Language testing-informed SLA? SLA-informed language testing?

J. Charles Alderson
Department of Linguistics and English Language, Lancaster University

The aim of SLATE is to bring together language testing researchers and researchers of second language acquisition to create synergies to our mutual benefit. This volume bears testimony to the interesting and varied research that SLATE has inspired, and there is much to be learned from this, both by SLA researchers and by language testers. In this invited evaluative chapter, I will attempt to draw some of the lessons and to answer the questions of my title.

The first, and to me impressive, lesson is that second language acquisition is about much more than English. The various chapters here report on studies into the development of aspects of proficiency not just in English, but also in Finnish, French, Dutch, Italian, Norwegian and Spanish. That clearly reflects the (West) European nature of SLATE and is to be applauded. However, there is a notable lack of studies of other European languages, particularly the Central and East European Slavic and Baltic languages as well as Greek, Hungarian, Romanian and others. Even German, Portuguese and Swedish are missing. Let us hope that as SLATE's work extends and becomes better known, this situation will change.

The second lesson is that, so far, SLATE has not paid much attention to the relationship between the learners' first language and their target second or foreign language, and it is to be hoped that, in future, attention can be paid to this important matter. Currently, most informants in SLA studies seem to be from so many different L1s that it is impossible to draw conclusions about cross-linguistic transfer or influence. Future studies will need to be specifically designed to yield such information, rather than to hope that it might emerge from opportunistic samples.

In relation to this issue of L1, it is interesting to note that one or two of the studies reported looked at the performance and proficiency of native speakers of their target language, but it was not always clear whether like was being compared with like, i.e. informants of similar ages, educational background and so

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 239-248

on. The very notion of a native speaker has, of course, been questioned and problematised (see Davies, 2003, for example), but there are obvious benefits to seeing how (similar) native speakers perform on the measures used by SLA researchers, even though we are conscious of the comparative fallacy (Bley-Vroman, 1983).

The third lesson is that second language acquisition and language testing need to conduct research with a range of different informants, not just the ubiquitous university student. Obviously such captive populations are convenient and attractive for reasons of practicality, but it is important that younger learners, learners outside formal education, migrants in second language settings, and learners who are not simply taking a test, be studied. It is to the credit of authors of this volume that SLATE members have already begun the study of younger learners and those who are not simply conveniently available because they have taken a public examination.

The fourth point, albeit not perhaps a lesson, is the importance of the Common European Framework of Reference, CEFR, (Council of Europe, 2001) in virtually all of the chapters. This is not surprising, given both SLATE's explicit aims to examine the relationship between the communicative approach of the CEFR and the linguistic development of learners, and the growing importance of the CEFR in Europe and beyond. Not everybody agrees that this is a desirable state of affairs (Fulcher, 2004; McNamara, personal communication), and there are concerns that the CEFR as it currently stands is probably not suitable for, and not intended for, younger learners. Nor is it suitable on its own for the development of language tests, or even of textbooks and curricula (Alderson et al., 2006; Byrnes, 2007b; Hulstijn, 2007; Little, 2007; Weir, 2005; Westhoff; 2007, and others in the MLJ Perspectives edited by Byrnes, 2007a). However, there can be no doubt that the existence of the CEFR has given an important impetus to language education, to language testing and examining in particular, and to research into the development of language competence. That the CEFR needs to be adapted to particular contexts not only should go without saying, but is explicitly stated in the CEFR itself, particularly in the boxed texts that frequently begin "Readers might like to consider to what extent...". It is regrettable, but it was predictable, that claims are made about the CEFR level of curricula, textbooks, tests, and more, that have no empirical basis, but which are often produced for marketing or political purposes. But that should not detract from the importance of the CEFR; indeed it emphasises the importance of research that investigates and challenges the claims of both the language education profession (or industry) and of the CEFR itself. Such research, as is beginning to be attested in SLATE, can only enhance our understanding of language proficiency and its development.

It is, however, important that such research be properly conducted, based upon knowledge of best practice in, and theories of, second language acquisition and language testing. Too much research has based itself on unsatisfactory measures of "proficiency" like years of study in school, first, second and subsequent years of study at university, the Vocabulary Size Placement Test of DIALANG, or a cloze or C-test. However, several authors in this volume have been careful to avoid the circularity of using CEFR linguistic scales alone.

