*Applied Linguistics 20/2: 221–236* © Oxford University Press 1999

# Assessment in English for Academic Purposes: Putting Content Validity in Its Place

GLENN FULCHER

University of Surrey

Testing and assessment in English for Academic Purposes (EAP) contexts has traditionally been carried out on the basis of a needs analysis of learners or a content analysis of courses. This is not surprising, given the dominance of needs analysis models in EAP, and a focus in test design that values adequacy of sampling as a major criterion in assessing the validity of an assessment procedure. This article will reassess this approach to the development and validation of EAP tests on the basis of the theoretical model of Messick (1989) and recent research into content specificity, arguing that using content validity as a major criterion in test design and evaluation has been mistaken.

## INTRODUCTION

In an English for Academic Purposes (EAP) context, we may wish to test for a number of reasons. The most likely of these is to select students for entrance to academic courses (proficiency testing), place students into appropriate courses before or during their academic studies (placement testing), and to measure their achievement on EAP courses (achievement testing). Whatever the purpose, there is a widespread assumption that EAP tests should be based on an analysis of students' needs, similar to those undertaken using the Munby (1978) model. Indeed, in much of the English for Specific Purposes (ESP) literature, there is an assumption that the specifications for an EAP test should flow as naturally from needs analysis as the EAP course itself (McDonough, 1984: 111). Carroll (1980: 13) saw the design of a test to comprise three elements: describing the participants, analysing their 'communicative needs', and then specifying test content.

Much of the discussion on EAP testing and assessment stems from this basic assumption that there should be a direct link between course and assessment at the level of content. This article investigates this assumption with reference to theory and relevant research. It argues that the content on EAP assessment has detracted from the main question of how we draw valid inferences from test scores. Finally, consideration is given to the likely future of EAP tests.

# CONTENT VALIDITY IN EAP

## Content validity: Defining the concept

Although Carroll (1980: 66–8) recognized the need for construct validation studies to be conducted in EAP contexts, it is content validation that has received the greatest attention in the literature. The argument for the centrality of content is summarized by Ebel (1983: 8, cited in Messick 1989: 41):

> The evidence for intrinsic validity [content validity] will consist of an explicit rationale for the test: a written document that (a) defines the ability to be measured, (b) describes the tasks to be included in the test, and (c) explains the reasons for using such tasks to measure such an ability. The explicit rationale indicates what the test is measuring. If that is what the user intends to measure, the test is a valid test for the user's purposes.

In EAP testing, this translates into the degree to which the test accurately samples from the EAP course of study and/or some future study domain. This has been termed the 'real life' approach to validity, in which all test tasks should be 'a representative sample of tasks from a well-defined target domain' (Bachman 1990: 310). It is clear that this relates not only to *test content*, but also to *test format*. The target domain should be described in the needs analysis in terms of both content and the language use situations in which the student will be expected to survive, and the analysis is then translated directly into test content and task type. Hutchinson and Waters (1987: 144), for example, state that in any ESP context 'assessment takes on a greater importance . . . because ESP is concerned with the ability to perform a particular communicative task'.

Within this understanding of EAP assessment the issues of validity and sampling are almost indistinguishable. This situation is represented diagrammatically in Figure 1. The needs analysis is intended to be the accurate description or definition of the target domain, and the test samples 'real' content and 'real' tasks from the target domain.

Within this approach it is also important to recognize that a number of other terms have become inextricably linked to the concept of content validity. These terms have become 'buzz words' in EAP testing, and so their use needs clarification. The most important are 'authenticity' (also see the discussion in Douglas 1998), and 'face validity'.

The degree to which sampling is successful is frequently expressed as the degree to which the test is 'authentic'. If the test looks authentic, it is then said to have 'face validity'. Thus, Bachman (1990: 307) can say that 'face validity is the appearance of real life (and) content relevance is the representation of real life and predictive utility is essentially precluded without authenticity.' The term 'authenticity' in language testing has therefore come to mean 'the degree to which the outside world is brought into the testing situation'.

