Introduction
In this chapter, we discuss the role of diagnostic feedback in language assessments. In the traditional way of examining diagnostic feedback, a narrow scope test specifically designed for diagnosis was considered necessary to provide such feedback. Such tests are rare because the attention of most language test development and research has been on proficiency and achievement testing and, as a result, the concept of diagnostic feedback has not been developed sufficiently. In a forward-looking approach, it is argued that diagnostic feedback can be incorporated into achievement and proficiency testing. Although this has happened somewhat unsystematically in the last decade, language testing researchers have recently shown that it is a feasible approach. The chapter provides examples of this approach and then concludes with the challenges that face researchers interested in diagnostic feedback.

Definitions and Scope
Over forty years ago, Davies (1968) presented traditional definitions of the purposes of tests as follows:

we speak of Proficiency (Aptitude) for or in something to do something else; we speak of Achievement (Attainment) in something by itself; and we speak of Diagnosis of something. Thus in this usage Proficiency (Aptitude) tests the student’s present ability for future learning, Achievement (Attainment) tests his present knowledge as indicative of past learning, and Diagnosis is the teacher’s concern of what has gone wrong. (pp. 6–7)

Davies’ own usage was slightly different: “it distinguishes four uses combining Achievement and Attainment, in terms of time and subject matter symbolized by X (p. 7). He showed the purposes of tests as follows:

The Handbook of Language Teaching  Edited by Michael H. Long and Catherine J. Doughty
In this view, achievement tests are concerned with the past, proficiency tests are concerned with past and the future, and aptitude tests look forward. Diagnostic tests, on the other hand, are concerned with the past in terms of performance and the future in terms of providing information to instructors, students, and parents regarding the strengths and weaknesses of the students’ performance.

In contrast, Bachman (1990) argued that language tests can be classified according to the type of decision to be made. He stated “we can speak of selection, entrance, and readiness tests with regard to admission decisions, placement and diagnostic tests with regard to identifying the appropriate instructional level or specific areas in which instruction is needed, and progress, achievement, attainment, or mastery tests with respect to decisions about how individuals should proceed through the program, or how well they are attaining the program’s objectives” (p. 70). Bachman further argued that

Virtually any language test has some potential for providing diagnostic information. A placement test can be regarded as a broad-band diagnostic test in that it distinguishes relatively weak from relatively strong students so that they can provide learning activities at the appropriate level. Similarly, a readiness test differentiates students who are ready for instruction from those who are not. When we speak of a diagnostic test, however, we are generally referring to a test that has been designed and developed specifically to provide detailed information about the specific content domains that are covered in a given program. Thus, diagnostic tests may be either theory or syllabus-based. (p. 60)

Alderson, Clapham, and Wall (1995) stated their position as follows:

Diagnostic tests seek to identify those areas in which a student needs further help. These tests can be fairly general, and show, for example, whether a student needs particular help with one of the four main language skills; or they can be more specific, seeking perhaps to identify weaknesses in a student’s use of grammar. However, achievement and proficiency tests are themselves frequently used, albeit unsystematically, for diagnostic purposes. (p. 12)

More than a decade has passed since these statements, but they are arguably still accurate, as there are very few specifically created diagnostic tests in second or foreign language testing. However, recently there have been a few attempts to provide some diagnostic information from achievement and proficiency tests.

Large-Scale Assessment Context

A large-scale assessment context is typically aimed at serving summative purposes, for example, evaluating what students have learned and whether they are
Antony John Kunnan and Eunice Eunhee Jang

ready to move to the next level of education. When assessment is used for this particular purpose, assessors concentrate on the linkage between the curricular detail in the target domain being taught and tested items in a test. This approach also conforms to the Standards for educational and psychological testing (AERA, APA, & NCME, 1999) guidelines:

When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both tested and target domain should be described in sufficient detail so their relationship can be evaluated. The analyses should make explicit those aspects of the target domain that the test represents as well as those aspects that it fails to represent. (p. 145)

