THE CHALLENGE OF (DIAGNOSTIC) TESTING: DO WE KNOW WHAT WE ARE MEASURING?

J. Charles Alderson
Lancaster University, UK

Abstract

The language testing literature is confused about the nature of diagnostic tests. Diagnosis is a frequently used but under-problematized concept and a debate is needed that might lead to a research agenda. This chapter aims to begin that debate by sketching out a possible set of dimensions of such a research agenda.

How does foreign language proficiency develop? Test-based diagnosis of language development should be informed by reference to theories of language use and language ability, even though second language acquisition research has failed to deliver a usable theory of development of foreign language proficiency. Research into formative and teacher-based assessment should be explored, both in language education and in education generally, for useful insights. Above all, we need to clarify what we mean by diagnosis of foreign language proficiency and what we need to know in order to be able to develop useful diagnostic procedures.

Introduction

In this chapter, I will argue that diagnostic testing is a much neglected area within the general field of language testing, both in terms of its possible function and in terms of the content and constructs that should underly diagnostic tests. Frameworks of language use and ability such as the Common European Framework of Reference (Council of Europe, 2001), American Council on the Teaching of Foreign Languages (ACTFL) (1985), and the International Second Language Proficiency Ratings (ISLPR) (Wiley and Ingram, 1995, 1999) provide encyclopaedic taxonomies and scales of relevant dimensions of language use and language learning that are likely to be implicated in and affected by language development. However, there is a relative scarcity, despite the scales that are presented in such frameworks, of evidence rather than speculation about their known relevance. Similarly, standards of language achievement, common in outcomes-based assessment (Brindley, 1998, 2001), which supposedly define the levels of attainment expected of (typically school-based) language learners, are often vague, ill-defined, lack any empirical base, and bear little relation to theories of second language acquisition. In short, it is far from clear exactly what changes as learners develop and therefore what
diagnosis of second language development (or lack of it) should be based on, or how diagnostic tests might be validated.

Although I contend that such problems are of global relevance, increasingly authors and researchers emphasize the situated nature of knowledge, and the contextual constraints on how and what we measure. Therefore, I first need to contextualize my thinking, which has developed from my own work in Western and Central Europe. Before going on to discuss language development, diagnosis, theories of language use, and language ability, as well as other aspects of research in language education that might inform how we go about diagnosing foreign language proficiency, I therefore need to say a little about developments in Europe.

Context: Europe and the CEFR

For the past 15 years or so there has been increasing interest in Europe in language education and the transparency of certified language competence, for a variety of reasons. One major manifestation of these concerns and one major contribution to developments has been the Common European Framework of Reference for Languages, sometimes known as the CEFR or the CEF for short. The idea of such a framework is not new, both in Europe, where work began in the 1970s on developing definitions of a level of language proficiency that indicated that learners could operate independently in a foreign language — the Threshold level — and elsewhere in the world, as seen in the ISLPR, ACTFL, the Foreign Service Institute (FSI) and Interagency Language Roundtable (ILR) scales, the Canadian Language Benchmarks, and more.

Elsewhere the use of such frameworks, or outcomes-based statements, benchmarks, or national standards, as they are variously known, is controversial. Brindley in particular (1998, 2001) has shown the problems and dangers in using outcomes-based assessment, in terms of its impact on instruction. However, the CEFR is not part of national governmental policy; rather, it has been developed by an NGO, the Council of Europe, to broaden understandings of what is involved in language education. In point of fact, the Council of Europe has no power to impose anything on any member state, and it is at pains to emphasize that the CEFR is a point of reference, not a means of coercing teachers, nor even a basis for measures of accountability.

Nevertheless, the CEFR has already had enormous impact, and anybody working in or thinking about developments in language testing in the European context has to confront or come to terms with the CEFR.

The aim of the CEFR was to bring together a wide range of thinking and research in language education under a common umbrella, in order to contribute to increased common understanding of what it means to learn, teach, and assess a foreign language, and to give curriculum developers, teacher trainers, textbook writers, language test developers, and classroom teachers a common
framework within which to communicate, to cooperate, and to develop independently.

The most immediate and pervasive application of the CEFR has been in the area of assessment, specifically that of portfolio assessment, proficiency testing, and, latterly, diagnostic testing. Indeed the levels of examinations provided by members of ALTE—the Association of Language Testers in Europe, to which Cambridge ESOL, the Goethe Institut, the Institute Cervantes, and so on belong—are now expressed in terms of the CEFR.

The CEFR provides an encyclopaedic taxonomy of relevant dimensions of language use and language learning that are likely to be implicated in and affected by language development, which attests to the enormous complexity of foreign language acquisition. The CEFR itself is divided essentially into two: the so-called Descriptive Scheme and a series of scales. The former emphasizes the complexity of what it means to learn and use a foreign language. The CEFR views the language learner as a social agent, operating in specific social contexts. The learner’s competence is described within a model of communicative language competence which owes a great deal to the Bachman model (Bachman, 1990), but much is made in the CEFR of tasks, their description, their performance, and the social purposes for which learners engage in tasks.