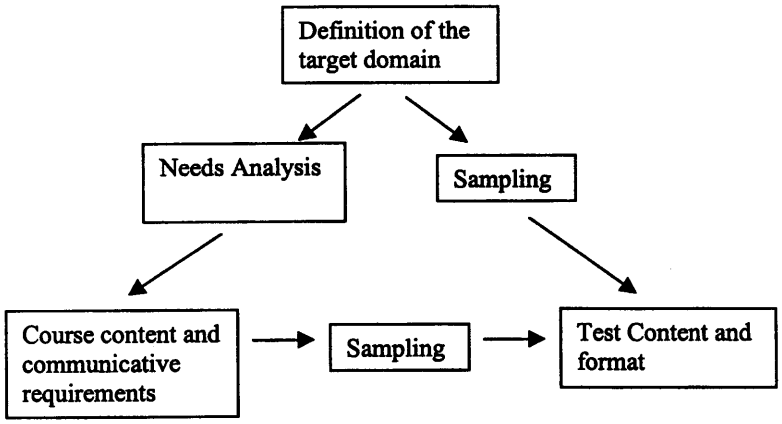


*Figure 1: Validity as sampling*

The relationship between the concept of authenticity and sampling had been prefigured in a discussion reported by Alderson (1981: 57–8), although the close link between the two was not overtly established. However, unlike most other writers in the late 70s and early 80s, Alderson realized that defining why a learner was capable of successfully completing a task (in a test situation or in 'real life') was an extremely complex issue. Alderson clearly points to the fact that there was no attempt to ask the deeper questions about ability or competence. Yet, for most writers of the time, describing the communicative situation and simulating it in the test task had been enough.

## Content validity: The problems

The central problem with the popular notion of content validity is that the exclusive use of the principles of content, authenticity and sampling in testing has led to a simplistic view of validity. Messick (1989: 36) states that:

> in practice, content-related evidence usually takes the form of consensual professional judgments about the relevance of item content to the specified domain and about the representativeness with which the test content covers that domain.

It has frequently been claimed, for example, that authentic tests are valid simply because they are authentic, and *look good*. At the extreme, the notions of content and face validity become touchstones of test validity:

> Construct validity is not an easy idea to work with, and indeed may not have much value outside language testing research. To reduce it to its simplest statement it says: does the test match your views on language

> learning? In practice, there may be little difference between construct and content validity. (Underhill, 1987: 106)

For Underhill, content validity is the same as asking whether the test content reflects the syllabus of a course, the aims of a programme of study, or the 'needs' as set out in a needs analysis. He concludes (ibid. 106) that 'Validation must then rely to a great extent on the test designer's intuitive knowledge of the implicit objectives of the programme.'

These 'naïve, face-valid judgements' (Stevenson, 1985: 111) about what tests are measuring are problematic for two major reasons.

Firstly, they reinforce the mistaken assumption that we *validate tests*. Indeed, in the preceding discussion, we have talked overtly about *test validity*, and the *touchstones of test validity*. But it is not tests that are valid or invalid. It is the inferences we make and the actions we take on the basis of test scores for a specific purpose that are valid or invalid. As Messick (1989: 13) states:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences* and *actions* based on test scores or other modes of assessment. (Emphasis in the original)

Messick (1989: 41) goes on to state the problem succinctly:

> The major problem is that so-called content validity is focused upon test forms rather than test scores, upon instruments rather than measurements.

For Messick, selecting content is an act of classification, which is in itself a hypothesis that needs to be confirmed empirically.

The second problem is that it is extremely difficult to define a target domain. For example, Wiliam (1997) considered situations where the domain seems fairly easy to describe, and provides the example of life saving. Here, the test taker is judged on whether they can rescue a drowning person from the sea. However, there are additional variables that must be taken into account. On the one hand there are personal variables such as stamina, strength and energy; on the other, there are external factors such as wind speed and direction, water temperature, and the strength of currents. Whatever decision is taken as to the 'standard' setting for any of these variables during the test will result in one person being favoured over another. In criterion referenced assessment, the procedure is to look at what the 'normal' or 'average' person can achieve in 'normal' or 'average' conditions. The process of description of the domain therefore requires that observers take decisions, and the resulting test is dependent upon the judgments of the observers, and their values.

