Validating Tests of English for Academic Purposes

Daniel DUNKLEY

Abstract

This paper introduces English Language for Academic Purposes Tests, which are a type of test of English for specific purposes, and investigates the process of validation. After providing a definition of and rationale for this type of test, we examine how they could be proved to be meaningful tests to satisfy all the stakeholders. One test validation project is described in detail and concluding recommendations are presented.

01. LSP test fundamentals

A Language for Special Purposes (LSP) test is one which assesses a candidate's ability to use language in a particular context, narrower than normal daily life. It is usually associated with professional activity. Thus for example the Occupational English Test, used in Australia, is for health care workers.

Why are LSP tests needed? The answer is that many workers are not native speakers of the language of the country in which they are working. They may be in the country temporarily, as, for example retirement home assistants in Japan who come from south-east Asian countries such as Malaysia or the Phillippines, or long-term immigrants, as are most foreign–born workers in the U.S.A. To a greater or lesser degree all of these workers need to speak and understand the

working language of the country in which they are working.

The lack of language skills can be simply an irritant, as in the case of manual workers, who are slowed down in their learning of tasks. On the other hand, insufficient local language knowledge can be a matter of life and death in some professions. Thus, for example a case occurred in England in 2010 (Daily Telegraph 2010) of a doctor whose incorrect prescription led to the death of a patient. One cause of the mistake was that the doctor was not a native speaker and had in fact failed an English language test. The correct checks both of his linguistic and professional competence might have avoided the accident. Similarly, miscommunication due to insufficient English language skill has been noted as a contributory factor in some fatal aircraft accidents (Cookson 2009).

This leads us to consider who uses ESP tests. Naturally it is those organizations which are responsible for overseeing the professions. In the case of the Occupational English Test it is the health care administration boards in each state in Australia which need to regulate the profession. At a more local level, individual hospitals demand to see an applicant's OET certificate as proof of linguistic competence when hiring workers. In addition, people wishing to immigrate to Australia need to have professional skills which are in demand, but in addition to professional qualifications specific language competence certification such as the OED is required by the immigration authorities.

Historically speaking, LSP tests appeared much later than general purpose language tests. In the UK for example, the international test of English known as the English Language Test Battery, the predecessor of IELTS, was created in the early 1960s and used by the British Council around the world from 1965 to 1980. On the other hand, one of the first ESP tests, the English language test for medical practitioners known as the Temporary Registration Assessment Board (TRAB) examination was only introduced by the General Medical Council in

1975. This date is when a growing number of medical staff were arriving in the UK from countries where English was not spoken, or where English was a common language of public and business life but not of home life.

How specific is a specific purpose test, one might well ask. Surely there is a vast overlap between the English needed in general life and that needed in a specific profession. However, in spite of this overlap, the special knowledge seems to take priority and determine the outcome of an ESP test. In this connection, a seminal study by Clapham suggested that background knowledge was very important in Business English tests. "Subject area familiarity made a significant contribution to test scores" (1996, 193). Thus, even though the boundaries of a topic, such as business or medicine, are ill-defined, there is a clear body of language which is unquestionably in the topic area, and an equal body of language outside. Candidates with knowledge of the target language (known as the target language domain) are favored, while laymen will score poorly. Accordingly we notice a salient feature of ESP tests: whereas on a general language test specific knowledge of the topic is simply a matter of luck, in an ESP test the specific knowledge requirement is desirable and fair.

02. English for Academic Purposes (EAP) tests

First we present a word of justification for this type of test. It might seem at first sight strange to suggest an English for Academic Purposes (EAP) test. Surely academic English is just advanced general English. No-one would disagree that a native speaker at age 10 and one age 18 are both competent in their language, although of course this competence is very different in many respects. One might suggest that academic language by extension is simply that which one naturally speaks at age 21 or 25. However, in practical terms

English language universities have found that a general qualification in English does not equip students who are not native speakers to deal with the demands of university life. As the number of international students in English-speaking universities has grown, this requirement has become more acute.

What does an Academic English Test look like? Two essential features of ESP tests are, firstly, that the type of processing of the question should mirror that from the target domain which the test replicates, and secondly, that the text or recording with which the candidate is confronted is similar to that occurring in real life. The first is sometimes called interactional authenticity and the second contextual authenticity. Thus for example, in an airline pilots' ESP test processing especially of spoken language is vital, because real-time decisions based on language heard over the radio in the cockpit involving accurate comprehension, fast judgment and precise expression are essential. Equally, an example of context authenticity is that the language in a ESP medical test must be up to date medical jargon, and not just what a non-medical item writer imagines is medical language. Weir (2005) calls the degree of authenticity of the processing "cognitive validity" and the accuracy of the language "context validity".

