasks require learners to perform at a limited discourse level, connecting sentences into a paragraph and creating a logically connected sequence of two or three paragraphs. Tasks respond to pedagogical directives, lists of criteria, outlines, and other guidelines. Genres of writing include brief narratives and descriptions, short reports, lab reports, summaries, brief responses to reading, and interpretations of charts or graphs. Under specified conditions, the writer begins to exercise some freedom of choice among alternative forms of expression of ideas. The writer has mastered the fundamentals of sentence-level grammar and is more focused on the discourse conventions that will achieve the objectives of the written text. Form-focused attention is mostly at the discourse level, with a strong emphasis on context and meaning.

4. Extensive. Extensive writing implies successful management of all the processes and strategies of writing for all purposes, up to the length of an essay, a term paper, a major research project report, or even a thesis. Writers focus on achieving a purpose, organizing and developing ideas logically, using details to support or illustrate ideas, demonstrating syntactic and lexical variety, and in many cases, engaging in the process of multiple drafts to achieve a final product. Focus on grammatical form is limited to occasional editing or proofreading of a draft.

MICRO- AND MACROSKILLS OF WRITING

We turn once again to a taxonomy of micro- and macroskills that will assist you in defining the ultimate criterion of an assessment procedure. The earlier microskills apply more appropriately to imitative and intensive types of writing task, while the macroskills are essential for the successful mastery of responsive and extensive writing.

Micro- and macroskills of writing

Microskills

1. Produce graphemes and orthographic patterns of English.
2. Produce writing at an efficient rate of speed to suit the purpose.
3. Produce an acceptable core of words and use appropriate word order patterns.
4. Use acceptable grammatical systems (e.g., tense, agreement, pluralization), patterns, and rules.
5. Express a particular meaning in different grammatical forms.
6. Use cohesive devices in written discourse.

Macroskills

7. Use the rhetorical forms and conventions of written discourse.
8. Appropriately accomplish the communicative functions of written texts according to form and purpose.
9. Convey links and connections between events, and communicate such relations as main idea, supporting idea, new information, given information, generalization, and exemplification.
10. Distinguish between literal and implied meanings when writing.
11. Correctly convey culturally specific references in the context of the written text.
12. Develop and use a battery of writing strategies, such as accurately assessing the audience's interpretation, using prewriting devices, writing with fluency in the first drafts, using paraphrases and synonyms, soliciting peer and instructor feedback, and using feedback for revising and editing.

DESIGNING ASSESSMENT TASKS: IMITATIVE WRITING

With the recent worldwide emphasis on teaching English at young ages, it is tempting to assume that every English learner knows how to handwrite the Roman alphabet. Such is not the case. Many beginning-level English learners, from young children to older adults, need basic training in and assessment of imitative writing: the rudiments of forming letters, words, and simple sentences. We examine this level of writing first.

Tasks in [Hand] Writing Letters, Words, and Punctuation

First, a comment should be made on the increasing use of personal and laptop computers and handheld instruments for creating written symbols. Handwriting has the potential of becoming a lost art as even very young children are more and more likely to use a keyboard to produce writing. Making the shapes of letters and other symbols is now more a question of learning typing skills than of training the muscles

of the hands to use a pen or pencil. Nevertheless, for all practical purposes, handwriting remains a skill of paramount importance within the larger domain of language assessment.

A limited variety of types of tasks are commonly used to assess a person's ability to produce written letters and symbols. A few of the more common types are described here.

1. Copying. There is nothing innovative or modern about directing a test-taker to copy letters or words. The test-taker will see something like the following:

Handwriting letters, words, and punctuation marks

The test-taker reads: Copy the following words in the spaces given:

bit	bet	bat	but	Oh?	Oh!
____	____	____	____	____	____
bin	din	gin	pin	Hello, John.	
____	____	____	____	____	

2. Listening cloze selection tasks. These tasks combine dictation with a written script that has a relatively frequent deletion ratio (every fourth or fifth word, perhaps). The test sheet provides a list of missing words from which the test-taker must select. The purpose at this stage is not to test spelling but to give practice in writing. To increase the difficulty, the list of words can be deleted, but then spelling might become an obstacle. Probes look like this:

Listening cloze selection task

Test-takers hear:
Write the missing word in each blank. Below the story is a list of words to choose from.