In EAP the situation is more difficult, as expert judges disagree on what constitutes the content of academic English, and what levels of performance are necessary for academic success. We cannot therefore take it for granted

that we can achieve anything near the exhaustive content definition that is required for the assessment of content relevance.

The use of terms like 'content validity', 'authenticity' and 'face validity' have all too often become the marks of 'good practice'. But the context of their use is usually non-theoretical, and the practice associated with them often anecdotal at best. In what follows, I intend to outline the relevant elements of Messick's theoretical framework for test validation as a basis for evaluating work on assessment and testing in an EAP context, and review the empirical evidence for content specificity, with reference to the International English Language Testing System (IELTS) and the Test in English for Educational Purposes (TEEP) as examples.

## PROVIDING A THEORETICAL FRAMEWORK

The theoretical framework used in this analysis is based on the work of Messick (1989). This is summarized in Table 1 below, which has been slightly adapted from Messick (1989: 20). The discussion that follows in 'Evidential basis for test interpretation' and 'Evidential basis for test use' considers construct validity and content relevance and representativeness, which are the parts of the model most relevant to the argument. Consequential validity is considered in relation to concept of face validity in 'How specific is specific?' below.

*Table 1: Dimensions of validity*

| Justification | Interpretation | Use, or test Utility |
|---|---|---|
| Evidential basis | Construct Validity (substantive and structural aspects) | Construct Validity (including content relevance and representativeness) |
| Consequential basis | Value Implications | Social Consequences |

### Evidential basis for test interpretation

Firstly, we will consider the evidential basis for test interpretation. That is, the empirical evidence that is required for the meaningful interpretation of test scores.

Construct validity is the overarching concept in the evaluation of testing practice. To paraphrase Messick (1989: 35), there are two major threats to construct validity and hence score interpretation. The first of these is construct under-representation, and the second is construct irrelevant variance. Construct under-representation occurs when the test does not adequately represent the construct it is meant to be measuring: something important

about the nature of the construct is left out. In such a case, the test scores cannot be properly interpreted in the light of the construct. Construct irrelevant variance is reliable test variance that is not related to the construct the test is intended to measure, but something else that we do not wish to measure. In this second case, it is possible to misinterpret test scores that systematically differ according to some factor that is irrelevant to the purposes of the test. Most validity research is an attempt to reduce the effects of these two threats to construct validity.

The substantive aspect of construct validity is 'the extent to which the content of the items included in (and excluded from?) the test can be accounted for in terms of the trait believed to be measured and the context of measurement' (Loevinger 1957: 661, cited in Messick 1989: 43). When test items have been written, they are piloted. Decisions are taken on which items to retain and which to reject on the basis of evidence collected during the pilot trials of the test. This evidence consists of measures of internal consistency, facility and discrimination, Item Response Theory (IRT) data, or any factor studies that may be conducted. In other words, researchers look for unidimensionality. This is to be understood as statistical homogeneity for the purposes of interpretation, and not the psychological simplicity of constructs, as demonstrated by Henning (1992).

This approach does not mean that any items can be written and placed in the pilot tests. There must be a theory underlying item construction, so that we may ask whether and to what extent that (or another) theory can account for the remaining items once poor items have been rejected. Advocates of 'communicative testing' may balk at this notion, as the selection of items on statistical grounds weakens the hold of the notion of content validity held within the communicative testing circles. But as Messick (1989: 43) makes clear, rejecting poor items has the effect of 'minimising construct contaminants' and strengthening the empirical domain structure that 'underlies the final test form and ultimate score interpretation'. Piloting tests, item analysis and item revision, is therefore a key part of the process of defining content and the domain to which inferences can be drawn. In other words, it is an attack on construct under-representation and construct irrelevant variance.

The structural aspect of construct validity is concerned with the structure of the construct, the structure of the test, and the structure of the scoring model to be used. If both the construct and the test are unidimensional, it is appropriate to report single scores. However, if the construct is multidimensional, and can be demonstrated to be multidimensional from evidence collected from sub-test scores, then profile scoring is more appropriate. Whatever the decision, the emphasis should always be upon the interpretability of the test scores. If a test is known to be multidimensional, reporting a single score is meaningless, as it represents an addition of dissimilar things.