Naturally all tests, while using natural authentic or real-life language and relevant cognitive tasks, are severely limited. In only a short time the test is trying to collect evidence to judge how well the candidate will deal with a particular language use situation in the future. To take the example of reading questions, in an EAP test the passage will be far shorter than even one chapter of a real text book. Equally, the questions must try, quite artificially, to prompt a variety of different types of processing in a short space of time.

Faced with these constraints on authenticity of both language and processing, to what extent can we construct a worthwhile EAP test? It has been suggested (Fulcher 1999) that because of the social baggage which authentic texts bring,

artificial but fair texts should be employed. In effect authenticity is a mirage. Thus, ironically, we will test candidates' ability to read real text books by creating fake book passages. We are aiming to achieve two goals here, namely both construct validity and content validity. The position is that we must motivate the candidates to take the test seriously, so we must create in their minds the "perception of authenticity" (Bachmann and Palmer 1996, 23–4). In effect we create pseudo authentic texts to increase the test's face validity (or "response validity" (Henning 1987, 92)). Fulcher (1999, 234) argues for an approach in which "content relevance … can be considered in the light of construct validity." In other words, to ignore the construct we are assessing, namely academic English ability, in favor of the content would be to reduce the effectiveness of the test.

03. Academic Language

If we are going to select or create academic texts (to be read or heard by the candidates), then the first requirement is an understanding of the characteristics of academic language. This has several aspects, of which vocabulary, grammar and rhetorical structure are the most basic.

In the field of vocabulary, there are several useful academic word lists to guide both test writers and test takers. A practical example of an academic vocabulary resource is Coxhead's Academic Word List (Coxhead 2000).

Thanks to computer technology, counting vocabulary frequency has become quite manageable. An academic English text corpus of 3,500,000 words from books, course workbooks and other texts in both arts and sciences was used, and to give an equal weight to each subject area and topic, each area was divided into seven subject areas. For example the Science subject area contains the

topics biology, chemistry, computer science, geography, geology, mathematics and physics. Nearly all (94%) of the list's words occur in 20 of the 24 subject areas, and as a consequence whatever university subject is to be studied, the AWL words are useful for all students. On the other hand, the only words to be excluded are the 2000 most frequent words in West's 1953 General Service List. This seems a rather arbitrary boundary, but in practical terms a precise cut-off point must be agreed.

In addition to this computer-generated word list, there are other methods of comparing and grading texts. Several projects from the field of computational linguistics have resulted in software tools which can both assess student essays and distinguish between authentic and specially written examination texts. These have greatly helped to improve our understanding of text difficulty. Indeed, previous readability formulae such as the Flesch-Kincaid Grade Level, which were objected to because they could not accurately show the complexity of a text, have been superseded by new readability formulae based on a wider range of criteria, such as that of Crossley et al. (Crossley Greenfield Mc Namara 2008).

Using these tools and word lists it is possible to answer the question "Is this text really academic?" or more precisely "Is it like a current textbook?" We are aiming to provide evidence to prove that the text is representative, in its form, content and style, of the academic texts which students will encounter on their future academic courses.

Academic vocabulary is one obvious characteristic of academic texts, but there are several others. In the last 20 years several researchers have analyzed academic texts and produced lists of characteristics. These have two functions: one is to orient the text writer to real academic style, and the other is to express both qualitatively and quantitatively what underlies text difficulty. For example, a set of criteria specifically designed to help examination writers to accurately

create test text passages at different proficiency levels was developed for the Cambridge ESOL examining organization by Khalifa and Weir (2009).

Let us present the broad outline of this taxonomy. The first feature of an academic text is that it is long. A university student must read many lengthy texts in a short time, and weakness in this skill hampers a student's progress. Not surprisingly this is one feature of a text that unfortunately cannot be replicated in an examination. Secondly the vocabulary and grammar of EAP tests is specific. Academic word lists have been produced, as noted above, to list and prioritize academic vocabulary. Additionally, academic grammar is special, and in fact grammatical knowledge has been shown to be as important as vocabulary knowledge in success on EAP tests (Shiotsu and Weir 2007). Thus at a low level of proficiency (Common European Framework of Reference (CEFR) level A2) only simple sentences appear in texts, whereas many complex sentences are a typical feature of high proficiency level tests (CEFR levels C1 C2). The third feature which helps testers to classify texts is cohesion, indicated by the use of transitional phrases such as "on the other hand" or "by contrast". Enright (1991) has shown that when the frequency of these devices increases, candidates' scores increase, and inversely, when asked to read a text in which they are rarer, scores decrease. Fourthly genre, defined as the communicative purpose of a text, plays a role. In academic texts we expect to meet genres such as classification, process description or arguments, rather than narrative or exhortation.