In practice, however, achievement and proficiency tests have been used for many purposes; and hence, the linkage between test items and curriculum standards is often unclear. For example, in the school context, achievement tests are used to monitor student progress through standardized test administration, scoring, and reporting, to collect uniform baseline information from a large group of students across geographical areas, and to provide rough diagnostic information to all stakeholders (teachers, students, parents, school administrators). Proficiency tests are used in different ways, such as in college and university level entrance examinations to provide a ranking among students (as only the high-scoring are rewarded with admission to colleges and universities). Often, these same tests are used to measure student achievement across geographical regions and to evaluate language programs for government or private accountability purposes. As these purposes are varied, the links between test items and curriculum standards is achieved by focusing on test reliability and validity of test score interpretations, in addition to uniformity of testing practice (including test administration, test time, test forms, test raters, scoring, reporting, and score interpretation) across geographical regions. While this focus on reliability, validity, and uniformity has served large-scale achievement testing reasonably well, there have been criticisms regarding the lack of useful diagnostic feedback to test takers and test score users (including teachers, parents and others) (Kunnan, 2004, 2008).

This approach to traditional testing has also resulted in focusing on quantitative assessments of an individual student’s general language ability relative to other students in the normative group (Brown & Hudson, 2002; Glaser, 1994). Such norm-referenced interpretations of test results have been criticized for their lack of pedagogically meaningful information with which teachers and students can better understand the meaning of test scores and students’ strengths and weaknesses in a specific academic domain, and for the lack of constructive guidance for instructional remediation. Further, while considerable theoretical and practical efforts have been made in advancing the concept of communicative language, authentic materials, and improved scoring into assessments (for example, Bachman & Palmer, 1996; Purpura, 2004), the use of aggregated test scores as
an overall measure of language proficiency or achievement levels, with little or no diagnostic feedback, has made such efforts less profitable to test score users, such as principals, teachers, parents, and students.

There have been a number of exhortations and suggestions to increase the usefulness of test results from researchers. Spolsky (1990) argued that it is the tester’s moral responsibility to ensure interpretability of test information, as well as accuracy. He suggested the creation of “profiles” that show multiple skills tested in more than one way as a more valuable scoring reporting method. Shohamy (1992) proposed a collaborative diagnostic feedback model in which tests provide useful information for teaching and learning advancement. Alderson, Clapham, and Wall (1995) provided extensive discussions about various practical issues concerning how to prepare instructionally useful score reports.

In practice, however, the most common and limited diagnostic feedback, if it can be called that, is still the total score for the test, and in some cases, section or paper subscores (such as scores for different skill areas, like listening, speaking, reading, or writing, and/or language components, such as grammar or vocabulary). But large-scale language proficiency tests used for admission to English-medium universities in Australia, Canada, the UK, and the US have recently begun to provide some limited diagnostic feedback, in addition to total and section or paper scores. A brief description of the diagnostic feedback from three well-known tests is provided.

The Test of English as a Foreign Language (TOEFL)

The current TOEFL provides the following information, according to its website:

Your scores are based on your performance on the questions in the test. You must answer at least one question in each Reading and Listening section, write at least one essay, and complete at least one Speaking task to receive an official score report. For the Internet based test, you will receive four section scores and a total score: Reading (0–30), Listening (0–30), Speaking (0–30), Writing (0–30), and Total Score (0–120). In addition to numeric scores, your examinee score record also includes performance feedback that indicates your performance level and a description of the kinds of tasks that test takers within the reported score range can typically do.

