Diagnostic Assessment in Theory and Practice

Daniel DUNKLEY

Most students in Japan experience a wide range of language tests in their ten-year long language-learning career. These include tests in different situations, starting with classroom tests which range from short vocabulary tests to long end of term tests. Secondly there are university entrance exams, where the stakes are high. Finally there are optional proficiency tests such as the *eiken* graded tests in seven different grades suited to junior high to university major level students, or the TOEIC test for university students and older candidates.

Students themselves are normally aware of two characteristics of a test, namely the skills tested and the purpose of the test. The former is normally expressed by the students in their question to the teacher "What's it on?" and the latter is tacitly understood: for school tests, getting a good grade in an end of term test will result in a good grade for the course, while university entrance tests are self-explanatory, and optional tests both motivate the candidates to study and increase their educational chances. In some situations a clear pass mark criterion is set, as in the *Eiken* proficiency tests, whereas in others, such as classroom tests, there is no need to think of pass and fail.

From an academic perspective, there are four broad types of test: proficiency, placement, achievement and diagnostic. (Alderson et al. 1987) For his part, Brown focuses on the type of decision rather than the type of test. (Brown

1996) Thus, proficiency decisions are about broad standards of knowledge, related to joining or leaving a certain course: "the general knowledge or skill prerequisite to entry or exit from some type of institution." (Brown 1996, 9) As a result, proficiency tests tend to be broad in every sense, demanding a range of skills (such as reading, writing and listening) and using a variety of question types (multiple choice, essay) and to draw on a wide range of knowledge—grammar and vocabulary. A further purpose of tests is placement. Placement decisions are about dividing students into homogeneous groups for study purposes. Accordingly these tests are fairly similar to proficiency tests, but it is better if they are created with the future course content in mind. The other two types of test are sometimes called *assessment* to emphasize that they they are low-stakes and small-scale, aimed to encourage learning rather exclude low-achievers. Especially at the classroom level, achievement and diagnostic decisions are important. Achievement decisions are about students' success or failure on a specific course: "about the amount of learning the students have done." (Brown 1996, 14) As a result the content will be related very precisely to the course syllabus and the item types will be familiar to the students. Finally diagnostic decisions are needed at certain points in the educational process. As with the analogy to physical health, the purpose of a diagnostic test is not to judge success or failure, but to draw up a list of weaknesses which need to be addressed, and, equally importantly, to establish in which areas the student has been successful. As Brown puts it, "... diagnostic testing often requires ... detailed information about the very specific areas in which students have strengths and weaknesses." (Brown 1996, 15) For this reason the timing of diagnostic tests is usually before a course begins or at an early stage in a course, whereas achievement tests are more likely at the end.

Humanistic testing

Diagnostic testing can be seen as one point on a continuum from behavioristic testing to educational testing. Behavioristic testing is easy to parody as part of social engineering whereby science and statistics are applied to classify students precisely and to assign them to roles in society. It is associated with items which focus on detailed points of grammar or vocabulary (*discrete point* tests) and with multiple-choice "objective" answers. At the other end of the continuum are communicative tests which aim to replicate real-life language use such as writing a report or reading an advertisement. However, these two points of view are not necessarily mutually exclusive: in fact, in real-world large-scale tests both techniques are used to elicit data on the candidates' knowledge. To take two well-known proficiency examinations, for example in IELTS and TOEFL there are single-skill multiple choice questions, in reading for example, but at the same time communicative and human-graded essay-writing questions.

A recent example of the clash of these two points of view can be seen in an exchange between two academics in the pages of a the ELT Journal, a major publication in the field, though not a testing specialist publication. (Tomlinson 2005, Figueras 2005) The soft or humanistic tester was a British lecturer in TESOL, and the hard or scientific tester was a Spanish language testing administrator. The soft tester argued that current tests assess but do not encourage, and that tests should be educational: "… the main point of language testing is to provide opportunities for learning." (Tomlinson 2005, 39) The hard tester objected that his opponent was guilty of imprecise use of terminology, and accused him of being "more emotional than rational" (Figueras 2005, 46), quoting a washback study in this question: "Can a test be held responsible for the ways in which some teachers teach towards it?" (Alderson and Hamp-Lyons

1996, 295) His solution was to propose more training of teachers in the use of assessment.

