4 Assessing Academic English Language Proficiency: 40+ years of U.K. Language Tests

Alan Davies
University of Edinburgh

Abstract

The paper offers an explanatory account of the progress of academic language proficiency testing in the U.K. (and later Australia) from the British Council's English Proficiency Test Battery (EPTB) through the revolutionary English Language Testing Service (ELTS) to the present compromise of the International English Language Testing System (IELTS). The three stages of academic language testing in the U.K. over the last 50 years move from grammar through real life to features of language use. At the same time, comparison of predictive validities suggests that all three measures account for very similar shares of the variance (about 10%) and that therefore the choice of an academic language proficiency test is determined only in part by predictive validity: other factors, such as test delivery, test renewal in response to fashion, research and impact on stakeholders, and assessment of all four language skills, are also important. Implications are drawn for our understanding of academic language proficiency.

Introduction

In this paper we trace the development of academic English language proficiency testing in the U.K. since the 1950s, paying particular attention to three tests, the English Proficiency Test Battery (EPTB, 1964), the English Language Testing Service test (ELTS, 1980) and the International English Language Testing System (IELTS, 1989). It is suggested that these tests embody changing views (or paradigms) of language. We explain these changes as showing, first, a strong influence of the communicative competence construct in the move from EPTB to ELTS and, as doubts about the meaning and use of communicative competence grew, a fall back towards a compromise position (which we see in IELTS). A convincing argument for this reversal was the similarity of variance across each of the tests and a common predictor of academic success, such as end-of-year degree or diploma examination results. An equally convincing argument was the growing acceptance that test delivery requirements should be included within the scope of a wider understanding of validity (Messick, 1989). Within the tradition described in this paper, tests of academic language proficiency are seen as primarily assessing skilled literacy, the literacy of the educated, based on the construct of there being a general language factor

relevant to all those entering higher education whatever specialist subjects they will be studying.

Over the years, much of my work in language testing has concerned language proficiency, especially the proficiency of foreign/international students entering higher education in English-speaking countries. In the early 1960s, when I was a post-graduate student in the University of Birmingham, I was offered an appointment on a project set up to investigate English proficiency on behalf of the British Council. The project intrigued me and without much regret I abandoned the research I was conducting into Anglophone negritude and spent the following two years developing an English proficiency test, which was given the name English Proficiency Test Battery (EPTB). In due course this test was put into operation by the British Council, at first targeting their own scholars and Fellows but over time used more widely by British universities and other post-secondary institutions. The advantage for these institutions was that the test would be conducted by the British Council in a student's home country and the result used as part of the selection and admissions procedure. Furthermore, the students themselves bore the costs. British universities were well served by the procedure, which continued in use until 1980.

Academic Language

Before describing the U.K. experience of testing academic language proficiency, it will be helpful to consider views of academic language. While academic language is taken for granted as a construct, attempts to describe it as a single domain raise even greater doubts than those which query the unitary nature of academia. Do science, music, the humanities, engineering, and dentistry all share some idea of knowledge and investigation or do we just assume they do because they are all studied and researched in universities? And for us, the harder question: do they all have a language in common which is different from other language uses?

Logic (Ravelli and Ellis, 2004), literacies (Zamel and Spack, 1998), language functions (Chamot and O'Malley, 1994), range (Short, 1994), intertextuality (i.e., Gibbons, 1998), specialized vocabulary common across academic disciplines (Cunningham and Moore, 1993)—these have all been considered in the search for an explanation of the nature of academic language. Bailey and Butler (2004, p. 186) conclude that "academic language ... implies ability ... to express knowledge by using recognisable verbal and written academic formats." "Moreover," they say, "academic language use is often decontextualised whereby students do not receive aid from the immediate environment to construct meaning" (p. 186). They suggest that the "development of test specifications that focus on both oral and written academic language will serve the long-term goal of developing a test framework that is based on empirical data culminating in academic language proficiency prototypes" (p. 189). Van Lier

(2004, p. 161) agrees: "In terms of academic development, learners need to be able to talk about the concepts required with their teacher and peers, to participate in conversations about the issues before they can be expected to apply the concepts and the modes of reasoning in literate products." And he warns that "narrow test-based accountability cultures cut off (for lack of time, since test preparation is of the essence) the very means by which academic success is established. ... Of course, in the short term, students may achieve good test scores, but in the long run, they will end up unprepared for the challenges that they will face in their professional life" (p. 161).