A sample TOEFL Internet-based Test Examinee Score Report (available on the TOEFL website) illustrates how the enhanced performance feedback is provided. In the sample, “typical” performance feedback is provided based on the test taker’s scaled scores, which are 17 points for reading, listening, and writing each, and 14 for speaking. The feedback for the reading and listening skills is more general and speaking and writing skills feedback is more detailed, as these are related to task types. Here is an example of a performance descriptor for reading and listening:

Test takers who score at the low level typically “have difficulty identifying the author’s purpose, except when that purpose is explicitly stated in the text or easy to infer from the text.”
Such reading and listening proficiency descriptors were developed based on a scale-anchoring study (Y. Sawaki, personal communication, April 2007). The scale-anchoring steps included the following: determining several cut points on a scale, mapping items on the scale in terms of difficulty, identifying items that define each ability level, and specifying the abilities and other characteristics of the items that define each level. Proficiency descriptors for speaking and writing skills were developed based on the rubrics and advice from a teacher panel with the help of categories (for example, speaking about familiar topics, about campus situations, and about academic course content) and levels of performance (good, fair, limited, and weak).

**The International English Language Testing System (IELTS)**

IELTS test takers receive scores on a band scale from 1 to 9 based on their performance on each of the four individual modules (listening, reading, speaking, and writing) which are equally weighted. The overall band score is calculated by taking the mean of the total of the four individual module scores. Band descriptors for writing and speaking have been developed to help stakeholders better understand the level of performance required to attain a particular band score in each of the criterion areas. Samples of a writing task descriptor and a speaking descriptor (available on the IELTS website) illustrate the type of feedback that is provided at each of the band levels.

**The Michigan English Language Assessment Battery (MELAB)**

MELAB test takers receive score reports based on their performance on composition, listening, grammar, cloze, vocabulary, reading, and an optional speaking test. Reports include scores of performance on the different parts of the MELAB, and these are averaged to produce a final MELAB score. Samples of composition descriptors according to score level and a speaking descriptor (available on the MELAB website) illustrate the type of feedback provided at each of the score levels.

While these examples of diagnostic feedback provide the necessary link between previous instruction and the test, and help test takers interpret their scores and performance, scale-anchoring with descriptors based on group performance at a score level cannot provide specific individualized feedback. Therefore, while this is a major step toward providing feedback to test takers and other stakeholders, the feedback is so “typical” for a group of test takers in a particular score range that its usefulness to individuals is limited. Another approach is to provide a profile score on a scale based on percentile ranks, as determined by positioning an individual score against the norm. But profiling a students’ proficiency in this
way provides little diagnostic information other than his or her relative position with respect to other test takers on the standardized scale.

**Classroom Assessment Context**

Classroom-based assessment is aimed at providing teachers with information needed to evaluate students’ level of achievement with reference to curricular goals or standards and students with diagnostic feedback regarding their performance. This type of assessment is generally referred to as formative assessment. Teachers can use this information to inform their instruction, evaluate resources, and provide feedback to students to promote their learning; students can use this type of information to understand their strengths and weaknesses. Many formal and informal techniques are used in this type of assessment. A range of test response formats, such as short answer, fill-in-the-blanks, multiple-choice, and alternative performance-based, are widely used to assess student achievement. While traditional multiple-choice response format tests are time-efficient and objective in scoring, they are limited to measuring the knowledge state and comprehension abilities in a discrete manner. Extended production and performance-based assessments are considered to serve our needs to assess students’ achievement in the context of language use better, particularly because of the complex nature of second and foreign language processing and production. Examples of performance-based achievement assessment include both oral and written tasks, such as speeches and essays, portfolios, and performance tasks, like drama and role-play. Other observational techniques that are increasingly used to assess learning outcomes in the classroom include teacher’s direct observation. This technique could provide information not only about achievement in the cognitive domain, but also about non-cognitive outcomes or changes, such as students’ motivation, attitudes, and personal/social development. For example, a teacher’s anecdotal records can provide a systematic source of information on second and foreign language development.