> When rating with reference to the CEFR levels, our aim was to rate learners' performances on the basis of their ability to do things with the language. Paying too much attention to linguistic features could introduce circularity in the reasoning underlying a study such as Cefling: proficiency levels are determined on the basis of linguistic features, and these features are, in their turn, used in defining the levels. (Alanen, Huhta, & Tarnanen, this volume)

Crucially, the fifth and substantive lesson is the importance of paying attention to the construct to be investigated, and of widening the range of constructs. Inevitably, for reasons of practicality, there is a tendency to investigate development through studying learners' written productions. There is no need to transcribe speech, and, increasingly, data can be available in digital form if the informants have word-processed their writing (as in computer-based testing, for example). There is less emphasis in the research reported in this volume on examining learners' oral performances, and no research looking at how learners' reading and listening abilities develop. This is hardly surprising, given the difficulty of studying what are essentially internal processes, but it is nevertheless to be hoped that future research will pay more attention to these so-called receptive skills.

SLA research has in the past tended to pay much more attention to morphosyntax than to other aspects of language, so the chapters on the development of vocabulary (Milton, this volume) and cohesive devices like discourse connectives (Carlsen, this volume) are especially welcome. It is to be hoped that other areas of language use might be studied in the future, like the development of the pragmatic features of politeness, for example, and sociolinguistic and cross-cultural competences. Although some studies still use convenient but crude indices like errors per T unit, or the number of subordinate clauses per clause, it was refreshing to see much more attention than in the past being paid to variables of more convincing construct validity.

But perhaps the most important lesson of all for me, as a language testing researcher, was the importance of research into SLA paying much more attention to its methodology, and the validity and reliability of the instruments and proce-

dures used. In this regard the chapter by Alanen et al. (this volume) was exemplary in its account of the design of their study. More and more SLA studies use electronic corpora as the data for their investigations, either specially created by the researchers, or pre-existing learner corpora, such as the International Corpus of Learner English (Granger, 1998), The Longman Learner Corpus (2008) or the Cambridge Learner Corpus (2010). The Cefling project, reported on in two chapters in this volume, is an interesting case of both. The researchers made use of a corpus of examination scripts from the National Certificates of Finland, but also created their own corpus of young learners' writing on specially designed tasks.

Pre-existing corpora, although convenient, may have their drawbacks. The International Corpus of Learner English was highly innovative in its time and gave rise to numerous interesting studies, but it has two rather serious limitations. First, there is no indication of the learners' proficiency level, be that according to the CEFR or any other measure. Rather, they are classified according to their year of university study. Unfortunately, that is not a valid measure of their level of development. Secondly, the tasks on which the data were based were "persuasive or argumentative essays", with no standardised rubrics, on a wide range of topics, and so the comparability of essays from different sources must be in some doubt. In addition, different genres are ignored, as are learners at different ages. Even the Cambridge Learner Corpus (CLC), the subject of one chapter in this volume, has problems, as shown by Kim (2009). Data in the CLC were collected over a long period of time, based upon the Cambridge Main Suite of ESOL Examinations. However, not only do the tasks in these examinations change from one administration period to the next, but all the examinations have undergone more or less substantial changes over time. This makes it difficult to be sure that the tasks are parallel. In addition, no grade is given for the writing tasks, but only for the overall grade achieved on the whole examination. Moreover, the examinations have not been formally linked through a standard-setting process to the CEFR, and therefore it is difficult to make statements about progression through the CEFR levels, rather than through the various Cambridge examinations. Similar problems exist with the Longman Learner Corpus.

Researchers who create corpora specially for SLA and testing research, such as the Young Learner part of Cefling, need to pay careful attention to their design, as Alanen et al. (this volume) describe. Learners should be asked to perform a variety of tasks, which have been specially designed for their capacity to elicit relevant performances, preferably related to the target language use situation. These need to be thoroughly scrutinised, piloted on suitable numbers of target learners, analysed and revised in light of the results. They should then be administered in conditions which are appropriate to the task, not necessarily, however, under examination conditions. The written or oral performances should be rated on

relevant scales, avoiding the danger of circularity, and the raters should be familiar with the scales, trained in their use, and benchmark performances should be available for the guidance of the raters. Such scripts should always be rated by at least two raters, and only those scripts should be incorporated into the corpus on which raters agree, within clearly specified margins of disagreement. Ideally, there would also be available an independent measure of CEFR-related proficiency.

Most corpora, and indeed SLA studies, are cross-sectional, but there is a strong case for longitudinal studies. Such studies need to allow sufficient time for relevant development to occur, and thus are probably better undertaken with learners in the range A1 to B1, initially at least, than with more advanced learners. Inevitably there are difficulties gathering data from a sufficient number of informants for the results to be of more general interest, but the existence of a group like SLATE may facilitate larger-scale studies than are possible in doctoral research. Much SLA research, even cross-sectional research, has used rather small datasets, which limits the value of such studies.

Hulstijn, Alderson, and Schoonen (this volume) contextualise and present the initial research questions of interest to SLATE. An important question is: to what extent are these research questions addressed in this volume?