Academic corpora have been analyzed to show a common academic vocabulary. Coxhead (1998, p. 159), researching a corpus of academic texts containing 3,500,000 running words, extracted "a compilation of 570 word families which occurred with wide range and high frequency." However, the use of the same lexeme in different academic contexts does not necessarily mean that they always have the same meaning: "vocabulary which is characteristic of a particular context of use cannot be identified just by looking for unusual and distinctive terms, because words from a general or a sub-technical list may have technical meanings that justify including them in a specific list as well" (Flowerdew, 1993, p. 236).

There is some consensus in the notion of an integrated set of language skills required to socialize students into the acqusition of academic language: "writing ... is not ... a stand-alone skill but part of the whole process of text response and creation; when students use both reading and writing in crucial ways, they can become a part of the academic conversation — they signal their reponse to academic ideas and invite others to respond to their ideas in turn" (Hamp-Lyons and Kroll, 1997, p. 19).

Testing Academic Language Proficiency: The U.K. Experience

The construction and development of the English Proficiency Test Battery (EPTB), referred to above, is recalled in some detail in a volume in the Cambridge series *Studies in Language Testing* (Davies, 2007), in which I look back at the developments in academic English language testing in the U.K. (and more recently in Australia), developments that were not paralleled by the similar activity in the USA, no doubt because its strong tradition of psychometric reliability put a premium on test improvement rather than test change. While the U.K. revised and rewrote its test materials, on the basis both of principle and fashion over the 40-year period, in the USA, the Test of English as a Foreign Language (TOEFL) has (until very recently) remained as steady and unchanging as the northern star. I do not here examine other situations in detail. The North American experience has been discussed by Spolsky (1995; see also Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, and Schedl, 2000; Davidson

and Cho, 2001; and Snow, 2005). Brindley and Ross (2001) and Hyland (2004) examine other situations. What I try to do in this chapter is to explain why the British proficiency tests seemed to change so radically. Very recently TOEFL itself has changed dramatically. As well as employing a Web-based delivery, it has focused on English for Academic Purposes. It may be thought that this change has been influenced by the IELTS example, but the general shift in the climate of opinion regarding proficiency in higher education has also made itself felt (Douglas, 2000). TOEFL iBT describes itself thus:†

> The TOEFL Internet-based test emphasizes integrated skills and measures all four language skills, including speaking. The content on the test is authentic, and the language is consistent with that used in everyday, real academic settings. The test has four sections:
>
> - Reading measures the ability to understand academic reading matter.
> - Listening measures the ability to understand English as it is used in colleges and universities.
> - Speaking measures the ability to speak English.
> - Writing measures the ability to write in a way that is appropriate for college and university course work.
>
> Test content is based on a "corpus," or database, of spoken and written language that currently contains more than 2.7 million words, collected from educational institutions throughout the United States. (graduateshotline, 2006)

Sampling

The main problem facing a language test constructor is what to sample. If the domain under test is, let us say, ten vocabulary items, then it would certainly be possible to test the entire domain, the whole population that the test is targeting. But in the case of the kinds of proficiency tests where the domain consists of large areas of the language, it is just not possible to test everything. And so the test constructor must sample the domain and face up to the question of how to make rational choices. Should he/she select vocabulary items, and if so which ones; the grammar, again which parts; relevant texts, again which? And so on. Indeed, the only domain that could be completely covered for proficiency testing might be the phonology, but there again, the tester would have to choose which version of the phonology, which accent, which phonetic realizations.

Sampling is inescapable: that is the first of the problems. The second is related. It is what the sample eventually chosen is a sample of. That is to say, while the choice may be to sample linguistic features or forms, the tester still needs to be convinced that those features and forms have a connection (which

† [Ed. note: See also Cohen, Chapter 5, for a description of TOEFL iBT.]

may, of course, be indirect) with the kinds of uses of the language that successful candidates will be capable of. In other words, does the language sample for the test match the criterion?