Edelenbos and Kubanek-German (2004) provided an argument for language teachers to be equipped with a new concept, “diagnostic competence.” Diagnostic competence is “the ability to interpret students’ foreign language growth, to skillfully deal with assessment material and to provide students with appropriate help in response to this competence” (2004, p. 260). Based on their research in Germany and the Netherlands, they provided a working definition of a teacher’s diagnostic competence in three parts:

1. diagnostic competence is an attribute of teachers who aim to improve the quality of foreign language growth of their pupils; (2) diagnostic competence can be seen as a combination of pedagogical attitude towards the learner, hermeneutic abilities: seeing, observing, comparing, interpreting, evoking, self-distance, openness; scaffolding learning: as an application of the ‘diagnosis’; (3) diagnostic competence precedes assessment. It is what teachers need in order to assess. (p. 277)
Unfortunately, until traditional teacher training programs train new teachers with these capabilities, diagnostic competence will remain a distant goal.

In a related move, the concept of “dynamic assessment” reconfigures the role of teachers to include assessment as part of their responsibilities (Lantolf & Poehner, 2004). Leung (2007) presents the concept of dynamic assessment (DA) as assessment for learning (AFL), in contrast to static assessment, which could be characterized mainly as assessment of learning:

In conceptual terms, one of the fundamental differences between static assessment and DA is that the former would seek to measure pre-defined abilities with instruments and activities (tests and tasks) that would require unassisted, and in the main, solo performance, while DA is built on the dynamic interaction between the examiner and the examinee in which the examiner responds to the examinee’s difficulties with appropriate support in the form of leading questions, meta-cognitive prompts and other forms of feedback. (p. 260)

Of course, such a move by teachers would need a rethinking of how they would operate in the classroom context. Leung lists four of the ten principles of AFL that are relevant for the reconceptualization of teachers’ assessment role (Assessment Reform Group, 2002):

(1) Assessment for learning should be part of effective planning of teaching and learning. . . .
(2) Assessment for learning should focus on students’ learning. . . .
(3) Assessment for learning should be recognised as central to classroom practice. . . .
(4) Learners should receive constructive guidance about how to improve. . . .

(Leung, 2007, p. 266)

In addition to informal teacher observational techniques, self-assessments are recognized as important tools for assessing student achievement and receiving diagnostic feedback. Self-assessment is aimed at encouraging students to develop critical thinking and meta-awareness of their own language development. In this view, students are acknowledged as the primary critical assessor of their own learning. They are actively engaged in critically assessing their own learning by making sense of information, relating it to their prior knowledge and experience, and using it for planning for new learning. Assessment is not done to or for them, but with and by them (Afflerbach, 2002; Earl & Katz, 2006). Although the merits of self-assessment are obvious, it is not widely used as part of achievement testing in the school context partly because of the difficulties with student underestimation and overestimation of their abilities and partly because school districts have not begun to see these tools as valuable. Therefore, standardized achievement tests designed and developed by testing agencies (most often with no assistance from teachers) continue to be used in the school context.

A new approach that is being used in school contexts is to construct proficiency profiles. This approach determines whether or not a student has performed
at “proficient” or non-proficient levels in his/her tests using a cut-off point based on content experts’ judgments. Although this method allows for input from teachers and content experts, the standards-setting procedure has been heatedly debated because of its uncertainty in determining cut-off points distinguishing students’ mastery levels. Researchers often call this feature of the standard-setting procedure its *Achilles’ heel* (Jang & Ryan, 2003; Kane, 2001). As has been well documented, standards-setting involves interpretations of test results and use (Brennan, 2001; Cizek, 2001) and is a value-laden and judgmental activity that inevitably faces problems, such as subjectivity, human biases, and ambiguity. These concerns have reduced the value of diagnostic feedback and have led educational measurement professionals to pay considerable attention to technical procedures for creating defensible standards.

**New Approaches**

Over the past decade, cognitive skills diagnostic assessment has drawn much attention among educational researchers and practitioners who have faced increasing demands for formative diagnostic feedback through a more fine-grained reporting of examinees’ skill mastery profiles (Alderson, 2005; Buck & Tatsuoka, 1998; DiBello, Stout, & Roussos, 1995; Frederiksen et al., 1990; Hartz, 2002; Shohamy, 1992; Tatsuoka, 1983). As there is a growing body of research into the impact of testing on teachers, learners, educational curriculum, and society (Cheng, Watanabe, & Curtis, 2005; Schwandt & Jang, 2004; Shohamy, 2001), testing consumers have been calling for more descriptive test information that allows meaningful interpretations and fair use of test results for improving instructional design and guiding students’ learning.