The over-arching research question was:

Which linguistic features of learner performance (for a given target language) are typical at each of the six CEFR levels?

Clearly this has been or is being addressed in virtually all the chapters presented in this volume. However, this research is in its early stages. No definitive answers have yet been provided, and much more work remains to be done.

The more specific research questions and goals of research were:

1. What are the linguistic profiles at every CEFR level for the two productive language skills (speaking and writing) and what are the linguistic features typical of the two receptive skills (listening and reading) at every CEFR level?

The chapters in this volume begin to address the productive skill of writing, particularly at the A2/B1 divide. However, speaking has yet to be explored in detail across the CEFR levels, and the two receptive skills have as yet received no attention, as discussed above.

2. *To what extent do common or different profile features exist across the seven target languages investigated by researchers in the SLATE group (Dutch, English,*

Finnish, French, German, Italian and Swedish)? Do the profiles differ along language-family lines (Finnish versus the two Romance languages represented in SLATE (French and Italian), and the three Germanic languages Dutch, English and German,)? To what extent do the profiles reflect learners' L1?

Although research relevant to these questions has begun, the surface has barely been scratched. Cross-linguistic comparisons are only addressed in one chapter (Kuiken, Vedder, & Gilabert, this volume) but are not related to CEFR levels, language families have not yet been profiled or compared, and the study of learners' L1s with respect to the second/foreign language profiles has yet to begin.

3. *What are the limits of learners' performance on tasks at each of the CEFR levels?*

Research into what learners canNOT do is barely represented in this volume, but remains an important area for distinguishing performances on different tasks at different CEFR levels.

4. *Which linguistic features, emerging from our profiling research, can serve as successful tools in the diagnosis of learners' proficiency levels and of weaknesses that require additional attention and training?*

The diagnosis of learners' strengths and weaknesses at the different CEFR levels has to be an important aim of SLATE research, and several chapters mention the diagnostic potential of their research. However, it has to be admitted that not much progress has been made in this area, especially with respect to the CEFR levels rather than to individual examinations or examination suites. The area of diagnostic testing is, however, very under-developed (but see Alderson, 2005, 2007; Alderson & Huhta, 2005). It is the firm belief of this author that SLATE has much to contribute in the future, if its efforts are appropriately focussed.

5. *Are there commonalities and differences between the linguistic profiles of foreign-language learners (learning the target language in the formal setting of a school curriculum or a language course) and those of second-language learners (learning the target language without formal instruction)?*

Given that linguistic profiles have yet to be achieved through empirical research for any language, this is clearly an ambitious, albeit interesting, question that is not addressed in this volume. Which is not to say that it will not be possible to address it in the future, as part of a long-term agenda.

And so to return to the questions of my title: *Language testing-informed SLA? SLA-informed language testing?* The assumption behind the questions is that both are important and equally relevant aims for SLATE, as is implicit in the acronym. However, in order to start a constructive and fruitful dialogue between the two disciplines, it is important to recognize some fundamental differences between SLA and testing.

As many chapters in this volume testify, SLA research is mainly concerned with variability of linguistic performance - be it variability related to the tasks, to the L1 or to a host of individual factors. While this variability is certainly a concern for language testers as well, it is clearly the case that testing, by its very nature, or, at least, proficiency testing, is mainly concerned to measure what is stable in language ability. There is little point in measuring something, especially if the results of that measurement will affect lives, if it is highly likely that that ability will change in the next ten minutes or three months, or if the very fact of measurement will distort the results. We need to have as stable as possible a picture of what we are measuring. Testers have to look for stability - one aspect of which is reliability - and we need 'stability of judgement' as much as we need 'stability of performance' (or ability, if you must). Testers also seek generalisability (whether you see this as a facet of reliability or validity is immaterial to the present argument). Thus, we are not interested in knowing whether the learner can understand this particular story about Prince Charles in today's edition of the Daily Mirror as compared with a rather different story about the same person in the Daily Telegraph, as compared with their ability to understand an article about the Dutch Royal Family. We typically want to make more general statements about reading ability than that. Thus testers - proficiency testers at least - are much less concerned with the particular than with the general underlying ability (I would argue if I had the space that diagnostic testers might well be more interested in the particular than the general, but that is another paper – see Alderson, 2005).