Have you ever visited San Francisco? It is a very nice city. It is cool in the summer and warm in the winter. I like the cable cars and bridges.

Test-takers see:

Have _____ ever visited San Francisco? It _____ a very nice _____. It is _____ in _____ summer and _____ in the winter. I _____ the cable cars _____ bridges.

is	you	cool	city
like	and	warm	the

3. *Picture-cued tasks.* Familiar pictures are displayed, and test-takers are told to write the word that the picture represents. Assuming no ambiguity in identifying the picture (cat, hat, chair, table, etc.), no reliance is made on aural comprehension for successful completion of the task.

4. *Form completion tasks.* A variation on pictures is the use of a simple form (registration, application, etc.) that asks for name, address, phone number, and other data. Assuming, of course, that prior classroom instruction has focused on filling out such forms, this task becomes an appropriate assessment of simple tasks such as writing one's name and address.

5. *Converting numbers and abbreviations to words.* Some tests have a section on which numbers are written—for example, hours of the day, dates, or schedules—and test-takers are directed to write out the numbers. This task can serve as a reasonably reliable method to stimulate handwritten English. It lacks authenticity, however, in that people rarely write out such numbers (except in writing checks), and it is more of a reading task (recognizing numbers) than a writing task. If you plan to use such a method, be sure to specify exactly what the criterion is, and then proceed with some caution. Converting abbreviations to words is more authentic: we actually do have occasions to write out days of the week, months, and words like *street, boulevard, telephone,* and *April* (months of course are often abbreviated with numbers). Test tasks may take this form:

Writing numbers and abbreviations

Test-takers hear: Fill in the blanks with words.

Test-takers see:

9:00 _____ 5:45 _____

Tues. _____ 5/3 _____

726 S. Main St. _____

Spelling Tasks and Detecting Phoneme–Grapheme Correspondences

A number of task types are in popular use to assess the ability to spell words correctly and to process phoneme-grapheme correspondences.

1. *Spelling tests.* In a traditional, old-fashioned spelling test, the teacher dictates a simple list of words, one word at a time, followed by the word in a sentence, repeated again, with a pause for test-takers to write the word. Scoring emphasizes correct spelling. You can help to control for listening errors by choosing words that

the students have encountered before—words that they have spoken or heard in their class.

2. Picture-cued tasks. Pictures are displayed with the objective of focusing on familiar words whose spelling may be unpredictable. Items are chosen according to the objectives of the assessment, but this format is an opportunity to present some challenging words and word pairs: ***boot/book, read/reed, bit/bite,*** etc.

3. Multiple-choice techniques. Presenting words and phrases in the form of a multiple-choice task risks crossing over into the domain of assessing reading, but if the items have a follow-up writing component, they can serve as formative reinforcement of spelling conventions. They might be more challenging with the addition of homonyms (see item #3 below). Here are some examples.

Multiple-choice reading-writing spelling tasks

Test-takers read:

Choose the word with the correct spelling to fit the sentence, then write the word in the space provided.

1. He washed his hands with _____.
   A. soap
   B. sope
   C. sop
   D. soup
2. I tried to stop the car, but the _____ didn't work.
   A. braicks
   B. brecks
   C. brakes
   D. bracks
3. The doorbell rang, but when I went to the door, no one was _____.
   A. their
   B. there
   C. they're
   D. thair

4. Matching phonetic symbols. If students have become familiar with the phonetic alphabet, they could be shown phonetic symbols and asked to write the correctly spelled word alphabetically. This works best with letters that do not have one-to-one correspondence with the phonetic symbol (e.g., /æ/ and a). In the sample below, the answers, which of course do not appear on the test sheet, are included in brackets for your reference.

Converting phonetic symbols

Test-takers read:

In each of the following words, a letter or combination of letters has been written in a phonetic symbol. Write the word using the regular alphabet.

1. tea /tʃ/ er _____ [teacher]
2. d /e/ _____ [day]

DESIGNING ASSESSMENT TASKS: INTENSIVE (CONTROLLED) WRITING

This next level of writing is what second language teacher training manuals have for decades called **controlled** writing. It may also be thought of as form-focused writing, grammar writing, or simply guided writing. A good deal of writing at this level is **display** writing as opposed to **real** writing: students produce language to display their competence in grammar, vocabulary, or sentence formation, and not necessarily to convey meaning for an authentic purpose. The traditional grammar/vocabulary test has plenty of display writing in it, since the response mode demonstrates only the test-taker's ability to combine or use words correctly. No new information is passed on from one person to the other.