While it may be ideal to design a diagnostic assessment instrument to be used specifically for the purpose of diagnosis, many researchers working in large-scale educational assessments originally designed for purposes other than diagnosis are interested in examining whether large-scale assessments can provide useful diagnostic feedback for test takers and test-score users. In addition, the quality of the feedback is important. If diagnostic feedback provided to students is not dependable; its practical usefulness is cast into question. Further, students’ perception and the actions taken to close the gap between test scores and their abilities and their desired learning goals must be meaningful. Therefore, feedback provided to students needs to be sufficiently diagnostic in order to allow learners to reset their own learning goals by breaking down goals into manageable tasks (Black & Wiliam, 1998; Stiggins, 2001). But, in planning to make feedback more diagnostic and dependable, we need to be aware that not all feedback has positive effects on learning and self-esteem (Dweck, 1986). Differential effects of diagnostic feedback can be caused by students’ different ability levels, their learning attitudes, goal orientation, or learning contexts. Usefulness of feedback can also vary depending on focus on either strengths or shortcomings or both and either evaluative or descriptive purpose (Tunstall & Gipps, 1996). Thus,
diagnostic feedback needs to consider learners’ beliefs about learning goals, about their own ability, and cognitive and metacognitive learning styles (Kunnan, 1995). It needs to further consider the link between assessment tasks and learners’ cognitive skills in light of the kinds of inferences made for diagnosing learners’ current knowledge/skill mastery state and actions taken for facilitating the learning progress.

In the field of educational measurement, various cognitive models have been proposed to measure a test taker’s mastery level of a set of skills from an administered test (see DiBello, Roussos, & Stout, 2007; Embretson, 1991; Hartz & Roussos, 2005; Leighton & Gierl, 2007; Tatsuoka, 1983, for details). Several of these models have been applied to second language assessment in reading and listening. For example, the Rule Space Model was applied to a short-answer listening comprehension test administered to Japanese college students (Buck & Tatsuoka, 1998), to TOEFL reading subtests (Kasai, 1997), and was recently used to provide “Score Report Plus” to students who took the Preliminary SAT and National Merit Scholarship Qualifying Test.

Jang’s (2005) diagnosis report card called DiagnOsis is another example of the cognitive diagnostic assessment approach. Diagnostic feedback contained in DiagnOsis consists of four major components: Review your answers; Improve your skills; How to interpret skill mastery; and Skill descriptors (see Appendix). The most interesting component of DiagnOsis is skill mastery standing expressed in a bar graph. It shows an individual student’s strengths and weaknesses in assessed subskills. Figure 32.1 shows Yoshi’s (pseudonym) skill mastery probabilities in nine reading skills which were estimated through application of the Fusion Model (Hartz & Roussos, 2005) to large-scale L2 reading comprehension test data. Skill mastery probability for each skill ranges from 0 to 1, and the gray area in the graph is an indifference region. When a learner’s skill mastery probability falls in the indifference region, mastery standing is undetermined. Otherwise, skill mastery status is determined as either a master or a non-master.

![Figure 32.1 Skill mastery probabilities in nine reading skills](image-url)
of a specific skill. Detailed descriptions of the subskills are presented with test items rank-ordered by the magnitude of diagnostic information in the full report.

Jang (2005) also shows how self-assessment can be combined with psychometric diagnostic information for student skill profiling through a study of five cases selected from 1,372 test takers. The five cases shared similar observed scores ranging from 24 to 26 out of a total of 41 points. However, their estimated skill mastery probabilities varied to great extent. She also considered test takers’ background information, such as gender, first language, reason for taking the test, and self-ratings of reading comprehension skills. Table 32.1 shows a summary of the information collected for skill profiling. The analysis of the five cases clearly shows that learners’ self-assessment is in agreement with statistically estimated skill mastery probabilities, although Case 3 shows disagreement.