Such an approach necessarily takes account of argument-based approaches to validity (Kane, 1992): since the interpretive construct for a test involves an argument leading from the scores to score-based decisions, it follows that the language sample for the test acts itself as a corroboration of the interpretive construct.

Over the past 50 years there have been three significant attempts in the U.K. to develop a measure of academic English proficiency: they take up quite different positions on this sampling issue. The first attempt, the English Proficiency Test Battery (EPTB), took a structural approach, sampling grammar and lexis (Davies, 1965). The second, the English Language Testing Service (ELTS), took a strong communicative approach, assuming that proficiency has to be represented by "real-life" examples of specific language uses (Carroll, 1980). And the third, the International English Language Testing System (first and second IELTS), eventually took a more abstract view of communicative competence, sampling what has been called communicative ability (Clapham, 1996).

All three attempts made claims on construct validity, EPTB supported by a structural model, ELTS by a communicative competence model and IELTS by a Bachman interactional authenticity (IA) rather than a real life (RL) authenticity model (Bachman, 1990). IA provides the rationale for determining the most appropriate combination of test method characteristics, thereby offering the compromise we find in IELTS between the claimed spontaneity (or *real life*) of ELTS and the structural generality of EPTB.

The story we can narrate begins in the late 1950s in the heyday of the structuralist approach to language. We note that although the communicative movement was already underway in the 1960s, the inevitable institutional lag meant that the EPTB continued to be used as the main British Council (and therefore U.K.) measure until the end of the 1970s.

The communicative revolution eventually swept all before it, first in language teaching and then in language testing (where it is well to note it was less widespread). In proficiency testing the outcome was the ELTS, which was launched by the British Council and eventually operated jointly with UCLES, the University of Cambridge Local Examinations Syndicate. This test dominated U.K. English language proficiency testing until the end of the 1980s. (It is also worthy of note that, as far as we are aware, no comparable test was developed for any other language).

The revolution had eventually, like all revolutions, to be hauled back, and from about 1990 ELTS gave way to the IELTS, which borrowed a great deal from ELTS but simplified its structure (even more so after 1995, when IELTS

was revised), and greatly improved the delivery, analysis, and production of the test. And if number of candidates is a measure of a test's success, then IELTS has been very successful, with a tenfold increase in the ten-year period up to 2003, when there were more than 500,000 candidates. Below, we ask whether it can survive that amount of success and still remain an acceptable test of communicative ability.

We have also suggested that the explanation for these changes has to do with the view we take of language: it is that view that provides our construct and determines the sampling we employ. In the first period of our history, language was basically seen to be grammar: that eventually came to be regarded as too distant, too abstract. In the second period, language was reckoned to be a set of real-life encounters and experiences and tasks, a view that took "real-life" testing so seriously that it lost both objectivity and generality. In the third period, there has been a compromise between these two positions, where language is viewed as being about communication but in order to make contact with that communication it is considered necessary to employ some kind of distancing from the mush of general goings on that make up our daily life in language.

We can propose two alternative explanations for this development.

Explanation A

During the first (EPTB) period, the pre-ELTS period, from about 1960 to about 1980 (see Table 4.1), language was seen to be structure and hence in the test(s) grammar was given a central role. Lado's advice to "test the problems" was the slogan and so tests concentrated on the component parts of the language, parts such as phonology, stress and intonation, grammar, and so on (Lado, 1961).

The receptive skills were prominent (reading and listening), with reading dominating. After all, language teaching was still under the influence of the classical languages and hence the purpose of all language teaching, including EFL and modern languages, was seen to be to ensure that learners became literate. The model was very much that of the classical languages, but it was also (perhaps itself a spin-off from Latin and Greek) influenced by the teaching of the mother tongue, which again was heavily into literacy, genres, and textual registers. Speaking was sometimes tested, but not in the EPTB; writing was also not included in the EPTB. Indeed, the policy in TOEFL, the contemporary of EPTB, was that both writing and speaking were optional and could be tested in the Test of Spoken English (TSE) and the Test of Written English (TWE) if desired. The TSE, which took 20 minutes to administer individually, came into operation in 1979. The TWE, which began in 1986, took 30 minutes. Over time it became clear that this TOEFL model, with optional speaking and writing components, was no longer considered authentic in terms of the