Dictation and Dicto-Comp

In Chapter 6, dictation was described as an assessment of the integration of listening and writing, but it was clear that the primary skill being assessed is listening. Because of its response mode, however, it deserves a second mention in this chapter. Dictation is simply the rendition in writing of what one hears aurally, so it could be classified as an *imitative* type of writing, especially since a proportion of the test-taker's performance centers on correct spelling. Also, because the test-taker must listen to stretches of discourse and in the process insert punctuation, dictation of a

paragraph or more can arguably be classified as a controlled or intensive form of writing. (For a further explanation on administering a dictation, consult Chapter 6, pages 131–132.)

A form of controlled writing related to dictation is a **dicto-comp**. Here, a paragraph is read at normal speed, usually two or three times; then the teacher asks students to rewrite the paragraph from the best of their recollection. In one of several variations of the dicto-comp technique, the teacher, after reading the passage, distributes a handout with key words from the paragraph, in sequence, as cues for the students. In either case, the dicto-comp is genuinely classified as an intensive, if not a responsive, writing task. Test-takers must internalize the content of the passage, remember a few phrases and lexical items as key words, then recreate the story in their own words.

Grammatical Transformation Tasks

In the heyday of structural paradigms of language teaching with slot-filler techniques and slot substitution drills, the practice of making grammatical transformations—orally or in writing—was very popular. To this day, language teachers have also used this technique as an assessment task, ostensibly to measure grammatical competence. Numerous versions of the task are possible:

- Change the tenses in a paragraph.
- Change full forms of verbs to reduced forms (contractions).
- Change statements to *yes/no* or *wh*-questions.
- Change questions into statements.
- Combine two sentences into one using a relative pronoun.
- Change direct speech to indirect speech.
- Change from active to passive voice.

The list of possibilities is almost endless. The tasks are virtually devoid of any meaningful value. Sometimes test designers attempt to add authenticity by providing a context ("Today Doug is doing all these things. Tomorrow he will do the same things again. Write about what Doug will do tomorrow by using the future tense."), but this is just a backdrop for a written substitution task. On the positive side, grammatical transformation tasks are easy to administer and are therefore practical, quite high in scorer reliability, and arguably tap into a knowledge of grammatical *forms* that will be performed through writing. If you are only interested in a person's ability to produce the forms, then such tasks may prove to be justifiable.

Picture-Cued Tasks

A variety of picture-cued controlled tasks have been used in English classrooms around the world. The main advantage in this technique is in detaching the almost ubiquitous reading and writing connection and offering instead a nonverbal means to stimulate written responses.

Vocabulary Assessment Tasks

Most vocabulary study is carried out through reading. A number of assessments of reading recognition of vocabulary were discussed in the previous chapter: multiple-choice techniques, matching, picture-cued identification, cloze techniques, guessing the meaning of a word in context, etc. The major techniques used to assess vocabulary are (a) defining and (b) using a word in a sentence. The latter is the more authentic, but even that task is constrained by a contrived situation in which the test-taker, usually in a matter of seconds, has to come up with an appropriate sentence, which may or may not indicate that the test-taker "knows" the word.

Ordering Tasks

One task at the sentence level may appeal to those who are fond of word games and puzzles: ordering (or reordering) a scrambled set of words into a correct sentence. Here is the way the item format appears.

Reordering words in a sentence

Test-takers read:

Put the words below into the correct order to make a sentence:

1. cold / winter / is / weather / the / in / the
2. studying / what / you / are
3. next / clock / the / the / is / picture / to

Test-takers write:

1. *The weather is cold in the winter.*
2. *What are you studying?*
3. *The clock is next to the picture.*

While this somewhat inauthentic task generates writing performance and may be said to tap into grammatical word-ordering rules, it presents a challenge to test-takers whose learning styles do not dispose them to logical-mathematical problem solving. If sentences are kept very simple (such as #2) with perhaps no more than four or five words, if only one possible sentence can emerge, and if students have practiced the technique in class, then some justification emerges. But once again, as in so many writing techniques, this task involves as much, if not more, reading performance as writing.