Other examples of self-assessment as learning and assessment tools include DIALANG, which is a computer-delivered diagnostic language assessment system (Alderson & Huhta, 2005). Users can take tests of reading, listening, writing, vocabulary, and structures in 14 different European languages. Test results are reported on the six levels of descriptors of communicative activities included in the Common European Framework of Reference (CEFR; Council of Europe, 2001). Alderson and Huhta (2005) claim that DIALANG is the first large-scale language assessment system that aims at diagnosing rather than selecting or certifying language proficiency. Notable is its capacity to provide learners with responsibility for the assessment process through self-assessment. Descriptors of communicative activities from the CEFR are used for self-assessment using “can-do” statements, as well as for diagnostic reporting scales. Learners are provided with the opportunity to review any mismatch between their test results and their self-assessed CEFR levels. DIALANG offers a variety of feedback. It reports users’ performance levels, along with advice and awareness-raising information aimed at helping them to take action for further language learning. For example, feedback in “Check your answers” can be used to review a user’s answers at both the item and subskill levels (see, Alderson & Huhta, 2005, for a more comprehensive description of DIALANG).

Another instance of the use of self-assessment is found in the European Language Portfolio (ELP). A portfolio is a collection of a student’s work, usually selected by the student from a larger corpus and often with a reflective note on the selected work. Self-assessment is an essential component of the ELP, which consists of three components: Language Passport, Language Biography, and Dossier (Little, 2005). The Language Passport provides an overview of the learner’s language proficiency in terms of CEF-referenced skills and any significant language learning experiences. The Language Biography emphasizes the learner’s involvement in planning, self-assessing and monitoring his or her own learning process. Learners are encouraged to state what they can do in one or more than one language. The Dossier allows learners to document all materials related to achievement and experiences appearing in the Language Biography or Language Passport. The overarching goal of the ELP is to make the language learning process more transparent to students and help them develop the ability to self-assess and monitor their own learning.
Table 32.1  Summary of skill profiles (from Jang 2005)

<table>
<thead>
<tr>
<th>Cases</th>
<th>Reason</th>
<th>Self-assessment of reading</th>
<th>Mastered skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Female, Venezuela)</td>
<td>To study abroad (undergraduate)</td>
<td>“Reading is not as good as other skills.”</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“I have some difficulty taking courses taught in English due to problems with reading.”</td>
<td></td>
</tr>
<tr>
<td>2 (Male, Indonesia)</td>
<td>To study abroad (graduate)</td>
<td>“Reading is not as good as other skills.”</td>
<td>1 skill (Summarizing)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“I am very good at understanding graphs and charts in academic text.”</td>
<td></td>
</tr>
<tr>
<td>3 (Female, Columbia)</td>
<td>To study abroad (undergraduate)</td>
<td>“I have no difficulty with reading in English.”</td>
<td>2 skills (Vocabulary, Summarizing)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Reading is my best skill.”</td>
<td></td>
</tr>
<tr>
<td>4 (Female, Vietnam)</td>
<td>To demonstrate English proficiency to company</td>
<td>“Reading is my best skill.”</td>
<td>7 skills (Vocabulary; Comprehension of explicit and implicit textual information; Summarizing; Contrasting ideas, etc.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“My weak areas are understanding charts and graphs in academic text and how to relate different ideas to each other.”</td>
<td></td>
</tr>
<tr>
<td>5 (Female, Lebanon)</td>
<td>To demonstrate English proficiency to company</td>
<td>“Reading is my best skill.”</td>
<td>8 skills (Comprehension of explicit and implicit textual information; Negation; Inferencing; Summarizing; Contrasting ideas, etc.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“I am good at vocabulary and understanding relative importance of ideas.”</td>
<td></td>
</tr>
</tbody>
</table>
Challenges

Skills diagnosis and feedback raise some critical challenges concerning pedagogical, ideological, and technological barriers (Linn, 1986, 1990). From the pedagogical perspective, while the behaviorist view of learning is still widely held in much of current large-scale assessment practice, contemporary perspectives on learning and knowledge acquisition agree that learning takes place through interaction of an individual’s mind with physical, social, and cultural contexts (Greeno, Collins, & Resnick, 1996). Considering different beliefs about how learning takes place and which instructional approach best suits learning, cognitive skills diagnostic assessment needs to be discussed in terms of its theoretical/philosophical assumptions about the learning process and the roles of assessment in broader educational contexts.