Table 4.1: English Proficiency Test Battery (EPTB), in operation from 1965 to 1980

Test	Duration	To test	No. of items	Test Contents
1. Phonemes in isolation	LV/SV[1] 12 mins.	perception	65	phoemic discrimination (triplets)
2. Phonemes in context	LV 6 mins.	perception	25	sentences offering phonemic contrasts
3. Intonation and Stress	LV/SV 20 mins.	perception	50	offering intonation and stress cues in conversation
4. Listening Comprehension	LV 18 mins.	understanding of spoken academic texts		items offering 3 texts: general, science, and non-science
5. Grammar	LV/SV 15 mins.	knowledge	50	multiple choice; testing knowledge of syntax
6. Reading speed	LV/SV 10 mins.	reading comprehension and speed	196	using cloze elide items inserted into a 1500-word text
7. Reading Comprehension	LV/SV 15 mins.	understanding of written academic texts	50	using modified cloze; 3 texts: general, science, non-science

Notes:

1. LV = Long Version; SV = Short Version.
2. It was established by regression that the variance shared with criteria would be only minimally reduced if a shorter version of the test was available. The table indicates which sub-tests were presented as forming the Short Version. Given the saving in time and expenses, it is not surprising that the Long Version was rarely if ever used in EPTB testing diets.

growing orthodoxy of the communicative competence approach, which put a heavy premium on real-life language use. EPTB, on the other hand, appeared to be operating at a more abstract level, attempting to assess control over systems and structures rather than real-life language use. It seemed to be too distant from the acts and experiences of communication that we engage in every day and for which teaching (and testing) of the component parts do not seem to prepare us. It was thought to be too remote.

In the second period (the 1980s), the English Language Testing Service (ELTS), which had replaced EPTB, emphasized so-called *real-life* language use (see Table 4.2). Language was seen to be purposeful: hence the field-specific orientation of the test, built on what was called English for Specific Purposes, a cult concept in the communicative language teaching materials of the time.

Table 4.2: English Language Testing Service (ELTS) test, in operation from 1980 to 1990

Choice of 6 Modules covering 5 broad areas of study plus one non-specific area:	
Life Science	Technology
Social Studies	Medicine
Physical Sciences	General Academic
The test consisted of 5 elements:	
General tests:	Modular tests:
G1: Reading: 40 items in 40 mins.	M1: Study skills: 40 items in 55 mins.
G2: Listening: 35 items in 35 mins.	M2: Writing: 2 pieces of work in 40 mins.
	M3: Interview: up to 10 mins.

Notes:
1. G1, G2, and M1 were multiple-choice.
2. For the modular tests (M1–M3) the candidate was given the relevant source Booklet (one of the 6 options), which contained extracts, including bibliography and index from appropriate academic texts. The correct responses to all items in M1 were found in the source Booklet; the tasks in M2 were derived from the Source Booklet and the core of M3 was discussion of material in the Source Booklet.

If the rallying cry for EPTB was "test the problems," for ELTS it was "test the purposes." To that end, ELTS offered a set of modular choices, based on what were thought to be the main academic divisions. However, the appeal to real life revealed itself as all mouth and no trousers. This was especially the case for language assessment. With language teaching it may have been less of a problem because the teacher was always there to provide the necessary context and explain the cultural references. This was not the case for language testing.

If EPTB had been too distant, ELTS was too close for comfort. All intervention (and this includes both teaching and testing) involves some degree of abstraction: it is never real life simply because real life is fugitive and too full of noise. And a sample of real life is not really representative of all other possible encounters, which is why sampling real life is so difficult; we might think impossible.

IELTS (see Table 4.3), increasingly dominant in the third phase (from 1989 to 1995 for the first IELTS and then post-1995 for the revised IELTS, the current model), offered a clever compromise between the EPTB's testing of the component parts and the ELTS's field and purpose testing by its approach to testing communicative ability (or abilities).