## Evidential basis for test use

Test *use*, as opposed to *interpretation*, although clearly closely related, involves the process of actually making decisions on the basis of test scores.

It is content relevance and representativeness within this framework that is the closest to the concept of 'content validity' in communicative language testing circles. Content relevance is related to ensuring that the test is relevant to the knowledge, skills or abilities important in the domain, as described in the domain analysis. This implies that there must be an analysis of the domain to which the test must generalize, and that the description should inform the test specifications. It is here where a needs analysis, or content analysis of some kind, is most appropriate. The purpose of the analysis and the description is to take every precaution at the stage of test construction to reduce irrelevant variance from as many sources of contamination as possible. Once items have been tested and the final test administered, this makes inferences more sound and decisions more just.

The test items must not only be *relevant* to the domain, they must also be *representative* of the domain. That is, the domain boundaries must be defined in such a way that we can say what constitutes the domain, and what lies outside it. If there are any facets of the domain, in terms of distinctive behaviours or traits, these should be described, and their relative importance within the domain should be estimated. This final point is important when it comes to deciding how many test items should be constructed to tap each facet of the domain.

From this position, it is clear to see that the process of domain specification and ensuring relevance and representativeness is a key component in providing evidence for construct validity. The domain specifications are crucial at the phase of test/item construction, and in test use, where the relevance of the scores to specified domains is important. This limits the possibility that construct irrelevant sources of variance will enter into score interpretation.

Finally, it should be noted that the difficulty of items or task types, the quality of the rubrics and the accuracy of keying for scoring, all come under this category, even though these are considered to be technical aspects of item construction. The reason for this is that inaccuracy or inappropriate difficulty levels create systematic construct irrelevant variance, which is a threat to test validity.

Having considered content validity as it is often understood and practised, outlined the major problems associated with content validity, and then seen how content validity should be placed within a larger theoretical framework, we now turn to consider the empirical evidence that has been collected over two decades of research into EAP testing. The review is not intended as a criticism of the tests to which reference is made. On the contrary, it is the development and evolution of the tests selected as examples that have provided the data necessary for recent research to assess the principles of

EAP test design in ways that embrace theoretical approaches to content validity.

## CONTENT VALIDITY IN PRACTICE

### The ELTS test and the TEEP

Carroll (1980) was among the first to produce test specifications for a 'communicative' EAP test: the British Council English Language Testing Service (ELTS). The specification framework consisted of:

- Details of the participant
- Purpose of study
- Settings for English (physical and psychological)
- Interactions involved (social roles)
- Instrumentality (face-to-face, text based, receptive/productive)
- Dialects of English
- Proficiency Target Levels
- Communicative Events and Activities
- Attitudinal tones
- Language Skills (a taxonomy of 54 subskills)
- Micro-functions (persuading, advising etc.).

The development of the ELTS test was the first attempt to produce a communicative language test of EAP on a Munby type model, with six modules: business studies, agricultural science, social survival, civil engineering, laboratory technician and medicine. For each of these roles, profiles of a 'typical' student were to be described. Eventually, the six modules to be produced were: Physical sciences, Life Sciences, Social Sciences, Medical Sciences, Technology, and General Academic.

At the time, doubts concerning the ELTS project were voiced, many of which have since proved to be correct. The first was that the needs analysis was never carried out; rather, the test designers constructed profiles of the 'typical' student's needs on the basis of armchair reflection (Clapham 1981). In other words, there was never any empirical data. Secondly, the notion that input, particularly in the form of texts specific to agriculture or medicine covering all branches of a discipline, was questioned (Criper 1981). The search for discipline-specific texts was unlikely to satisfy anyone, as they could never be specific enough. This observation has generated one of the most significant ongoing debates in EAP testing: namely, how specific is specific? And, how specific do we need to get? By the early 1980s it was recognized that 'The fact is that the Communicative Needs Processor does not help one to select texts or items for a test' (Alderson 1981: 128).