Short-Answer and Sentence Completion Tasks

Some types of short-answer tasks were discussed in Chapter 8 because of the heavy participation of reading performance in their completion. Such items range from very simple and predictable to somewhat more elaborate responses. Look at the range of possibilities.

Limited response writing tasks

Test-takers see:

1. Alicia: Who's that?
   Tony: _____ Gina.
   Alicia: Where's she from?
   Tony: _____ Italy.
2. Jennifer: _____?
   Kathy: I'm studying English.
3. Restate the following sentences in your own words, using the underlined word. You may need to change the meaning of the sentence a little.
   3a. I never miss a day of school. always
   3b. I'm pretty healthy most of the time. seldom
   3c. I play tennis twice a week. sometimes
4. You are in the kitchen helping your roommate cook. You need to ask questions about quantities. Ask a question using *how much* (#4a) and a question using *how many* (#4b), using nouns like *sugar, pounds, flour, onions, eggs, cups.*
   4a. _____.
   4b. _____.

DESIGNING ASSESSMENT TASKS: RESPONSIVE AND EXTENSIVE WRITING

In this section we consider both responsive and extensive writing tasks. They will be regarded here as a continuum of possibilities ranging from lower-end tasks whose complexity exceeds those in the previous category of intensive or controlled writing, through more open-ended tasks such as writing short reports, essays, summaries, and responses, up to texts of several pages or more.

Paraphrasing

One of the more difficult concepts for second language learners to grasp is paraphrasing. The initial step in teaching paraphrasing is to ensure that learners understand the importance of paraphrasing: to say something in one's own words, to avoid plagiarizing, to offer some variety in expression. With those possible motivations and purposes in mind, the test designer needs to elicit a paraphrase of a sentence or paragraph, usually not more.

Scoring of the test-taker's response is a judgment call in which the criterion of conveying the same or similar message is primary, with secondary evaluations of discourse, grammar, and vocabulary. Other components of analytic or holistic scales (see discussion below, page 242) might be considered as criteria for an evaluation. Paraphrasing is more often a part of informal and formative assessment than of formal, summative assessment, and therefore student responses should be viewed as opportunities for teachers and students to gain positive washback on the art of paraphrasing.

Guided Question and Answer

Another lower-order task in this type of writing, which has the pedagogical benefit of guiding a learner without dictating the form of the output, is a guided question-and-answer format in which the test administrator poses a series of questions that essentially serve as an outline of the emergent written text. In the writing of a narrative that the teacher has already covered in a class discussion, the following kinds of questions might be posed to stimulate a sequence of sentences.

Guided writing stimuli

1. Where did this story take place? [setting]
2. Who were the people in the story? [characters]
3. What happened first? and then? and then? [sequence of events]
4. Why did _____ do _____? [reasons, causes]
5. What did _____ think about _____? [opinion]

Paragraph Construction Tasks

The participation of reading performance is inevitable in writing effective paragraphs. To a great extent, writing is the art of emulating what one reads. You read an effective paragraph; you analyze the ingredients of its success; you emulate it. Assessment of paragraph development takes on a number of different forms:

1. Topic sentence writing. There is no cardinal rule that says every paragraph must have a topic sentence, but the stating of a topic through the lead sentence (or a subsequent one) has remained as a tried-and-true technique for teaching the concept of a paragraph. Assessment thereof consists of

- specifying the writing of a topic sentence,
- scoring points for its presence or absence, and
- scoring and/or commenting on its effectiveness in stating the topic.

2. Topic development within a paragraph. Because paragraphs are intended to provide a reader with "clusters" of meaningful, connected thoughts or ideas, another stage of assessment is development of an idea within a paragraph. Four criteria are commonly applied to assess the quality of a paragraph:

- the clarity of expression of ideas
- the logic of the sequence and connections
- the cohesiveness or unity of the paragraph
- the overall effectiveness or impact of the paragraph as a whole

3. Development of main and supporting ideas across paragraphs. As writers string two or more paragraphs together in a longer text (and as we move up the continuum from responsive to extensive writing), the writer attempts to articulate a thesis or **main idea** with clearly stated **supporting ideas.** These elements can be considered in evaluating a multi-paragraph essay:

- addressing the topic, main idea, or principal purpose
- organizing and developing supporting ideas
- using appropriate details to undergird supporting ideas
- showing facility and fluency in the use of language
- demonstrating syntactic variety

Strategic Options

Developing main and supporting ideas is the goal for the writer attempting to create an effective text, whether a short one- to two-paragraph one or an extensive one of several pages. A number of strategies are commonly taught to second language writers to accomplish their purposes. Aside from strategies of freewriting, outlining, drafting, and revising, writers need to be aware of the task that has been demanded and to focus on the genre of writing and the expectations of that genre.