In addition to the Descriptive Scheme the CEFR also contains a series of scales across the four skills that describe language ability in terms of the six main levels of the CEFR and in a variety of settings (reading for information and argument, listening as a member of a live audience, writing reports and essays, informal discussions with friends, transactions to obtain goods and services, and so on). Such calibrated scales can be seen as providing a snapshot of development, and thus could be used as the basis for test development at the level of interest.

DIALANG

One application of the CEFR with which I have been closely involved is DIALANG, a project which has developed computer-based diagnostic tests in 14 European languages, in reading, listening, writing, grammar, and vocabulary. The test framework and specifications for all the languages were based upon the Common European Framework, as that offered the most recent and most European basis for test development. It was relatively uncontroversial at the time and thus was most likely to be acceptable to the various testing cultures represented in the Project. Unlike other tests and examinations, the DIALANG tests were directly based ab initio on the CEFR rather than being merely linked to it post hoc, and DIALANG test results are reported in terms of the CEFR scales.

The DIALANG suite of tests was intended to be diagnostic and freely available over the Internet. Anybody can take the tests at any time and thus
they are intended to be low-stakes — indeed no-stakes. They are intended to be diagnostic in at least two senses. On the one hand, they report results on each test in terms of the CEFR — from A1 to C2 — without giving any score. Thus they give learners some idea of where they are within the framework of the CEFR. Secondly, they are intended to diagnose ability within each macro skill area, in terms of sub-skills, which are reported in profiles immediately after the learner has taken the test. Thirdly, learners can explore their responses to individual items, to see what they got right and wrong and to speculate on why that might be so. Since the tests also encourage learners to assess their own abilities in terms of the CEFR (using I can statements) feedback is also provided on the match or mismatch between self-assessment and test results, explanations are provided for why there might be a mismatch, and advice is given to learners on how they can improve from one CEFR level to the next.

One of the problems the Project encountered was that, while the CEFR provided material to help define a number of content categories for item writer checklists, the Project had to complement the CEFR itself with material from the more detailed publications of the Council of Europe (the Waystage, Threshold, and Vantage levels), as well as from many other sources when designing the detailed task and test specifications — see Huhta, Luoma, Oscarson, Savaraga, Takala, and Teasdale (2002). However, detailed analysis of the results of the piloting on the English DIALANG tests (Alderson, 2005) reveals that there are virtually no significant differences across CEFR levels in terms of the difficulty of the diagnostic sub-skills that DIALANG endeavoured to test. For example, Alderson (2005) concludes with respect to the reading tests:

> 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 increased ability to make inferences, for example, rather than to understand the main idea. (p. 137)

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. 152)

To summarize, a set of diagnostic tests has been developed, based on the CEFR, which has proved to be very popular across Europe, but whose ability to diagnose development in terms of the CEFR has so far proved problematic.

The Dutch CEFR Construct Project

A better understanding of development might be achieved by a systematic inspection of language proficiency tests that have been developed and calibr-
ed on the CEFR scales, to see differences across levels according to the test specifications and the test tasks. A recent project (Alderson, Figueras, Nold, North, Takala, and Tardieu, 2004, 2006) attempted such an analysis, in two ways. First, by taking all the descriptors common to a CEFR level for a skill (Reading and Listening) and analyzing these descriptors into their component parts, the Project developed a grid that could be used both to characterize texts, tasks, and items from a range of tests and to identify common features at each level of the scale across tests. However, so far, admittedly with a limited number of test tasks (75 in all) the Project has failed to identify such common features. Yet it may be that a larger-scale project, using the Project Grid, might accumulate enough evidence to contribute to the identification of key features at each level on the scale, which could then be examined for their power to predict development in the skills in question, and hence to have diagnostic potential. One promising attempt has recently been made by Kattandjieva and Takala (2006).

The second part of the same Project collected test specifications and guidelines for item writers from a range of examination bodies, all of which claim to test language proficiency at different levels of the CEFR. Once again it proved impossible to identify elements in specifications and guidelines that were common to a particular level, and which were clearly distinguished from other levels. We conclude (Alderson et al., 2006):

There appear to be no systematic differences in the test specifications examined, in terms of most of the dimensions included in the Grid, as CEFR level changes. The specifications examined barely distinguish among CEFR levels in terms of content. (pp. 17-18)

When analyzing the results of expert judgments about texts and tasks using the grid, as well as examining the results of the inspection of test specifications, the Project found very little information on how different dimensions may affect difficulty, or how the dimensions may vary across CEFR levels. It was concluded that item writers' understanding of test specifications seems to rely in most cases on exemplification (previous exams) and local expertise rather than on any explicit construct of language development.