IELTS exploits neither features of language (as EPTB did) nor instances of language use (like ELTS). Instead it brings them together by aiming at features of language use. Therefore it quite deliberately eschews any claim to specificity because what it wishes to claim is that the test is generic, potentially

Table 4.3: International English Language Testing Service (IELTS) test, in operation since 1989

1989 Version
Modular tests:
Module A: Physical Science and Technology
Module B: Life and Medical Science
Module C: Business Studies and Social Sciences
Four elements:
Reading: Module A, B, or C or the General (non-specific) test
Writing: Module A, B, or C or the General (non-specific) test
Listening: Non-specialized module: two tasks: 60 mins.
Speaking: Non-specialized module: 10–15 mins.
1995 Version
The three specific modules were reduced to one Academic Reading and one Academic Writing Module: the reading and writing modules were no longer linked (as they had been in the 1989 version). The General Module became General Training for reading and writing and was deliberately made less academic.
Reading: 3 tests: 60 mins.
Writing: 2 tasks: 60 mins.
Modules (including General Training): 60 mins. each
Speaking: 10–15 mins.

generalizable to any type of academic language use. The emphasis has been on tasks and on production. As with ELTS, one of the great selling points has been the obligatory test of speaking. There lies the heart of the communicative aspect of IELTS and it is in speaking tests that the real break is made with the structural tradition. No longer is the rallying cry: test the problems (EPTB) or test the purposes (ELTS). With IELTS it is "test the interactions." IELTS represents a kind of regression to the mean, a (good) compromise between the extremes of the structural and the communicative.

Explanation B

There is another, more complex, explanation of the development.

While grammar was certainly central to the EPTB, the test did in fact take up a somewhat elementary approach to work sampling. The construct included a linguistic component (grammar, phonology, intonation, and stress) and a work sample component (reading comprehension, reading speed, listening comprehension): the first sampled what language is (as understood in the 1960s), the second what language is used for. As has been pointed out, the approach was wholly receptive (only listening and reading): no attempt was made

to sample the productive skills of speaking and writing. In the first version (the long version) of the test there were alternative sub-tests of (a) scientific and (b) humanities texts. This choice was removed from the shorter operational version, largely because the work samples were redundant for predictive purposes. Grammar, along with reading comprehension, was central.

ELTS too was not nearly as pure a representative of the model it promoted since, as well as the field-specific modules it provided, there was also the core test of reading comprehension. Indeed, the prediction delivered by this test of reading comprehension on its own was more or less equivalent to that provided by the entire ELTS battery. What was being predicted was what at the time, in the 1970s and 1980s, was regarded as the criterion of success in higher education, the results at the end of the year examination in the student's academic discipline(s). Since language proficiency was one component only of the students' academic performance, Pearson correlations of the order of 0.4 between the test and a criterion indicative of academic success were regarded as important (see below). To that extent, and from a statistical point of view, the field-specific modules were redundant. However, since a monolithic test of grammar or reading comprehension has, it might be claimed, poor impact on language teaching, the modular apparatus was necessary to ensure good washback.

IELTS moved on from ELTS but not very far. The content of the two tests was similar — the major difference (especially after 1995) was that there were no longer field-specific modules — unless we accept that the Academic Module is specific to academia. And again, in that specificity, what dominates is the reading module. Evidence for matching to academic success is sparse but what there is suggests that, as with both EPTB and ELTS, the IELTS predictive validity correlation with performance at end of year degree/diploma examinations is about 0.3–0.4. In other words, all three tests do a very similar job, in spite of the changes in paradigm, the move back and forth between structural and communicative, and the inclusion of specific purposes testing. Nothing much has changed at the base. The variance shared by all three tests and academic success is still around 10–15%. The normal method for assessing the predictive validity of these proficiency tests was by simple correlation (product moment) between the test (usually taken at the start of the academic year) and the degree or diploma examination taken at the end of the same academic year. The only constant for all those tested was the English language proficiency test, since the subjects students were studying and therefore the examinations they were sitting ranged widely across the whole gamut of academic disciplines. No doubt this helps explain why the shared variance was typically 10–15%. While this is clearly not large, it is probably as large as one might reasonably expect, given the criterion variable used. After all, other factors such as intelligence, academic knowledge and ability, attitude, and health contribute to academic

success. Language is necessary but not a sufficient determinant. If it were so, then native speakers of English would always succeed in academic programs in English medium. Clearly they do not.