The mismatch between the real-world orientation of the content validity approach and the scoring system for the oral test was also called into question. Fulcher (1987) tried to show that the ELTS rating scales were incapable of

accounting for the speech of educated native or non-native speakers. Reviewing this challenge to content validity, Wood (1993: 236) says that:

> Fulcher manages to demonstrate . . . that the ELTS categories signally failed to capture 'real-life' communication. . . . Fulcher concluded that a fixation with content validity, brought about by excessive allegiance to the Munby taxonomy of skills, has resulted in construct validity being ignored. . . . If, as critics claim, the Munby taxonomy is merely a role call of desirable subskills, with strictly speculative status, then the ELTS is bound to be vulnerable to the kind of criticism Fulcher makes.

Finally, and perhaps most importantly, any attempt to address construct validity questions is notably absent from most of the early EAP test development literature.

The additional facts that the original test specifications for the ELTS were 'mislaid' (Alderson 1988b: 222), and that a validation study provided no empirical evidence for the content validity of the ELTS test (Criper and Davies 1988) as understood at the time, led to calls for its revision (Hughes *et al.* 1988).

The Test of English for Educational Purposes (TEEP) is the other major EAP test that has been devised in the last 20 years, based on a major empirical needs analysis undertaken by Weir (1983) for the Associated Examining Board (AEB). The test development process avoided the problems of armchair speculation that beset the ELTS test, but faced similar and related problems because of the reliance on Munby type models for establishing content validity. Alderson (1988b: 223) argues that the vast amount of data collected by Weir was simply too complex to form the basis of an operational EAP test, and the result was an attempt to find 'commonalities' between students of all subjects that could form the basis of the test. Commenting on Davies' (1965) view that little work had been done in the description of specialist uses of English that could be used for test construction, Weir (1988: 53) wrote:

> The situation has improved only slightly since the establishment of the English Proficiency Test Battery. Analyses of the discourse used in the vast variety of courses under review are still not available. Given this current lack of analyses in EAP/EST we were forced to compromise. . . . We found that there was a good deal of common ground between students in different disciplines and at different academic levels, in terms of the types of activity faced in the various study modes, the attendant performance constraints and the levels of difficulty encountered.

The TEEP test contained a common core test, followed by a choice between a science and engineering paper, or an 'everyone else' paper. The TEEP solution to the notion of specificity is very different to that of the original ELTS test, and one that has survived the test of time.

We now turn to look at the question of specificity in some detail.

## How specific is specific?

We have seen that the TEEP only had two options, whereas the ELTS had five specific modules and one general academic module. In the pilot version of the TEEP, however, there was a grammar test that was not included in the final operational version because it was not 'EAP' in nature. Weir (1988: 72) admitted that 'the test of grammar might be a sufficient indicator on its own of a student's ability to cope with the language demands made on students by English medium study.' This deserved more investigation at the time, but the push towards some degree of specificity, driven by the notion of content validity, precluded such an investigation. Alderson (1993) also indicates that a grammar test was the best predictor of other scores on the IELTS, but that this was not included in the final test because it did not look like a specific EAP sub-test. It would almost appear that the commitment to specific purposes testing as it has evolved since the late 1970s resulted in empirical findings being given less attention than they deserved.

Research by Alderson and Urquhart (1983; 1985a; 1985b) and Alderson (1988a) into specificity of content in reading tests also produced results that were difficult to interpret; because of the mixed results from their studies, they

> suggested that both linguistic proficiency and background knowledge in the most general sense, might have compensatory effects and their absence the converse, but were unable to clearly establish that such was indeed the case, nor to determine whether there might be said to be some threshold level of proficiency below which superior background knowledge might have a considerable compensatory effect but above which the trade-off might be less. (Alderson 1988a: 23)

In later studies using the ELTS test, Alderson and Urquhart again obtained mixed results. Engineering, science and mathematics students did better on their specific module and worse than others on general modules; liberal arts students performed better on all modules except technology; engineering, science and maths students did just as well on social studies as economists. Alderson (1988a: 24) claims that this research has demonstrated a broad-based support for EAP testing as practised, but also acknowledges (ibid. 26) that 'the uncharitable interpretation of these results would propose getting rid of the M1 [subject specific] modules altogether, since they yield rather similar results to the G1 and G2 (general) tests.' This 'uncharitable interpretation' was supported by the finding of Criper and Davies (1988) that the general (non-modular) part of the test correlated most highly with test total scores.