Indeed, I suggest that professional item writers might well have something to contribute to our understanding of how language develops. If item writers have some intuitive sense of what makes an item suitable for a given level of test, i.e., what will work at various proficiency levels, then an exploratory study would be worthwhile to see if experienced item writers can indeed predict item performance and then to develop ways to capture information on how they do it, if they can indeed do it. Anecdotal evidence suggests that experienced item writers can even write test items at an appropriate level for a language they do not know, combining their professional expertise with a native-speaker
informant of the language. How this works, and what light it might shed on (intuitive) theories of language development, could also contribute to a better understanding of diagnosis.

State of the Art?

In summary, a body of evidence is developing that shows that the dimensions contained in the CEFR itself do not describe language development. Given that the CEFR is based upon a view of language learners that sees them as social agents, given that the perspective taken on language ability in the CEFR is sociolinguistic in its emphasis on text types, discourse types, task features, and the like, and given that the CEFR is deliberately intended to be neutral as to target language (there are no scales that describe how specific aspects of individual target languages might develop) then perhaps it is not surprising that the usefulness of the CEFR for diagnosing language strengths and weaknesses is rather limited. In light of the DIALANG and Dutch Construct Project experiences, it is suggested that “the CEFR in its present form may not provide sufficient theoretical and practical guidance to enable test specifications to be drawn up for each level of the CEFR” (Alderson et al., 2006, p. 5).

Indeed, this is perhaps not unexpected in light of what Alderson (1991) says about language scales. He makes a distinction, since widely quoted and accepted, between assessor-oriented, user-oriented, and constructor-oriented scales. On reflection, it seems clear that the CEFR scales are a mixture of user- and assessor-oriented scales, but they are not constructor-oriented. This would account for the lack of guidance in the CEFR for the development of test specifications or guidelines to item writers. The user-orientation of the scales is apparent in the Can-Do statements: they are intended to inform users as to what a learner judged to be at level X (from A1 to C2) on the CEFR, can do. They are intended to help users — as in DIALANG — to interpret their test result. The fact that some of the scales are assessor-oriented (for example, several of the speaking and writing scales) has led to considerable confusion as to the status and orientation of the scales in general.

Language Development

In order to diagnose language development, we need to have a clear idea of how foreign language ability develops. It appears that the scales in the CEFR do not provide a clear picture of such development, and the Descriptive Scheme does not directly address the issue.

I would argue that any study of language development relevant to Europe should attempt to distinguish learners according to their competences as defined within the CEFR framework.
How does foreign language development take place? What do we know about development? We know that it is a long, slow, and complex business; that individuals vary greatly in how, how fast, and how far they develop; that development also varies by first language and probably by aptitude—at least in terms of how fast one learns. We know that development can be characterized in terms of how much one has learned—how many words known, how many structures mastered, how many phonemes produced accurately; we know that development will also vary by context of use: in slow, careful monologues, pronunciation, for example, may be more accurate (for certain phonemes and allophones) than in contexts where quick, spontaneous reactions are required, or on topics one has less knowledge of, or with interlocutors who are of higher status. We also know that development takes place in terms of quality, not just quantity: not simply the number of domains one can use the language in, but also the accuracy, appropriacy, and fluency of one's language.

We know that most learners do not develop native-like competence: the vast majority stop somewhere along the way. This fossilization of interlanguage can happen at any stage of development, and it may happen in some aspects of language use but not in others: pronunciation, for example, may fossilize earlier than lexis, while structural competence may cease to develop long before pragmatic competence.

So we know quite a lot, in general terms, about development, it would seem. We have frameworks, not only in Europe, of course, that purport to describe stages of development, usually in terms of the things that learners can do at a specific level, or tests they can pass. But what distinguishes a learner at one level from another learner at another level and how can a learner go from one level to another? Here there is less certainty: we describe learners at each level and say that learners go from A to B. But how do they do this? What do they have to do to get from A to B? What more do they need to learn or to do to get there (quantity)? How much better do they have to perform to develop from one stage to another (quality)? These are old questions, and it is common to characterize development as an ice cream cone in order to explain how the further up the vertical dimension you go, the more you have to do on the horizontal dimension, which sort of explains how development vertically takes longer the higher up you go—and which may even explain fossilization.

But what exactly do I have to put into my test or assessment procedure to decide whether a learner is at level X or Level Y on this ice cream cone? And, given that a test is necessarily only a tiny sample, in a very limited space of time, of the range of things that I could include in my language test, how can I be sure that I have sampled adequately?