The concept of *testing to learn* is more precisely defined in the notion of *formative assessment*. Here there are two ideas about assessment in a language programme, summative and formative. Summative assessment is similar to the achievement tests above, in other words, the purpose is to determine, at suitable points in the course, an especially at the end, to what extent the aims of the course have been achieved; it is concerned with "the accountability of the product." (Davies et al. 1999, 65) In contrast formative assessment is carried out to make changes to the teaching methods of a course; it "... attends to the process of a programme in order to provide immediate feedback which could lead to improvement." (Davies et al. 1999, 65) This sounds a little dry, benefitting the teacher rather than the learner. So a more student-centred approach comes closer to the "teaching to learn" idea outlined above. One example of this is the notion of *competency-based assessment* (Brindley 1995) in which diagnostic feedback to learners is a central concern. Brindley emphasizes that assessment should be an integral part of the learning process. Assessment and reporting are closely related to events in the classroom: "... what is taught is directly related to what is assessed and (in theory at least) what is assessed is, in turn, linked to the outcomes that are reported." (Brindley 1998, 52)

This all sounds very desirable, but we must not lose sight of classroom realities. It is undeniable that in spite of receiving regular updates on their progress through frequent assessment, many students fail to make progress This is explicitly stated by Sadler (quoted in Torrance and Pryor 1998, 13–14). So in conclusion we can say that diagnostic or formative assessment alone is not enough. Teachers should "structure classroom assessment to maximize the possibility that intended consequences are realized." (Torrance and Pryor 1998,

169) In other words, we need a system including this kind of assessment for the desired outcome of improved learning to be achieved.

A Large-scale Diagnostic Test: Dialang

So far we have discussed the broad nature of diagnostic testing and contrasted it with "traditional", suggesting that this is just a provisional way of characterizing it: we may find that it may be an unnecessary contrast, a somewhat crude and even misleading distinction. After investigating some examples we will be able to revisit the question.

We will first present a European diagnostic test of language ability. In common with many other parts of the world such as Asia and Africa, Europe has many indigenous languages spoken by populations of widely varying size. For example there are about 90 million native German speakers but only about 10 million native Hungarian speakers. In addition, there are several ethnic and linguistic minorities in many countries, some very ancient such as the Welsh in the UK, others a few hundred years old, such as the German-speaking minority in Hungary, while many are recently arrived such as speakers of North African languages in France, or of Turkish in Germany. As a result of this linguistic diversity, there is a need for language learning for several reasons. Linguistic minorities need to master their host country's language, then a common language is needed to facilitate international cultural and business interchange with neighboring countries, and finally a world language is needed to communicate with people throughout the world, including Asia, Africa and the Americas.

The Dialang test is a very ambitious project, covering many skills, languages and levels. It includes tests of reading, writing, vocabulary and structures. It

is available in 14 different European languages, from the more common, such as French, to those with relatively few native speakers, such as Icelandic. It is based on a system of language levels known as the Common European Framework of Reference (CEFR) (see Council of Europe 2003). This divides language ability into six levels, from A1, the lowest, to C2. One unusual feature of the test is that it is only available on line, and thus has the advantage of being suited to the computer age rather than having migrated from a paper test.

The writers of the test make two claims for the product. First they point out that it aims to describe the test takers' strengths and weaknesses, rather than to state the candidates' success or failure in achieving a specific level. It "aims at diagnosing rather than certifying English proficiency." (Alderson and Huhta 2005, 302) Then, it is created in the spirit of collaborative or autonomous learning. "It gives the taker responsibility for the assessment process". (ibid.)

The examination aims to be easy to use, giving the candidate a wide range of choices. Candidates may choose the language of the interface and rubrics, the language to be tested and the skill to be tested. In addition some aspects of the exam are chosen by the user: For example, users may opt to take a preliminary vocabulary placement test before starting the exam proper. Then they are invited to reflect on their own proficiency in the language, in a self-assessment exercise. Finally there is a wide choice of the type of feedback given.

One feature of the test which is unique to computer-based tests is that it is an adaptive test. The level of the test differs according to the student's answers to the preliminary vocabulary test. Early versions of the test adapted to one of three bands of ability, but adaptivity at the item level is planned.

The part of the test which is most characteristically diagnostic is the feedback. This includes both statements of students' strengths and weaknesses in some detail, and also gives hints for future study. Candidates first see a statement about their CEFR level, and a description of features of the level.

Then the "check your answers" section presents the student's right and wrong answers in tabular form. "Placement test" reports on the preliminary vocabulary test on a 1–100 scale, with a general explanation of the meaning of the score. The self-assessment feedback screen tells the user the extent to which the self assessment and the test results matched. The final section is advice, in which the users see advice on their level and the level above. This is intended to help students to think about their learning.

An Academic test: DELNA

Whereas the DIALANG test is a vast undertaking intended for a wide range of different users, the DELNA (Diagnostic English Language Needs Assessment) test (Read 2008) is for a very specific population. It was developed in Australia and New Zealand to help students beginning their studies at Australian and New Zealand universities. Two categories of students are targeted: one is overseas students, who have never studied in an English medium institution before, and the second is local citizens who for various reasons, normally because they arrived in the country quite recently, do not have a strong command of English. The output from the test is a report on the candidate's strengths and weaknesses in English, along with recommendations for remedial courses if necessary.