From the point of view of language assessment, there are specific substantive challenges. Alderson (2006) argued, based on his experience with the “Can-do” statements used in the Common European Framework of Reference for Languages, (CEFR or CEF) and the DIALANG project, that there “are virtually no significant differences across CEFR levels in terms of difficulty of the diagnostic sub-skills that DIALANG endeavoured to test” (p. 4). He elaborated further

Learners who achieved scores indicating they were at higher CEF levels showed weaknesses in all three sub-skills. It appears not to be the case that as one’s reading ability develops, this is associated with an increasing ability to make inferences, for example, rather than to understand the main idea . . . Similar conclusions were reached with respect to listening. Even low-level learners are able to answer some questions that test inferencing abilities, as well as items testing the ability to understand main ideas. (p. 4)

Alderson (2006) concluded that in order to diagnose language development,

we will need to have a much better understanding of foreign language ability. If we can then incorporate such understandings into assessment and make them useful to learners through the provision of meaningful and useful feedback and follow-up, then diagnosis will truly have become the interface between learning and assessment. (p. 15)

Further, while cognitive diagnostic assessment is aimed at providing formative diagnostic feedback for advancing teaching and learning, we need to be aware of any unintended consequences of the use of skills diagnostic information. Collins (1990) warns us that unintended menacing side-effects might occur in the use of skills diagnostic information in a high-stakes testing situation. There is ample evidence that assessment merely serving as an accountability system makes teachers teach to tests and makes students pay attention to tested subjects or topics.
In addition, most currently available diagnosis models do not have much capacity to allow for various test formats, such as constructed response items. As testing technology drives education, the use of skills diagnostic information in standardized high-stakes testing situations might limit current educational measurement to a narrow behaviorist array of discrete skills. Thus, the use of standardized tests for diagnostic purposes without a strong alignment with the curriculum may create psychometric challenges and concerns about different principles guiding the construction of tests and, also, concerns regarding test washback.

**Conclusion**

The main vision in using diagnostic assessment in large-scale and classroom assessment contexts is to help assess students’ abilities and understanding with feedback not only about what students know, but about how they think and learn in content domains, to help teachers have resources of a variety of research-based classroom assessment tools, to help recognize and support students’ strengths and create more optimal learning environments, and to help students become critical evaluators of their own learning (Pellegrino, Chudowsky, & Glaser, 2001). In order to bring diagnostic assessment to full fruition, it is hoped that the integration of technology into assessment will enable teachers and students to share learning goals, design individualized assessments, and engage in real-life problem-solving tasks. Also, it is hoped that the reconceptualization of educational assessment and measurement may encourage students to show evidence of their understanding in many different ways, not only by performing on traditional tests, but also by writing essays, presenting projects, transforming their new knowledge into alternative expressions such as drama, poetry, or visual arts. Thus, collaboration among the various educational constituencies by informing each other using their own expertise and experience and sharing responsibilities is the key to successful integration of three essential components: curriculum, instruction, and assessment.

The main limitations that need to be overcome so that diagnostic feedback is routinely possible include the following: continued use of the traditional reliability-obsessed, deficiency-oriented approach to diagnosis instead of broadening assessment into performance assessment; better understanding of language ability and language development to be incorporated into a diagnostic feedback model; and the critical need to expand the total and subscore-based reporting with meaningful diagnostic feedback. When these limitations are fully overcome, diagnostic feedback can routinely be offered in all assessments, not just in so-called diagnostic tests, and diagnostic feedback can reach its full potential of integrating assessment with teaching, learning, and the curriculum.
NOTE

1 Scale points and score descriptors for holistic and analytical scales were available for extended discourse tests such as writing and speaking (examples, the Test of Written English and the Test of Spoken English) in the 1980s and 1990s.