Does this then mean that there is no way of choosing among the three tests?

Best Test?

The EPTB and the ELTS were both good tests, both set out to test proficiency in English for academic study, and, although their approaches are (or seem to be) quite different, they both have had much the same degree of success. However, from today's standpoint, both are out of fashion and for the sake of stakeholders, there is much to be said for keeping up with the fashion. They both had very poor delivery, largely because they were produced and delivered (and administered) as part-time activities, the first by a university department, the second by the British Council. There was no program in either case for the production of new versions, and as candidate numbers increased, it became more and more necessary to ensure proper procedures for administration, analysis, and training. EPTB and ELTS were largely one-off operations, they were not maintained with new material on a regular basis, and they did not have the advantage of being informed by new (and ongoing) research. ELTS, unlike EPTB, did test all four skills, it is true, but here again we meet the problem of maintenance: there was no proper professional training program. And they both had weak impact — or, if they had more, that was never known since there was no project in place to check.

IELTS is an improvement in all these features. True, its predictive validity (on the little evidence we have) is much the same as the two other tests. But in all the other aspects it is a superior product. Its communicative ability model is now fashionable. Its delivery (even now with the extra imposition of fixed date testing) is impressive. It is well maintained and research-led. It tests, very deliberately, all four skills. And it has ensured from the mid-1990s that its impact is monitored and the information from that project acted on. Its partnership status is also new and important. It is no longer just a British (or just a British Council) test. The partnership between the University of Cambridge Local Examinations Syndicate (UCLES), now more properly known as Cambridge ESOL, and the Australian International Development Programme (IDP) and of both with the British Council has been generally positive and now, it seems, no partner would consider going it alone or separating off. I suppose the question is whether there are other possible partners that might join, New Zealand and South Africa, perhaps. And then there may be the question of whether a World Englishes community (Singapore, Hong Kong, India) might be interested in sharing. Such a development would be innovative, given that it would mean a move away from the anglophone inner circle hegemony. But it

would speak well to those who still view the British (and the English language) as wishing to continue imperialism by other means.

What the tenfold increase in candidature for IELTS over the ten year period up to 2003 (from under 50,000 to over 5,000,000) suggests is that the test has been successful. This success calls both for rejoicing and for vigilance. Rejoicing, because it demonstrates that virtue does indeed reside in minute particulars, that paying very close attention to details does pay off over time to produce a successful testing operation. But vigilance is also called for, particularly with regard to the increasing uses to which IELTS is put. Its very flexibility could cause it to lose its niche audiences and dedicated stakeholders. Furthermore, from a professional testing point of view, two crucial issues need early attention. The first is the relation between the Academic and the General Training modules. In our view, a decision needs to be taken as to whether they should be far more clearly distinguished from one another or whether they should be combined and outcomes determined on the basis of differential cut-offs. The second issue has to do with the continuing unease about the reliability of both the Speaking and the Writing components. Cambridge ESOL have made serious attempts to develop procedures that will assure stakeholders that IELTS Speaking and Writing are reliable measures, but it does seem that the doubts will continue as long as single marking is retained.

Nevertheless, we may conclude that for prediction alone, grammar, however tested, is good: hence our choice of a test of academic language proficiency would be the EPTB (perhaps brought up to date in terms of content). For face validity in academia (especially with subject specialists), an ESP approach is good: hence ELTS. And for general appeal, we would favour IELTS. But we should be aware that our combining sub-tests or modules together does not of itself add to the prediction: a test of grammar would be adequate on its own.

But a language proficiency test needs to offer more than prediction and therefore it is very important not to end this section with such a reductionist statement. Prediction, we might say, is only one part of what an academic language proficiency test is for. It also needs those qualities we have listed above so that it can be welcomed with the seriousness it deserves by admissions officers, government officials, employers, and by the candidates themselves. In other words, test validity must now take account of washback or, even more widely of test impact (Hawkey, 2006).

What Is Academic Language Proficiency?