Despite Alderson's claim of broad-based support for the modular approach to EAP testing, as ELTS has changed to IELTS, and IELTS has undergone revision, the notion of specificity has been considerably changed. In IELTS there is now a general section, an academic module and a non-academic module. In other words, the five specific module format (the most specific EAP testing has ever attempted) has reverted to a single module format. The

only specificity comes in the provision of different reading passages and writing tasks in the second part of the IELTS, where test takers can choose from the three broad areas of: Business and Social Science (BSS), Life and Medical Science (LMS) and Physical Science and Technology (PST).

This trend to reduce specificity in EAP tests has come about precisely because of the lack of conclusive evidence to support specificity. Clapham (1993) summarizes research from around the world, much of which suggests that whilst background knowledge related to area of study does contribute something to score variance, language ability accounts for just as much, if not more, unique test variance. Of particular importance is the study of Tan in Malaysia, which demonstrated that language proficiency was the better predictor of text comprehension than subject specific background knowledge. Tan (1990: 222) argued that 'the weight of language proficiency is about twice the weight of subject specific background knowledge in the prediction of how well a reader can extract and interpret the meaning of a foreign language text'. Clapham (1993: 267) also concluded from her earlier studies that:

> the evidence from this study does not show the need for three academic subject modules in the test battery, and, second, if students are given academic modules outside their subject areas, they will not be placed at a disadvantage. If the forthcoming study bears out these initial findings, IELTS could, from an empirical point of view, satisfactorily offer just one academic module.

The most comprehensive study to tackle the issue of content specificity is that of Clapham (1996), in which she addresses a number of key research questions in relation to the impact of academic field of study, level of studies, subject specificity of reading passages, background knowledge and language proficiency, on test scores. This major study used the new format IELTS. Clapham's study has shown that it is difficult to tell whether or to what extent a text is 'specific', because this can only be judged in relation to the knowledge of the reader. However, when considering subject knowledge, Clapham (1996: 187) comes to the conclusion that:

> A multiple regression analysis of students' scores on the complete test module showed that the major contribution to their test scores appeared to be level of English proficiency. The students' field of study and their familiarity with the subject area were also significantly related to test scores, but much less strongly.

When considering the level of English proficiency needed to understand academic texts, Clapham (1996: 187) concludes that:

> results . . . showed that the students did not appear to be affected by background knowledge until they achieved scores of over 60% on the Grammar test. The results of the complete modules then showed that students with Grammar scores of over 80% appeared to make less use of background knowledge than the intermediate students. . . . It seems

> clear that the low proficiency students could not take advantage of their background knowledge, but from the results it is not clear whether as they became more proficient they were able to use linguistic skills to compensate for any lack in background knowledge.

This clearly echoes the views of Tan (1990), and shows that the empirical evidence that has been amassing since the early 1980s does point us to an inexorable conclusion; namely that whilst it is useful to have test content that you are familiar with, if the test is a test of English, and your English is not very good, then you are not going to do as well. Only at certain levels of proficiency does subject knowledge help compensate for lack of proficiency, and once proficiency advances again, subject knowledge becomes less important. Indeed, Clapham (1996: 193) reports that language proficiency accounted for 44 per cent of variance in her study, and the addition of background knowledge variables increased this by only 1 per cent to 45 per cent. When she revised the module to remove what was considered to be 'less specific' content, language proficiency accounted for 26 per cent of variance, which was increased by 12 per cent to 38 per cent by the addition of background knowledge variables. Subject familiarity was the most important of these, with knowledge of topic being insignificant. However, we must note that the revision of the modules was to remove 'less specific' sub-tests, and this was done on the basis of how much background knowledge-related variance the sub-tests generated. This approach was adopted because expert judges were incapable of deciding which texts were or were not subject specific. Indeed, it is stated that the decision for removing some of the sub-tests was taken on the basis of evidence from the Analysis of Variance results (Clapham 1986: 198–9). This process *appears* to be circular, but it is not. An empirical method has been used to increase variance that can be attributed to specificity; the problem lies in providing an applied linguistic theory to account for the texts that are left, *and* to use that applied linguistic theory to produce a new test that would demonstrate similar specificity variance, as argued in 'Evidential basis for test interpretation' above.