These are all questions relevant to diagnosis: “How do learners go from A2 to B1?” What do they have to learn to make this journey? How can we advise learners on what to do, what to learn or unlearn? And how can we
Diagnosis
We have seen that the CEFR does not of itself provide much insight into how language proficiency develops. Nevertheless it is increasingly used to characterize tests, proficiency levels, and learners. In addition, tests such as DIALANG have been developed not only to identify a learner’s level in terms of the CEFR but also to diagnose learners’ strengths and weaknesses. Indeed, DIALANG is fairly unique, at least in the European context, in attempting diagnosis. Language testers have long written about diagnosis as one of the main six purposes of language tests (proficiency, achievement, progress, diagnosis, placement, aptitude), yet we have presented no ideas how to design or research diagnostic tests. The literature on diagnostic testing is sparse and vague: it is common to assert that diagnostic tests are intended to probe the strengths and weaknesses of learners, but there is virtually no description, much less discussion, of what the underlying constructs might be that should be operationalized in valid diagnostic tests. Indeed, there is considerable confusion in the literature as to the difference between placement tests and diagnostic tests, and it is frequently claimed that diagnostic tests can be used for placement purposes and vice versa. The whole area of diagnostic testing is a much neglected area in language testing.

If diagnostic tests are supposed to identify a learner’s strengths and weaknesses in language knowledge and use, what strengths and weaknesses are relevant? How should we identify these? Are they the same for all language backgrounds, for all learner types, for all possible reasons why learners might be acquiring a foreign language? Is diagnosis of problems in reading or listening different from diagnosis of problems in speaking or writing? Is diagnosis of pragmatic competence possible? Above all, what do we know about language development that could be relevant to diagnosis of language level or progression?

These are all issues that are rarely discussed or researched (for one exception, see Alderson, 2005). Not surprisingly, perhaps, in the light of such confusion and absence of a practical or theoretical rationale, no evidence is ever provided as to the validity of diagnoses, and to my knowledge nobody has described or discussed how the validation of diagnostic tests might proceed. In short, few have problematized—or even thought much about—diagnostic testing.

What we appear to lack, in short, is any theory of what abilities or components of abilities are thought to contribute to language development, or whose
absence or underdevelopment might cause weaknesses. This is in marked contrast with other areas of diagnosis, be that in medicine, psychiatry, motor mechanics, or first language reading development, where not only do such theories exist, but also where there are well-established procedures for diagnosis of weaknesses and problems. I have been particularly struck by the long-established tradition in first language reading of numerous diagnostic test and assessment procedures. What many of these diagnostic assessment procedures have in common is that they tap into particular components of reading ability, for example, visual word discrimination, directional attack on words, recognizing sound-symbol relations, and more. They are also often noteworthy for the fact that they are administered individually rather than in groups and that detailed notes on performance are retained and referred to in interpreting results.

Diagnosis is also frequently related to remediation, where the results are acted upon by teachers, where weaknesses are addressed and efforts are concentrated on removing them. Indeed, in most diagnostic procedures I have examined, the administrator and interpreter concentrate on identifying weaknesses, not on strengths. Strengths seem to be taken for granted. At best they are seen as part of the background to the diagnosis rather than as leading to successful diagnosis and treatment. Whereas in our field we seem to be more interested in identifying strengths — what learners Can Do — rather than weaknesses — what learners Can Not Do. I suggest that in diagnostic language testing we should turn our attention much more to establishing what learners cannot (yet) do.

Many diagnostic procedures in other fields have less to do with the real world and holistic performance and more with being focused on isolating aspects of performance in a clinical setting. I infer from this that diagnosis need not concern itself with authenticity and target situations, but rather needs to concentrate on identifying and isolating components of performance.

A corollary of this is that diagnostic measures need not be integrated or task-based but might be more usefully discrete in nature, since the detailed interpretation that seems to be necessary for diagnosis is much more difficult in integrated tests or performance measures.

The evidence from what research has been done into diagnosis-related language assessment and the development of language use is that the components of test design that are typically included in proficiency tests and in frameworks like the CEFR are not predictive, at least in isolation, of stages of development. The evidence from the Dutch CEFR Construct Project is that dimensions of test design like text source, discourse type, text topic, and even the supposed mental operations tapped in test items do not discriminate tests that target different levels on the CEFR. The evidence from tests like DIALANG is that learners at A2 are clearly able to make inferences and understand the global meaning of texts, as are learners at C1. The difference lies not so much in the presence
or absence, use or non-use, of the sub-skill but in the texts to which they are applied, and in particular to the difficulty of the language in those texts and the density of information. Thus I conclude from this that fruitful diagnosis will not seek to establish whether learners can read editorials in newspapers or understand adverts or recipes but will look at what causes any difficulties learners may have with such text types. I suspect also that whilst density of information, discourse structure, or lack of relevant background knowledge may be part of the problem, we are also likely to find that lack of linguistic knowledge, or lack of the ability to deploy that linguistic knowledge, may well be better diagnostic indicators.