REFERENCES


Figure 32.2a  DiagnOsis scoring report card

Review Your Answers

| Question | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
| Your Answer | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Correct Answer | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Difficulty: e e e h m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m

Score: 1356 3.33

How to Interpret Skill Mastery

- Nine primary reading skills are assessed in this reading comprehension test. Please review skill descriptions and example questions attached to this scoring report.
- The graph on the left shows your probable mastery standing of each skill. The grey region indicates that your probable mastery standing cannot be determined.
- There may be some measurement error associated with the classification.
- This diagnostic information can be more useful when used in combination with your teacher’s and your own evaluation of your reading skills.
### DiagnOsis scoring report

<table>
<thead>
<tr>
<th>Skill Description</th>
<th>Example Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Skill 1: Deduce word meaning from context</strong>&lt;br&gt;Deducing the meaning of a word or a phrase by searching and analyzing text and by using contextual clues in the text</td>
<td>33, 34, 32, 4, 3, 11</td>
</tr>
<tr>
<td><strong>Skill 2: Determine word meaning out of context</strong>&lt;br&gt;Determine word meaning out of context with recourse to background knowledge</td>
<td>9, 27, 10, 20, 19, 21, 7</td>
</tr>
<tr>
<td><strong>Skill 3: Comprehend text through syntactic and semantic links</strong>&lt;br&gt;Comprehend relations between parts of text through lexical and grammatical cohesion devices within and across successive sentences without logical problems</td>
<td>3, 26, 12, 36, 4, 2, 22, 33, 24</td>
</tr>
<tr>
<td><strong>Skill 4: Comprehend meaning of text: explicit information</strong>&lt;br&gt;Read quickly across sentences within a paragraph and comprehend literal meaning of explicitly stated information</td>
<td>22, 18, 30, 17, 8, 24, 36, 20, 12, 25, 14</td>
</tr>
<tr>
<td><strong>Skill 5: Comprehend text: implicit information at a global level</strong>&lt;br&gt;Read selectively a paragraph or across paragraphs to recognize salient ideas paraphrased based on implicit information in text</td>
<td>6, 34, 26, 4, 25, 35</td>
</tr>
<tr>
<td><strong>Skill 6: Infer major arguments or a writer's purpose</strong>&lt;br&gt;Sum through paragraphs and make propositional inferences about arguments or a writer's purpose with recourse to implicitly stated information or prior knowledge</td>
<td>31, 16, 23, 15, 28, 2, 11, 7, 32</td>
</tr>
<tr>
<td><strong>Skill 7: Comprehend negatively stated information</strong>&lt;br&gt;Read carefully or expeditiously to locate relevant information in text and to determine which information is true or not true</td>
<td>22, 7, 28, 5</td>
</tr>
<tr>
<td><strong>Skill 8: Summarize major ideas from minor details</strong>&lt;br&gt;Analyze and evaluate relative importance of information in the text by distinguishing major ideas from supporting details</td>
<td>13, 5, 17, 25, 20</td>
</tr>
<tr>
<td><strong>Skill 9: Determine contrasting ideas through diagrammatic display</strong>&lt;br&gt;Recognize major contrasts and arguments in the text whose rhetorical structure contains the relationships such as compare/contrast, cause/effect or alternative arguments and map them into mental framework</td>
<td>37, 23, 35</td>
</tr>
</tbody>
</table>

*Note: All example questions are equally informative in assessing related skills. Questions are listed in the order from most informative to least informative. *’?’* indicates that these skills are weak areas you need to improve. *’?’* indicates that your mastery is not determined.

**Figure 32.2b** DiagnOsis scoring report card (cont.)