A fifth feature which relates to text difficulty is that of subject and cultural background. A neutral observer might well ask why, in the TOEFL and IELTS tests all academic disciplines are put together. If we have an English for Nurses test and a separate English for social workers test, why should aspiring university students not have a test tailored to their discipline, be it, for example, engineering or medicine? Is this concentration into one exam just a matter

of ease of administration and economy of test production, or does it serve a useful purpose? A glance at the history of the IELTS test is worthwhile at this point (see Davies, 2008). When the IELTS test was first produced in the 1980s there were subject specific modules, but they were abandoned in favor of a unified academic module after only 6 years of operation in 1995. The reason for this reversal of policy seems to be based on both practical considerations and research. On the research side, a series of studies described in Clapham (1996) found that subject knowledge had little effect on scores compared with linguistic ability. However, this remains a topic of debate, since several researchers have come to diametrically opposite conclusions, such as Khalifa (1997).

Though there is uncertainty about the exact role of subject knowledge in text difficulty, the role of cultural knowledge is clearer. Cultural knowledge, such as knowledge of traditions, religions and festivals or of local social behavior, has been shown to affect text comprehension markedly, as in Sasaki (2000). As a result of these considerations, current IELTS writers are instructed to avoid texts which assume knowledge of particular topic areas or cultures However, this paradoxically makes the text less authentic, since most text books are written and used in one cultural situation, with scant regard to readers from other cultures.

The sixth and final feature affecting reading difficulty is abstractness. As with native language learning, the early and easy parts of the language deal with concrete things, while the more difficult parts the language are reserved for abstract ideas. Indeed, academic discourse is largely concerned with theorizing, and accordingly EAP tests should be expected to contain preponderantly abstract and thus difficult language.

04. The process of validation

To satisfy the stakeholders' demand that a test is valid, the test maker needs to prove both that the language is typical of that in the target use domain, and also that the tasks demanded of the examinee are typical of those of people operating in that domain. Let us focus on the texts. Is it really possible to produce evidence sufficient to convince the test takers and stakeholders that the texts used are a fair reflection of real-life academic texts? This is vitally important to the standing of the examination. At present there are two main EAP tests, TOEFL, used by US universities, and IELTS (Academic Module) used by UK and Commonwealth institutions. They are publicized with bold claims of authenticity. For example, IELTS states the reading question texts are "based on authentic texts..designed to present the test taker with material similar to those which they might need to read on a university course." (IELTS 2007, 1)

A recent research project (Green, Unaldi, Weir 2010) set out to examine this claim by comparing IELTS exam texts with real academic texts. The researchers compared 42 passages from 14 undergraduate textbooks with 42 sample test reading passages. All passages were of the same length, that is to say about 900 words. Texts were analyzed in two very different ways: firstly by using software tools, and secondly by having them reviewed by a panel of expert judges. The results were statistically analyzed to compare the two sets of texts. The software tools gave many details of vocabulary, such as word length, lexical density (number of content words relative to grammatical words) and word frequency. In addition the grammatical complexity was measured by software, giving data such as the logical operator incidence. This is the number of words such as *and* and *or* relative to other words, a feature which has been shown to cause difficulty in reading. In addition readability statistics such as the

Flesch Reading Ease were collected.

The next area to be investigated that of cohesion, meaning the use of words that guide the reader. This was quantified by the anaphor reference function. Latent Semantic Analysis tools were also used at this point, in order to compare the degree of similarity between the test texts and real academic texts. When it came to genre and rhetorical task, the researchers relied on a panel of three human judges. Weir et al's (2009) taxonomy of reading activities was used to establish the list genres. As for rhetorical tasks, meaning the precise aim of the author in the text, Enright's (2000) categories of *exposition, argumentation* and *narrative* were adopted. To establish which subject area each text referred to, the judges were asked to assign each text to one of eighteen subject areas. Similarly, because the IELTS test specification explicitly states that they avoid cultural bias in the test, the judges gave each text a score for cultural specificity. For example, historical events which are specific to one country, or products which are not universally known increase the cultural specificity rating. The final feature of the texts to be investigated was their abstractness, a high level of which is typical of academic texts, especially in the social sciences. This was evaluated by software, which produced three different relevant statistics.

After these two types of analysis, one by computer, the other by human judges were performed, the results gave a comparison of IELTS texts with real academic texts. Comparing their academic vocabulary content, using Coxhead's Academic Word List as the measure, IELTS texts had only 2.2% compared with 4.3% for the academic texts and 10% for Coxhead's corpus. On the other hand the level of lexical density was similar for both IELTS and the academic texts. Equally, judging from readability measures, IELTS texts and academic texts were similar. Taking the measures of cohesion coherence and explicitness of rhetorical organization, the two groups of texts were similar except in their conceptual cohesion. This, the authors suggest, may be a result of the text being

shaped to allow many questions. In the investigation of genre, the researchers considered the sources of the test texts, namely books, research reports and newspapers. However, more of the IELTS texts were journalistic compared with the real academic texts; this could be attributed to the text writers' search for broadly academic but not too narrowly focused subjects. In respect of subject and cultural specificity the real academic texts were different from the IELTS passages. This however, is not an accident, but rather reflects IELTS' policy of avoiding technical topics and culturally specific texts. Ironically therefore, the more the test writers adhere to IELTS guidelines, the less the passages resemble real academic texts.