The DELNA test consists of two parts, named *Screening* and *Diagnosis*. The purpose of the screening is to identify students who need the diagnosis and those who do not. The second part is the diagnosis proper. The two parts of the test are very different lengths: the screening takes 30 minutes and the diagnosis two hours. The screening has two parts, vocabulary and speed reading, both administered by computer. The vocabulary test consist of word to definition

matching, while in the speed reading items candidates must identify which word in each line of a text is extraneous.

While the screening is computer based, the diagnosis is a paper and pencil test, with three sections: listening, reading and writing. The aim is to check candidates' specifically academic English proficiency, or to "provide a more extensive task-based assessment of their academic language skills." (Read 2008, 183) In the listening test they hear a short lecture about which they have to answer multiple choice items. The reading test is based on an academic passage and has a variety of item types such as cloze and true-false. Finally the writing task involves a 200-word commentary on a social trend, based on a graph. Naturally the written part is manually scored, with double rating to ensure validity.

Candidates receive one of three results for the screening. *Good* means no remedial courses are needed, *satisfactory* indicates that some extra course would be helpful, while *recommended for diagnosis* means that the diagnostic test should be taken. The procedure after the diagnostic test is quite complex. A global score from 9 (best) to 4 (weakest) similar to the IELTS scale is used. Students at the higher levels receive an email report, while those in the lower range are asked to discuss their situation with a DELNA language advisor. The reason that a face-to-face interview is used is that students have been found to be unwilling to take remedial courses unless they are urged to do so by a teacher in person.

The two diagnostic tests described have some features in common. First they both have a two–stage approach. Then the first stage involves vocabulary, because this has been found to be an effective means of judging approximate proficiency levels (see Read 2000, ch 5, Meara 1988). Finally the reporting is detailed.

There are of course many differences, mainly due to the different purposes

of the two tests. The first test is a broad measure of proficiency for a wide range of people and purposes, whereas the latter is an ESP test for a specific population. So with the latter there is less choice and a more goal-oriented form of reporting.

What then are the characteristics of a diagnostic test? It seems that we cannot point to one item type which is specifically diagnostic; what is diagnostic is a pair of characteristics: the breadth of skills and item types on one hand, and the style of reporting to the candidate on the other. The level of detail captured by the items is sometimes known as the *granularity* of the test. A fine-grained test picks up detailed data about many characteristics of the learner's proficiency, whereas coarse-grained tests only give the broad outlines. This has also been linked to group level and individual level feedback. Group level feedback is coarse-grained, in that everyone in the same level receives the same feedback, while individual-level feedback is fine-grained and more helpful to the learner. A particularly thorough approach to individual feedback is found in cognitive diagnosis (Lohman and Ippel 1993, Lee and Sawaki, 2010) in which cognitive psychology is applied statistically to test items.

Conclusion

It is clear that any proficiency test gathers a wide range of data on proficiency as expressed in test performance. However, it is important that the data should be put to good use. First this data needs to be reported to the candidate in a form that will both enlighten and motivate. As teachers we should like to inform students in a way that will lead them to achieve their best possible level. Accordingly we need to supply them with objective and detailed information rather than subjective and vague comments. It is here that diagnostic testing can

exert an influence on classroom procedures and set us on the path to providing the best possible education for our students.

Note

The author acknowledges with gratitude the receipt of an AGU research grant to visit the Language Testing Research Centre, University of Melbourne, Australia, in March 2011.

References

Alderson, J. C. (2005) *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. London: Continuum.

Alderson, J. C. and Huhta, A. (2005) "The development of a suite of computer-based diagnostic tests based on the Common European Framework". *Language Testing*, 22 (3): 301–320.

Alderson, J. C. and Krahnke, Stansfield (1987) *Reviews of English Language Proficency Tests*. Washington, D.C: TESOL.

Brindley, G. (1998) "Outcomes-based assessment and reporting in language learning programs: A review of the issues". *Language Testing*, 15 (1): 45–85.

Brown, J. D. (1996) *Testing in Language Programs*. Prentice Hall.

Council of Europe (2003) *Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Preliminary pilot version. Strasbourg: Council of Europe.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999) *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Lee, L-W. and Sawaki, Y. (2010) "Cognitive Diagnosis and Q-matrices in Language Assessment: The Authors Respond". *Language Assessment Quarterly*, 7 (1): 108–112.

Lohman, D. F. and Ippel, M. J. (1993) "Cognitive diagnosis from statistically based assessment toward theory based assessment". In Frederiksen, N., Mislevy, R. J. and Bejar, I. (eds.) (1993) *Test theory for a new generation of tests*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

Meara, P. (1987) "An alternative to multiple choice vocabulary tests". *Language*

Testing, 4: 142–154.

Read, J. (2000) *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2008) "Identifying academic language needs through diagnostic assessment". *Journal of English for Academic Purposes 2008*, 7: 180–190.

Torrance, H. and Pryor, J. (1998) *Investigating Formative Assessment: teaching, learning and assessment in the classroom*. Buckingham: Open University Press.