1. Attending to task. In responsive writing, the context is seldom completely open-ended: a task has been defined by the teacher or test administrator, and the writer must fulfill the criterion of the task. Even in extensive writing of longer texts, a set of directives has been stated by the teacher or is implied by the conventions of the genre. Four types of tasks are commonly addressed in academic writing courses: compare/contrast, problem/solution, pros/cons, and cause/effect. Depending on the genre of the text, one or more of these task types will be needed to achieve the writer's purpose. If students are asked, for example, to "agree or disagree with the author's statement," a likely strategy would be to cite pros and cons and then take a stand. A task that asks students to argue for one among several political candidates in an election might be an ideal compare-and-contrast context, with an appeal to problems present in the constituency and the relative value of candidates' solutions. Assessment of the fulfillment of such tasks could be formative and informal (comments in marginal notes, feedback in a conference in an editing/revising stage), but the product might also be assigned a holistic or analytic score.

2. Attending to genre. The genres of writing that were listed at the beginning of this chapter provide some sense of the many varieties of text that may be produced by a second language learner in a writing curriculum. Another way of looking at the strategic options open to a writer is the extent to which both the constraints and the opportunities of the genre are exploited. Assessment of any writing necessitates attention to the conventions of the genre in question. Assessment of the more common genres may include the following criteria, along with chosen factors from the list in item #3 (main and supporting ideas) above:

Reports (Lab Reports, Project Summaries, Article/Book Reports, etc.)

- conform to a conventional format (for this case, field)
- convey the purpose, goal, or main idea
- organize details logically and sequentially
- state conclusions or findings
- use appropriate vocabulary and jargon for the specific case

Summaries of Readings/Lectures/Videos

- effectively capture the main and supporting ideas of the original
- maintain objectivity in reporting
- use writer's own words for the most part

- use quotations effectively when appropriate
- omit irrelevant or marginal details
- conform to an expected length

Responses to Readings/Lectures/Videos
- accurately reflect the message or meaning of the original
- appropriately select supporting ideas to respond to
- express the writer's own opinion
- defend or support that opinion effectively
- conform to an expected length

Narration, Description, Persuasion/Argument, and Exposition
- follow expected conventions for each type of writing
- convey purpose, goal, or main idea
- use effective writing strategies
- demonstrate syntactic variety and rhetorical fluency

Interpreting Statistical, Graphic, or Tabular Data
- provides an effective global, overall description of the data
- organizes the details in clear, logical language
- accurately conveys details
- appropriately articulates relationships among elements of the data
- conveys specialized or complex data comprehensibly to a lay reader
- interprets beyond the data when appropriate

Library Research Paper
- states purpose or goal of the research
- includes appropriate citations and references in correct format
- accurately represents others' research findings
- injects writer's own interpretation, when appropriate, and justifies it
- includes suggestions for further research
- sums up findings in a conclusion

SCORING METHODS FOR RESPONSIVE AND EXTENSIVE WRITING

At responsive and extensive levels of writing, three major approaches to scoring writing performance are commonly used by test designers: holistic, primary trait, and analytical. In the first method, a single score is assigned to an essay, which represents a reader's general overall assessment. Primary trait scoring is a variation of the holistic method in that the achievement of the primary purpose, or trait, of an essay is the only factor rated. Analytical scoring breaks a test-taker's written text down into a number of subcategories (organization, grammar, etc.) and gives a separate rating for each.

Holistic Scoring

The TWE scoring scale above is a prime example of **holistic** scoring. In Chapter 7, a rubric for scoring oral production holistically was presented. Each point on a holistic scale is given a systematic set of descriptors, and the reader-evaluator matches an overall impression with the descriptors to arrive at a score. Descriptors usually (but not always) follow a prescribed pattern. For example, the first descriptor across all score categories may address the quality of task achievement, the second may deal with organization, the third with grammatical or rhetorical considerations, and so on. Scoring, however, is truly holistic in that those subsets are not quantitatively added up to yield a score.