SLA researchers, on the other hand, seem to be more interested in examining the particular - the use in English of the third person s, the use of the present perfect, of negation, or particular speech acts, and so on. This is in part no doubt because they are driven by linguistic theories and the predictions of such theories - hence the proliferation of studies on the development of grammatical morphemes - or by pedagogical applications of linguistic insights - hence the search for the effect of negative evidence. SLA research is more interested, I would argue, in some of the bits and pieces, the minutiae of linguistic competence or performance than they are in the much wider picture of the ability to communicate meaning in context. Language proficiency testers these days (not diagnostic testers, who scarcely exist as a breed yet) are almost obliged to exam-

ine how somebody would perform in a range of communicative settings, across a variety of tasks, texts, activities and all the other dimensions of communicative behaviour that are discussed, however briefly, in the CEFR. SLA researchers are not, *per se*, interested in what is in the CEFR, because there is no specific mention of specific languages or specific linguistic knowledge.

SLA researchers take 'the worm's eye view', immersed in the detail of the blades of grass around them, whilst language testers take not merely the bird's eye view (which might include trying to find worms to eat), but arguably the space satellite's view in order to discover the underlying features that determine the shape of the landscape. The different eye views might be incompatible - that remains to be seen - but they are certainly different, and I would argue that they serve different purposes. However, although much neglected and confusingly defined (see Alderson, 2005), diagnostic testing is one obvious possible meeting place between SLA and language testing. Diagnosis needs theories of language development, as provided, at least potentially, by SLA and SLA needs the insights provided by language testing's awareness of the variability in performance across test task facets.

How far has this volume advanced the informing of SLA through language testing principles and practice, and to what extent has testing been informed by SLA theory and results?

My impression is that the main focus of the volume has been on SLA questions and issues, and has not (yet) been relevant to language testing. Which is not to say that this will not be the case in future. Indeed, I would assert that, to the extent that SLA research heeds the principles and practice of language testing, it will eventually yield results that will be hugely relevant to language testing. But these results are much more likely to be relevant to the diagnostic testing of learners' strengths and weaknesses than they are to the communicative testing of language proficiency in target language use situations.

Language testing, at least in high-stakes contexts, has tended to concentrate on the testing of an individual's ability to communicate accurately and appropriately in relevant settings, because it has been concerned with assessing learners' proficiency, and not with diagnosing their levels, strengths and weaknesses. Language testing may thus at first sight not appear to offer much to SLA's interest in the development, specifically, of linguistic competences. However, I believe this would be a mistaken conclusion. SLA has much to learn about the development of valid, reliable and comparable language use tasks which can offer insights into the linguistic features of learners' performance and development. Indeed, I believe that this volume is evidence of the importance of SLA researchers paying attention to language testing's concerns. Thus, it would

appear to be of more relevance to SLA researchers than to language testers. Whilst some progress could be said to have been made towards language-testing informed SLA, the relevance to SLA-informed language testing has yet to emerge.

Lest this appear too pessimistic a conclusion, let me finish by claiming that, to the extent that SLA is informed by language testing principles and practice, SLA will indeed contribute important research insights that will eventually be relevant to and will inform language testing theory and practice, especially, but not exclusively, in the area of diagnostic testing. In addition, I would hope that, given a common interest in the CEFR, both as a statement about what it means to be able to learn and use any foreign language, and as a possible framework for understanding - or thinking about - what language development might mean, we might be able to find common ground. I look forward to that future.

References

Alderson, J. C. (2005). Diagnosing *foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C. (2007). The challenge of diagnostic testing: Do we know what we are measuring? In J. Fox et al. (Eds.), *Language testing reconsidered* (pp. 21–39). Ottawa: University of Ottawa Press.

Alderson, J. C., & 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., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of The Dutch CEFR Construct project. *Language Assessment Quarterly*, *3*(1), 3–30.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, *33*(1), 1–17.

Byrnes, H. (Ed.). (2007a). Perspectives [Special section]. *The Modern Language Journal*, *91*(4), 641–685.

Byrnes, H. (2007b). Developing national language policies: Reflections on the CEFR. *The Modern Language Journal*, *91*(4), 679–685.

Cambridge Learner Corpus [A collection of exam scripts]. (2010). Retrieved March 25, 2010, from http://www.cambridge.org/fi/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/?site_locale=fi_FI

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Davies, A. (2003). *The native speaker: Myth or reality?* Clevedon: Multilingal Matters.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, *1*, 253–266.

Granger, S. (Ed.). (1998). *Learner English on computer*. London and New York: Addison Wesley Longman.

Hulstijn, J. J. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language. *The Modern Language Journal*, *91*(4), 663–667.

Kim, J. (2009). *Development patterns of Korean learners corresponding to morphosyntactic items* (Unpublished Doctoral dissertation). UK: Lancaster University.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, *91*(4), 645–655.

The Longman Learner's Corpus. (2008). Retrieved March 25, 2010, from http://www.longmanusahome.com/dictionaries/corpus.php#aa

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300.

Westhoff, G. (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. *The Modern Language Journal*, *91*(4), 676–679.