Given our current lack of knowledge about what and how to diagnose, it is conceivable that insights into difficulties and weaknesses might usefully be gained from the learners themselves, through self-assessment and through awareness-raising activities. Asking learners directly what they think they have difficulty with, why they think they have problems in general or in particular, may well yield useful insights. DIALANG is one example of a supposedly diagnostic tool that offers learners the opportunity to assess aspects of their own proficiency, to receive feedback on their test performance, and to reflect on why there might be a discrepancy between their self-assessment and their test results. The system currently provides a set of possible reasons for such discrepancies, but it could conceivably be used in a more open-ended way in face-to-face dialogue between teachers/researchers and learners, both in terms of their understanding of the test results, their self-assessments, and the feedback the system provides, as well as in terms of what aspects of this array of feedback they think likely to be useful in their own language development, or which they have found useful in the recent past. Such dialogues may well provide new insights into which aspects of language knowledge and use can be diagnosed, and what sorts of diagnoses are understandable, relevant, and useful.

This latter point suggests directions for research agendas that might not only enhance our understanding of diagnosis but also might suggest ways in which diagnosis and diagnostic tests can be (partially) validated. I suggest that central to diagnosis must be the provision of usable feedback either to the learners themselves or to the diagnoser—the teacher, the curriculum designer, the textbook writer, and others. Thus the nature of feedback and the extent to which it can directly or indirectly lead to improvements in performance or in eradicating the weaknesses identified, must be central to diagnostic test design. Yet in language assessment in general we are all too often content with the provision of part or whole scores and a rather general description of what the scores might mean—in global terms. Diagnostic testing surely requires much more detailed feedback and explanation, and this represents a major challenge, not only to language testers but to applied linguists more
generally. If the feedback currently provided, for example, by DIALANG or other diagnostic instruments is seen as inadequate, what better feedback can be provided?

Theories of Language Use and Language Ability

Any attempt to identify learners' strengths and weaknesses must, if only implicitly, relate to theories of language learning and use, and thus any approach to diagnosis must take cognisance of current debates in this area. Previous emphases on the importance of linguistic knowledge in language learning and the development of proficiency have to a large extent been superseded by more sophisticated notions of what language ability might be. The influential Bachman model (1990) of communicative language ability (CLA) saw language competence as consisting of organizational and pragmatic competence, with the former dividing into grammatical and textual competences and the latter into illocutionary and sociolinguistic competences. All these and their sub-components are presumably candidates for diagnosis, although nobody has yet explored this in any meaningful or systematic way. In addition, Bachman envisages strategic competence contributing to language use, consisting of the components of assessment, planning, and execution, and it is conceivable that aspects of such competences could contribute to diagnosis, if they could be identified. However, more recently it has been suggested by both McNamara (1995) and Chalhoub-Deville (2003) that this CLA approach is limited by being cognitive-psycholinguistic in nature, whereas language use has to be seen in a social context. Chalhoub-Deville argues that what she calls the L2 construct is socially and culturally mediated, that performances in communicative events are co-constructed by participants in a dynamic fashion, and that language ability is not static, or part of the individual, but rather "inextricably meshed" with users and contexts (2003, p. 376). Quite what the implications are of such theories for the very possibility of diagnosis are far from clear, but research into diagnosis and the concept of strengths and weaknesses cannot avoid taking these views into account, if only eventually to reject them as irrelevant or incapable of operationalization.

A related discussion has to do with the relationship between the constructs we aim to tap in language assessment and the tasks, or the means by which we hope to tap them. I accept the idea that ability and task interact, but I am agnostic as to where the construct is to be found, parea Bachman (2002). Clearly, the difficulty of a task and the measure of ability of an individual is a result of the interaction between task characteristics and individual characteristics. The rub is in identifying and characterizing these.
Bachman's position is that we need to take account of both task and construct in identifying language ability (as indeed do most volumes in the Cambridge Language Assessment Series, edited by Alderson and Bachman), but it is not clear how this would contribute to the identification of features of either task or construct that could have diagnostic potential. It is even less clear how the social interactionalist perspective of Chalhoub-Deville or McNamara would permit the notion of diagnosis or what its elements might be. Nevertheless, given the ongoing nature of the debate, diagnosis research should take account of these issues.

As we have seen from the CEFR, which is essentially task-based, such approaches to assessment have little or nothing to say about development, other than via Can-Do statements. One may be able to read a recipe, phone for a taxi, write a letter of condolence, understand a lecture about nuclear physics, but such Can-Do statements, and associated tasks, do not seem to me on their own to help us understand why one cannot do such things, and thus have little to offer of diagnostic value. Or at least the relevance of such task-based approaches to diagnosis has yet to be demonstrated, and the research reported in Brindley and Slatyer (2002), and Elder, Iwashita, and McNamara (2002) suggests that the adherents of task-based approaches themselves have no explanation to offer for the lack of consistent relations between task feature variables and individual performances (i.e., task features could not predict difficulty of task).