Advantages of holistic scoring include

- fast evaluation,
- relatively high inter-rater reliability,
- the fact that scores represent "standards" that are easily interpreted by lay persons,
- the fact that scores tend to emphasize the writer's strengths (Cohen, 1994, p. 315), and
- applicability to writing across many different disciplines.

Its disadvantages must also be weighed into a decision on whether to use holistic scoring:

- One score masks differences across the subskills within each score.
- No diagnostic information is available (no washback potential).
- The scale may not apply equally well to all genres of writing.
- Raters need to be extensively trained to use the scale accurately.

In general, teachers and test designers lean toward holistic scoring only when it is expedient for administrative purposes. As long as trained evaluators are in place, differentiation across six levels may be quite adequate for admission into an institution or placement into courses. For classroom instructional purposes, holistic scores provide very little information. In most classroom settings where a teacher wishes to adapt a curriculum to the needs of a particular group of students, much more differentiated information across subskills is desirable than is provided by holistic scoring.

Primary Trait Scoring

A second method of scoring, **primary trait**, focuses on "how well students can write within a narrowly defined range of discourse" (Weigle, 2002, p. 110). This type of scoring emphasizes the task at hand and assigns a score based on the effectiveness of the text's achieving that one goal. For example, if the purpose or function of

an essay is to *persuade* the reader to do something, the score for the writing would rise or fall on the accomplishment of that function. If a learner is asked to exploit the imaginative function of language by expressing personal feelings, then the response would be evaluated on that feature alone.

For rating the primary trait of the text, Lloyd-Jones (1977) suggested a four-point scale ranging from zero (no response or fragmented response) to 4 (the purpose is unequivocally accomplished in a convincing fashion). It almost goes without saying that organization, supporting details, fluency, syntactic variety, and other features will implicitly be evaluated in the process of offering a primary trait score. But the advantage of this method is that it allows both writer and evaluator to focus on function. In summary, a primary trait score would assess

- the accuracy of the account of the original (summary),
- the clarity of the steps of the procedure and the final result (lab report),
- the description of the main features of the graph (graph description), and
- the expression of the writer's opinion (response to an article).

Analytic Scoring

For classroom instruction, holistic scoring provides little washback into the writer's further stages of learning. Primary trait scoring focuses on the principal function of the text and therefore offers some feedback potential, but no washback for any of the aspects of the written production that enhance the ultimate accomplishment of the purpose. Classroom evaluation of learning is best served through **analytic scoring,** in which as many as six major elements of writing are scored, thus enabling learners to home in on weaknesses and to capitalize on strengths.

Analytic scoring may be more appropriately called analytic *assessment* in order to capture its closer association with classroom language instruction than with formal testing. Brown and Bailey (1984) designed an analytical scoring scale that specified five major categories and a description of five different levels in each category, ranging from "unacceptable" to "excellent" (see Table 9.2).

At first glance, Brown and Bailey's scale may look similar to the TWE® holistic scale discussed earlier: for each scoring category there is a description that encompasses several subsets. A closer inspection, however, reveals much more detail in the analytic method. Instead of just six descriptions, there are 25, each subdivided into a number of contributing factors.

The order in which the five categories (organization, logical development of ideas, grammar, punctuation/spelling/mechanics, and style and quality of expression) are listed may bias the evaluator toward the greater importance of organization and logical development as opposed to punctuation and style. But the mathematical assignment of the 100-point scale gives equal weight (a maximum of 20 points) to each of the five major categories. Not all writing and assessment specialists agree. You might, for example, consider the analytical scoring profile suggested by Jacobs et al. (1981), in which five slightly different categories were given the point values shown on page 246.