As a result of combining these separate comparisons of many aspects of the IELTS with real academic texts, the authors found that the claims made for the texts as broadly speaking academic are justified. "… the texts appearing in the test do … have many features of the kinds of text encountered by undergraduates." However they point out that in some respects the IELTS texts are easier than real texts. "Texts at the level of the most challenging undergraduate textbooks are not represented on the test" (Green et al 2010, 207). Furthermore they recommend that test text writers should use software to check that their texts reach the official standards.

This example of a validation study is one of many ways of investigating an academic English test. Another approach is to compare results on the test with later academic progress, to investigate the test's predictive validity. Equally the relative scores of various groups of candidates, classified by features such as native language, ethnicity, age or gender can be compared to trace possible bias.

05. Conclusion: the future

In spite of the impressive rigour and seriousness of these validation studies, one crucial question remains. Most of this validation research is commissioned by the examining organizations, whether Cambridge ESOL tests in the UK, the Educational Testing Service ETS (in the USA) or other administering body. Thus a possible conflict of interest arises, often given the popular title of "Who polices the police?" The impartial observer is bound to suspect that there is a tendency for the employee-researchers to show their masters in the best possible light. On the other hand, the only defence of the commissioned papers is that they are published in academic journals and can thus be questioned freely by the academic community. Additionally, the various stakeholders in the test-the universities, candidates, teachers and future employers of the graduates-need to have channels of communication to the examination producers, with the ultimate aim of increasing the tests' fairness and effectiveness. As Douglas (2000, 258) states, test validation is not like a snapshot of the test but rather an ongoing process: "Validation is not a once and for all event, but rather a dynamic process in which many different types of evidence are gathered and presented in much the same way as a mosaic is constructed."

References

Bachmann, L. F., Palmer, A. S. (1986) *Language Testing in Practice*. Oxford: OUP.

Coxhead, A. (2000) A New Academic Word List. *TESOL Quarterly*, 34 (2), 213–23.

Clapham, C. N. (1996) *The Development of IELTS: A Study of the Effect of Background Knowledge on Reading Comprehension*. Cambridge: Cambridge University Press.

Cookson, S. (2009) Zagreb and Tenerife: Airline accidents involving linguistic factors. *Australian Review of Applied Linguistics*, 32 (3), 22.1–22.

Crossley, S. A., Greenfield, J., McNamara, T. (2008) Assessing text readability using

cognitively based indices. *TESOL Quarterly*, 42(3), 475–493.

Daily Telegraph (2010) Daniel Urbani Inquest 5th February 2010.

Douglas, D. (2000) *Assessing Language for Specific Purposes*. Cambridge: Cambridge University Press.

Ehrlich, M. F. (1991) The processing of cohesive devices in text comprehension. *Psychological Research*, 53 (2), 169–174.

Enright et al. (2000) *TOEFL 2000 Reading Framework: A Working Paper*. TOEFL monograph series 17 Princeton NJ ETS.

Fulcher, G. (1999) Assessment in English for Academic Purposes: Putting content Validity in Its Place. *Applied Linguistics*, 20 (2), 221–236.

Green, A., Unaldi, A, Weir, C. (2010) Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27 (2), 191–211.

Henning, G. (1987) *A Guide to Language Testing*. Cambridge MA: Newbury House.

Khalifa, H. (1997) *A Study in the construct validation of the reading module of an EAP proficiency test battery: Validation from a variety of perspectives*. Unpublished PhD Thesis University of Reading.

Khalifa, H. and Weir, C. (2009) *Examining Reading*. Cambridge: CUP.

Sasaki, M. (2000) Effects of Cultural Schemata on students' test-taking processes: a multiple data source approach. *Language Testing*, 17 (1), 85–114.

Shiotsu, T. and Weir, C. (2007) The relative significance of syntactic knowledge and vocabulary breadth in the prediction of second language reading comprehension test performance. *Language Testing*, 23 (4), 99–128.

Weir, C. (2005) *Language Testing and Validation: An evidence-based approach*. London: Palgrave.

Weir, C., Hawkey, R., Green, A. (2009) The relationship between the academic reading construct as measured by IELTS and the reading experience of students in the first year of study at a British university in Taylor. *IELTS Research Papers*, Vol. 9 (157–189). London: British Council.