The interesting thing about such negative findings is that they call into question the assumptions of second language acquisition (SLA) researchers like Skehan (1998) that task performance can be predicted by a combination of aspects of code complexity, cognitive complexity, and communicative stress. It would appear that such a (relatively simplistic) approach to performance and hence to diagnosis is not very productive and that what affects performance — and hence what is worth diagnosing — is far from being well understood.

What is abundantly clear from the task-based language assessment research is that complex interactions of variables are to be expected in determining or influencing how individual learners respond to test tasks. In helping us reach such a realization, both the SLA literature and the task-based learning, teaching, and assessment literature have made a valuable, albeit a yet somewhat negative, contribution: we are not yet in a position to say what our tasks are measuring, or how less than perfect performances can help us diagnose weaknesses.

In the European context, where the CEFR is such a powerful framework and where CEFR levels are in many ways important milestones of language development, we lack a suitable diagnostic framework. What we need is a diagnostic framework that can interface with the CEFR, or a future revised and updated version of it, that can help us to explore how learners develop from
one CEFR level to the next and how we can best diagnose problems in such development.

Formative and Teacher-Based Assessment

Since teachers are usually the ones who work most closely with learners, it makes sense to look at how they go about assessing their learners' strengths and weaknesses, and to explore what we can learn from them about diagnosis. Indeed, talking to teachers about how they diagnosed first language readers' strengths and weaknesses, as well as what sort of remedial action they took in light of diagnoses was how Clay (1979) began to develop her Diagnostic Survey.

The difficulties which were demonstrated by the 6-year-old children who were referred to the programme were diverse. No two children had the same problem. Procedures for dealing with these problems were evolved by observing teachers at work, challenging, discussing and consulting in an effort to link teacher and pupil behaviours with theory about the reading process. (p. 67)

Until recently, there has been very little research into teacher-based assessment or any form of formative assessment in foreign language learning. However, Cheng, Rogers, and Hu (2004) show that teachers do indeed claim to use assessment to diagnose their learners' strengths and weaknesses, and thus looking at teachers' assessment practices could be a profitable way forward.

McNamara (2001) says that as language assessment researchers we should broaden the scope of our study to encompass classroom assessment, in order to make our work more answerable to the needs of teachers and learners. I would argue that looking at how teachers diagnose the strengths and weaknesses of their learners would also contribute to a better understanding of what can or could be diagnosed. Rea-Dickins (2001) takes up that challenge and shows how teachers go about formative assessment with English as an additional language (EAL) learners. Although descriptions of the assessment process are interesting, we are given little information on what teachers focus on and describe, and so future research questions that could contribute to our understanding of diagnosis are: What evidence for learning do teachers identify? What strengths and weaknesses do they concentrate on and what evidence has diagnostic potential?

Leung and Mohan (2004) use discourse analysis to understand how teachers do assessment, showing how teachers encourage students to discuss their potential answers to tasks, to justify and debate reasons why they think their answers are appropriate, to arrive at a group answer, and to understand why they believe their answers to be correct; the process is at least as important as the product. Teachers do not simply say right or wrong but treat answers as
provisional and get students to reflect on why they might be correct or incorrect, often through a process of scaffolding students to find the correct answer.

Perhaps most relevant is the study by Edelenbos and Kubanek-German (2004), who develop the notion of a teacher's diagnostic competence, in other words "the ability to interpret foreign language growth in individual children" (p. 260). It is interesting to note that the authors claim that the advent of the CEFR and the related English language portfolio (ELP) "require language teachers to become familiar with new methods of assessment and testing" (p. 260). Specifically, they claim, teachers need to become "more aware of the fact that learners may be at different levels within various sub-domains of language competence" (p. 260). The ELP, they claim, "calls for keen observation and for a comparison between the perceptions — teacher and student — of a student’s achievement. It also challenges the teacher to take a student’s interpretation of individual progress into account" (p. 260).

However, yet again, the findings relate more to the process of teachers assessing learners, and much less to the actual content of the assessment. Future research into exactly what teachers focus on will be important in expanding our understanding of what can be diagnosed, given claims that teachers are in the best position to know their students and to have insights into learning. Insights gained from those closest to the learning — the teacher and the student — as to what changes as learners progress from level to level can only enhance our understanding of what to diagnose and possibly even how and when to diagnose, as well as what to do about the results.

**General Education**

If we go beyond the formative assessment literature into education more generally, there is a vast literature on learning that could be of relevance to diagnosis. I have already mentioned the literature in first language reading, for example, from which foreign language assessment could learn. Within our own field there is of course also a large literature on the learning of foreign languages, although much less of direct relevance to diagnosis. Broadfoot (2005) reminds us that learning is as much emotional as cognitive. She stresses that "if a learner likes and/or respects a teacher, if they are in a supportive group of peers, if the culture of the classroom is conducive to learning and, above all, how they see their own strengths and weaknesses, are factors that are likely to play a key role in the engagement and motivation of the individual concerned" (p. 131).