Table 9.2. Analytic scale for rating composition tasks (Brown & Bailey, 1984, pp. 39–41)

	20–18 Excellent to Good	17–15 Good to Adequate	14–12 Adequate to Fair	11–6 Unacceptable–not college-level work	5–1
I. **Organization:** Introduction, Body, and Conclusion	Appropriate title, effective introductory paragraph, topic is stated, leads to body; transitional expressions used; arrangement of material shows plan (could be outlined by reader); supporting evidence given for generalizations; conclusion logical and complete	Adequate title, introduction, and conclusion; body of essay is acceptable, but some evidence may be lacking, some ideas aren't fully developed; sequence is logical but transitional expressions may be absent or misused	Mediocre or scant introduction or conclusion; problems with the order of ideas in body; the generalizations may not be fully supported by the evidence given; problems of organization interfere	Shaky or minimally recognizable introduction; organization can barely be seen; severe problems with ordering of ideas; lack of supporting evidence; conclusion weak or illogical; inadequate effort at organization	Absence of introduction or conclusion; no apparent organization of body; severe lack of supporting evidence; writer has not made any effort to organize the composition (could not be outlined by reader)
II. **Logical development of ideas:** Content	Essay addresses the assigned topic; the ideas are concrete and thoroughly developed; no extraneous material; essay reflects thought	Essay addresses the issues but misses some points; ideas could be more fully developed; some extraneous material is present	Development of ideas not complete or essay is somewhat off the topic; paragraphs aren't divided exactly right	Ideas incomplete; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content	Essay is completely inadequate and does not reflect college-level work; no apparent effort to consider the topic carefully
III. **Grammar**	Native-like fluency in English grammar; correct use of relative clauses, prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences	Advanced proficiency in English grammar; some grammar problems don't influence communication, although the reader is aware of them; no fragments or run-on sentences	Ideas are getting through to the reader, but grammar problems are apparent and have a negative effect on communication; run-on sentences or fragments present	Numerous serious grammar problems interfere with communication of the writer's ideas; grammar review of some areas clearly needed; difficult to read sentences	Severe grammar problems interfere greatly with the message; reader can't understand what the writer was trying to say; unintelligible sentence structure
IV. **Punctuation, spelling, and mechanics**	Correct use of English writing conventions: left and right margins, all needed capitals, paragraphs indented, punctuation and spelling; very neat	Some problems with writing conventions or punctuation; occasional spelling errors; left margin correct; paper is neat and legible	Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas	Serious problems with format of paper; parts of essay not legible; errors in sentence punctuation and final punctuation; unacceptable to educated readers	Complete disregard for English writing conventions; paper illegible; obvious capitals missing, no margins, severe spelling problems
V. **Style and quality of expression**	Precise vocabulary usage; use of parallel structures; concise; register good	Attempts variety; good vocabulary; not wordy; register OK; style fairly concise	Some vocabulary misused; lacks awareness of register; may be too wordy	Poor expression of ideas; problems in vocabulary; lacks variety of structure	Inappropriate use of vocabulary; no concept of register or sentence variety

Content	30
Organization	20
Vocabulary	20
Syntax	25
Mechanics	5
Total	**100**

As your curricular goals and students' needs vary, your own analytical scoring of essays may be appropriately tailored. Level of proficiency can make a significant difference in emphasis: at the intermediate level, for example, you might give more emphasis to syntax and mechanics, while advanced levels of writing may call for a strong push toward organization and development. Genre can also dictate variations in scoring. Would a summary of an article require the same relative emphases as a narrative essay? Most likely not. Certain types of writing, such as lab reports or interpretations of statistical data, may even need additional—or at least redefined—categories in order to capture the essential components of good writing within those genres.

Analytic scoring of compositions offers writers a little more washback than a single holistic or primary trait score. Scores in five or six major elements will help to call the writers' attention to areas of needed improvement. Practicality is lowered in that more time is required for teachers to attend to details within each of the categories in order to render a final score or grade, but ultimately students receive more information about their writing. Numerical scores alone, however, are still not sufficient for enabling students to become proficient writers, as we shall see in the next section.

BEYOND SCORING: RESPONDING TO EXTENSIVE WRITING

Formal testing carries with it the burden of designing a practical and reliable instrument that assesses its intended criterion accurately. To accomplish that mission, designers of writing tests are charged with the task of providing as "objective" a scoring procedure as possible, and one that in many cases can be easily interpreted by agents beyond the learner. Holistic, primary trait, and analytic scoring all satisfy those ends. Yet beyond mathematically calculated scores lies a rich domain of assessment in which a developing writer is coached from stage to stage in a process of building a storehouse of writing skills. Here in the classroom, in the tutored relationship of teacher and student, and in the community of peer learners, most of the hard work of assessing writing is carried out. Such assessment is informal, formative, and replete with washback.