There is an irony here. The attempt to define proficiency seems to lead inevitably to aptitude since what we are concerned with in reaching for proficiency is how to predict future performance. That seems to require a measure

of language aptitude, the ability to learn the language of choice effectively when needed. This is not a new idea. In the late 1960s, my work on the EPTB led me to conclude that to measure proficiency we needed an aptitude test. And so with government funding I directed a large-scale language aptitude project. Work over several years led me to the opposite conclusion, that is that the best predictor of future performance is not an array of unconnected skills and abilities assembled to measure aptitude but present performance. What the aptitude project showed, convincingly, was that the best predictor of future language performance is present language performance. And so we can define proficiency in academic English as the ability to operate successfully in the English used in the academic domain. But what does this mean? A helpful approach is that of V. Jakeman (personal communication, September 28, 2005), who considers that IELTS assesses a candidate's current ability to study in an English academic environment. In other words, it measures pre-study rather than in-study ability.

Notice how far we have come from the communicative heyday. It may be too far since we have no way of knowing how we should test every individual candidate's future ability to study in an English-medium environment. This sounds remarkably like an appeal to a language aptitude test, although what we are in fact talking about is a test of final year secondary school language use, a pre-study test, which suggests that the distinction often made between achievement and aptitude may be a distinction too far. On the principle that present achievement is a good, perhaps the best, guide to future success, then it does appear that what IELTS offers is a measure of language aptitude. But, again as we have seen, IELTS has to be more than that if it is to be and remain the test of choice.

So what is academic language proficiency? We avoid both circularity and reductionism by suggesting that academic language proficiency is the language of argument, of analysis, and of explanation and reporting, in all cases not being specific to any particular academic area.

Academic language proficiency is skilled literacy and the ability to move easily across skills. As Pope says of physical agility:

> True Ease in Writing comes from Art not Chance,
> As those move easiest who have learn'd to dance,
>
> (Pope, 1711, ll. 390–391)

In other words, it is the literacy of the educated, based on the construct of there being a general language factor relevant to all those entering higher education, whatever specialist subject(s) they will study. For all three proficiency tests discussed, the core measures, the indispensable tests, are those to do with the written language and primarily with reading.

Van Lier (2004, p. 161), as we have seen, considers that academic discourse cannot be captured in (proficiency) tests: "narrow text-based accountability cultures cut off ... the very means by which academic success is established." He may well be right — indeed he probably is right because the bar of authenticity he is demanding of a test is just too high. Tests cannot be authentically real-life: the best they can do is to simulate reality. That may be what Hyland (2004) is reaching towards:

> Writers always have choices concerning the kinds of relationships they want to establish with readers, but in practice these choices are relatively limited, constrained by interactions acknowledged by participants as having cultural and institutional legitimacy in particular disciplines and genres. We communicate effectively only when we have correctly assessed the readers' likely response, both to our message and to the interpersonal tone in which it is presented. ... For teachers, helping students to understand written texts as the acting out of a dialogue, offers a means of demystifying academic discourse. (pp. 21–22)

These relationships, these interactions, this engagement that Hyland persuasively alludes to, are no doubt central to academic discourse and their representation in even the most valid proficiency test can only be a pale shadow. But unlike academic journals, textbooks, papers, and manuals, tests cannot by their nature use academic discourse tasks, since they require, as Hyland points out, true engagement between the reader/hearer etc. and the stimulus. What tests can do is to simulate academic discourse and incorporate aspects of academic language, its vocabulary, its formal sentence structure, its logical development and its reliance on proceeding by argument. And if we are willing to forego engagement, then the imperative to develop specific purpose tests fades away, vindicating tests of general academic proficiency.

Conclusion

In this paper we have argued that the changes in academic English language proficiency testing in the U.K. over the second half of the 20th century were not random. They were driven by the paradigmatic changes in the climate of opinion about language, following closely the movement over the period from a linguistic view of language, first to a communicative competence view and then, in the 1990s, moving back to a compromise view of language as communicative ability. As such, what the current British/Australian English language proficiency test, IELTS, claims is a general and dynamic capacity to reflect control over interactions in language use rather than a structural (and static) knowledge of language (as in EPTB) or an equally static communicative competence control (as in ELTS).