Thus, according to Broadfoot, it would be mistaken to see diagnosis as narrowly concerned with the technical aspects of language; we must be aware at all times of the emotional, the human, and, indeed, the social dimensions to learning and learning success. Broadfoot suggests that we need to see that "the context in which the learning is taking place, the degree of collaboration between teacher and student and between students themselves, the degree of
confidence possessed by students, the opportunity for effective communication around learning” (p. 132) all contribute to learning success. We need in language learning in particular to remember the variety of learning contexts: the adult learning a foreign language for pleasure or tourism, the university student needing to learn a language to graduate, the school child forced to study a foreign language for which they see no need: all these different contexts are “significantly different affective contexts in terms of motivation, confidence and anxiety levels” (p. 134), that need to be taken into account. Broadfoot suggests that we need to deconstruct the familiar vocabulary of assessment and testing — “ability, performance, standards, achievement” (p. 138) and to pay attention to an alternative lexicon of “context, collaboration, confidence, communication and coercion” (p. 128).

However far this takes us from the notion of diagnosis is a matter of personal philosophy and interest, but when thinking about diagnosis we need to be aware that other disciplines might have something of relevance for us to consider. As Broadfoot says, “We should recognise that all learners are first and foremost sentient human beings and hence that the quality and scope of their learning is likely to be at least as closely related to their feelings and beliefs as it is to their intellectual capacity” (pp. 138–139). Assessment for learning, or even assessment as learning (Earl, 2003) is increasingly recognized as a crucial consideration in the development and use of assessment procedures.

The Need for Research

From what we have seen so far, it is clear that there is a need for research into foreign language development and diagnosis, at the very least within the framework of the CEFR, but conceivably much more generally. Certainly work within the CEFR will have to look for useful research findings elsewhere and then seek to apply them in the European context. In the USA, for instance, the work on measuring growth in second language proficiency mentioned by Butler and Stevens (2001) is of potential interest. It also appears that both California, with its ELD (English Language Development) test, and Illinois, with its IMAGE (the Illinois Measure of Annual Growth in English) could provide insights into what changes as learners improve their English.

We should also be looking at existing studies that have contrasted more or less proficient students to see whether the variables identified could have diagnostic potential within the context of the CEFR (see, for example, Yamashita, 2003, for second language reading, and Wu, 1998, for second language listening). But what should future research into diagnosis concentrate on?

Clearly we should not underestimate the complexity of language learning, and indeed we cannot as long as we pay attention to accounts like those of the CEFR, which constantly remind us of this complexity. At the same time, we should not despair and say that things are so complex that we have no hope of
ever being able to describe language development or to diagnose components of it. There is always a danger of looking for single causes of development, of expecting that one or two variables will provide insight into strengths and weaknesses. It is highly likely, as with the diagnosis of medical or first language reading problems, that complex interactions amongst multiple variables will be of most interest, rather than simple bivariate correlations. Which is not to deny that there may well be useful indices or indicators of development, such as vocabulary size or pronunciation accuracy. These will probably not, however, provide useful diagnoses on their own.

We can learn from the diagnostic literature elsewhere which stresses that we need to understand the mental processes involved in learning a subject or acquiring a skill, something that applied linguistics rather lacks to date, with its current emphasis on social aspects of language use. For progress to be made in diagnostic testing, we will need to have a much better understanding of the psycholinguistics of foreign language learning and use. We need more of the sort of research being undertaken in listening, where considerable attention is being paid to bottom-up processes of speech perception and understanding, of short-term memory use, to attention and retention, where comprehension breakdowns are often shown to be associated with failure to segment sounds into words, and to lexical ignorance or deficiency, rather than exclusively concentrating on context and its supposed (and elusive) facilitating or inhibiting effects.

Clearly there are also non-linguistic factors that can help explain foreign language development, specifically aspects like motivation, both at a macro level in terms of so-called integrative or instrumental levels but also at the micro level of task motivation, and day-to-day, even minute-to-minute, motivation and attention. However, it is far from clear that motivation will be sufficient to explain specific strengths and weaknesses in development.

The SLA literature has suggested a host of other variables of interest in accounting for individual differences in acquisition, both cognitive as well as affective variables, including language aptitude variables, such as phonemic coding ability, language analytic ability, and memory. There are many potential candidates for further research, but it will be important to show that such variables have diagnostic potential in relation to language development. The same applies to matters of recent interest in applied linguistics such as language use strategies, or the difference between declarative and procedural knowledge.

Analysis of learner errors is not currently a fashionable area of research because of the well-known difficulties both of establishing what meaning the learner was trying to convey, and also the difficulty of describing and explaining the errors, but I have no doubt that the combination of error analysis with the creation and exploitation of suitably constructed large electronic corpora of learner language at defined stages of development will prove to be
an immensely valuable tool in the near future. Such tools will need to be both longitudinal as well as cross-sectional in nature — already interesting research in this area is being conducted at Lancaster by Franceschina (2006) and Banerjee and Franceschina (2006), who are looking at the development of writing ability.