Most writing specialists agree that the best way to teach writing is a hands-on approach that stimulates student output and then generates a series of self-assessments, peer editing and revision, and teacher response and conferencing (Raimes, 1991, 1998; Reid, 1993; Seow, 2002). It is not an approach that relies on a massive dose of lecturing

about good writing, nor on memorizing a bunch of rules about rhetorical organization, nor on sending students home with an assignment to turn in a paper the next day. People become good writers by writing and seeking the facilitative input of others to refine their skills.

Assessment takes on a crucial role in such an approach. Learning how to become a good writer places the student in an almost constant stage of assessment. To give the student the maximum benefit of assessment, it is important to consider (a) *earlier* stages (from freewriting to the first draft or two) and (b) *later* stages (revising and finalizing) of producing a written text. A further factor in assessing writing is the involvement of self, peers, and teacher at appropriate steps in the process. (For further guidelines on the process of teaching writing, see *TBP,* Chapter 19.)

Assessing Initial Stages of the Process of Composing

Following are some guidelines for assessing the initial stages (the first draft or two) of a written composition. These guidelines are generic for self, peer, and teacher responding. Each assessor will need to modify the list according to the level of the learner, the context, and the purpose in responding.

Assessment of initial stages in composing

1. Focus your efforts primarily on meaning, main idea, and organization.
2. Comment on the introductory paragraph.
3. Make general comments about the clarity of the main idea and logic or appropriateness of the organization.
4. As a rule of thumb, ignore minor (local) grammatical and lexical errors.
5. Indicate what appear to be major (global) errors (e.g., by underlining the text in question), but allow the writer to make corrections.
6. Do not rewrite questionable, ungrammatical, or awkward sentences; rather, probe with a question about meaning.
7. Comment on features that appear to be irrelevant to the topic.

The teacher-assessor's role is as a guide, a facilitator, and an ally; therefore, assessment at this stage of writing needs to be as positive as possible to encourage the writer. An early focus on overall structure and meaning will enable writers to clarify their purpose and plan and will set a framework for the writers' later refinement of the lexical and grammatical issues.

Assessing Later Stages of the Process of Composing

Once the writer has determined and clarified his or her purpose and plan, and has completed at least one or perhaps two drafts, the focus shifts toward "fine tuning" the expression with a view toward a final revision. Editing and responding assume an appropriately different character now, with these guidelines:

Assessment of later stages in composing

1. Comment on the specific clarity and strength of all main ideas and supporting ideas, and on argument and logic.
2. Call attention to minor ("local") grammatical and mechanical (spelling, punctuation) errors, but direct the writer to self-correct.
3. Comment on any further word choices and expressions that may not be awkward but are not as clear or direct as they could be.
4. Point out any problems with cohesive devices within and across paragraphs.
5. If appropriate, comment on documentation, citation of sources, evidence, and other support.
6. Comment on the adequacy and strength of the conclusion.

Through all these stages it is assumed that peers and teacher are both responding to the writer through conferencing in person, electronic communication, or, at the very least, an exchange of papers. The impromptu timed tests and the methods of scoring discussed earlier may appear to be only distantly related to such an individualized process of creating a written text, but are they, in reality? All those developmental stages may be the preparation that learners need both to function in creative real-world writing tasks and to successfully demonstrate their competence on a timed impromptu test. And those holistic scores are after all generalizations of the various components of effective writing. If the hard work of successfully progressing through a semester or two of a challenging course in academic writing ultimately means that writers are ready to function in their real-world contexts, *and* to get a 5 or 6 on the TWE, then all the effort was worthwhile.

REFERENCES

1. Language Testing and Assessment: An Advanced Resource Book, Glenn Fulcher and Fred Davidson, 2007, Routledge
2. Assessing Reading, J.C. Alderson, 2000, Cambridge University Press (lebih baru)
3. Assessing Listening, G.Buck, 2001, Cambridge University Press
4. Assessing Speaking, S. Luoma, 2004, Cambridge University Press
5. Assessing Writing, S.C. Weigle, 2002, Cambridge University Press
6. Testing for language teachers. A. Hughes, 2002. Oxford: Oxford University Press.