Alderson (2005) explores possible aspects of language knowledge — in particular grammatical and lexical knowledge and use — that might be worth exploring. It may be that the learner’s current level of language proficiency influences the diagnostic potential of many linguistic, psycholinguistic, cognitive, or even affective variables. Thus, vocabulary size may have more diagnostic value at lower levels of proficiency, where a minimal vocabulary might be crucial to language use, whereas grammatical abilities may only become diagnostic once a learner has passed a threshold of lexical knowledge. Above all, it is fairly clear that the learner’s first language will have a crucial role in mediating language development, and so diagnosis will have to take account of what is known about the language development and the linguistic strengths and weaknesses of learners from a range of different linguistic backgrounds.

There is no doubt that SLA has a contribution to make in this area, and indeed it has shown that it is fruitful to understand developmental sequences, for example the acquisition of negation, questions, tense and aspect, and more, as well as the development of pragmatic abilities. Clearly a learner’s individual characteristics, developmental history, first language background, language aptitude, intelligence, proficiency in other foreign languages, motivation, and other characteristics need to be taken into account in such research.

I also believe that in Europe, it will be important to situate such work within the CEFR. Obviously the levels put forward in the framework are arte-factual, and development is more a matter of progress along a number of different continua, and indeed backsliding along some of these at the same time as making progress along others. However, frameworks like the CEFR are useful in helping us to conceptualize and operationalize levels of proficiency, and since the CEFR has considerable currency in Europe in language education in general, it makes sense to attempt to relate descriptions of development and exploration of diagnostic utility, to the constructs of the CEFR. Ultimately it may well be the case that we need radically to revise the CEFR and our notions of what language development entails. But we can only do that by understanding the strengths as well as the weaknesses of existing tools, and seeking to expand their usefulness rather than dismissing them out of hand.

Alderson (2005) concludes by drawing attention to the advantages that computer-based assessment offers for the delivery of diagnostic tests, not least because of the possibility of the provision of individualized assessment, immediate feedback, advice, and even follow-up. One can imagine, for example, a learner inputting details of relevant personal characteristics like first
language, age, learning history, and so on, and being presented with a diagnostic assessment tailored to such backgrounds. Bennett (1998) has suggested how computer-based assessment will eventually allow the embedding of assessment within learning, such that assessment procedures are indistinguishable from learning procedures. In such a situation we would indeed have developed tests with learning validity. However,

before that day comes, we will need to have developed a much better understanding of foreign language development. 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. (Alderson, 2005, p. 268)

This seems like a huge undertaking, but given the claims for construct validity that underlie all language tests, it is surely indispensable. Given the need to relate diagnostic testing to learning, it is surely inadequate to consider that such coarse instruments as placement tests are capable of providing useful diagnostic information to learners. Much clearer thinking is needed in our field to define what we need to know in order to be able to provide adequate diagnoses for learners, on which useful feedback and advice can be given to assist with language development. Until we can do that I will be far from convinced that as a profession we have any idea what we are measuring.

Conclusion

In this chapter, I have argued that diagnostic language testing is a much neglected area. Although frequently named as one of the main purposes for language tests, there are very few examples of diagnostic tests, considerable confusion about the distinction between diagnostic and placement testing, and no theorizing about what might usefully be diagnosed. As a result we lack guidelines on how to construct diagnostic tests, what they might contain, or how to validate their use. In the absence of these, I speculated about what an adequate diagnostic test might look like and what it needs to achieve. A diagnostic test needs to be based upon a model of foreign language development, which we currently lack, which itself needs to be based on a theory of language ability and language use, about which there is still controversy. Faute de mieux, those constructing diagnostic tests in Europe have had recourse to a framework for language learning, teaching, and assessment that is widely referred to but which itself has limitations as the basis for diagnosis. We therefore need to complement our understanding of development by looking at other areas of assessment—in particular teacher-based formative assessment—for insight into what can be diagnosed, as well as to general education for current ideas on what facilitates and enhances learning. Taking account of these areas of
education will doubtless enhance our understanding of what we might diagnose, but at the end of the day we need to account for the learning of language, and therefore we must work together with applied linguists, in particular second language acquisition researchers, in a joint search for variables that will help us describe, diagnose, and possibly explain foreign language development.

Notes
1 The name DIALANG was a sort of blend of diagnostic and language testing; it is the name of a project, a suite of tests, and is the subject of a website: www.dialang.org.
2 See Council of Europe (2001) for information regarding this six-point scale, ranging from A1 (lowest) to C2 (highest) as follows: A1, A2, B1, B2, C1, and C2.