The empirical evidence to date, as reviewed here, suggests that:

- Language proficiency accounts for most of the variance in EAP test scores
- Subject knowledge can compensate to a small degree for lack of language proficiency, but this appears only to happen at intermediate levels
- What makes a text specific to a subject area cannot be defined by expert judges
- Increased specificity by module proliferation is unnecessary.

We must finally ask the question why, in the face of this evidence, EAP tests are still constructed on the basis of content analysis? This practice survives even though it is now acknowledged that: 'We have reasonably explicit test specifications, but they do not guarantee the content validity of the test' (Alderson 1993: 217). The answer would seem to be face validity, or how the

test looks to the test takers and the score users. Alderson (1993: 216) may provide the answer:

> from a face validity point of view, it is hard to justify a test of English for academic purposes that does not contain a test of reading, but many lay people, including admissions officers, might well argue that they do not care about a student's grammar provided that he or she can perform academic tasks (like reading and writing) satisfactorily in English.

Similarly, from Clapham (1996: 201):

> The texts used in the present TOEFL 'are taken from general reading materials rather than specialised textbooks.' Such texts, by definition, are not academic, and do not contain the variety of genres encountered in academic studies. For IELTS to include reading passages of this non-academic type would be to ignore the results of Weir's (1983) needs analysis and the IELTS content validation study, and would make any such test a test of general rather than university level reading. One of the strong points about the present is that the inclusion of academic texts encourages tutors to use study skills teaching methods when they are preparing students for the test. This gives students practice in the academic skills they will need in their further studies. The introduction of general non-academic texts might decrease this study skills element, and would also lessen the test's face validity.

This last quotation echoes Clapham (1993: 267), in which she wrote:

> the performance of the test is not the only criterion; the importance of face validity should never be underestimated. Students who were asked what they thought of the old ELTS test mostly liked it, and one of their reasons was the existence of the subject modules. . . . In addition, many universities and colleges particularly liked the ELTS test because it had subject-specific modules.

We have reached the stage where we can say that an EAP test should include prompts drawn from academic texts, or should have an 'academic setting' of some kind. But this is for the most part a matter of appearance. This is not, of course, to underestimate the importance of content, authenticity and face validity once they have been put in their place. Bachman and Palmer (1996: 23–4) rightly argue that the *perception of authenticity* encourages students to undertake the tasks to the best of their ability. If students do not take tests seriously their responses to test tasks are not likely to be adequate samples of their ability, which in turn threatens score meaning, and hence validity. This aspect of face validity, which may more appropriately be termed 'response validity' (Henning 1987: 92), should be taken seriously. Secondly, as Clapham (1996: 201) argued in the quotation above, the use of academic texts in EAP tests may have a beneficial washback effect upon teaching in EAP classes. This can be related directly to the consequential basis for test interpretation in Table 1. What is included (or excluded) from a test sends messages to test users about what is considered important, and what is considered to be

marginal. These messages are transmitted through the content of the test, what the test is called, and the labels attached to sub-tests.

There is therefore a realization that we are not testing *academic English*, but *English in an academic context*. With this realization the question of test content must be approached from a different perspective—a perspective in which content relevance and representativeness can be considered in the light of construct validity, rather than treating content validity as an aim and end in itself.

## CONCLUSIONS

Almost two decades ago, Widdowson (1983: 10) wrote:

> Instead of a theory we have an assumption that ESP is simply a matter of describing a particular area of language and then using this description as a course specification to impart to learners the necessary restricted competence to cope with this particular area.

EAP testing practice has also been slow to reverse the consequences of the lack of theory that has become associated with the focus on content. Widdowson (1983: 15) also claimed that:

> the work of ESP has suffered through too rigid an adherence to the principle of specificity of eventual purpose as determining criterion for course design. This has arisen, I suggest, because ESP has been removed from the context of language teaching pedagogy in general.

Similarly, in the field of EAP testing, we have seen that researchers are now questioning the focus on content that has caused some to lose sight of what is important in any assessment situation: that decisions made on the basis of test scores are fair, because the inferences from scores are reliable and valid.

What is the future of EAP testing? In a review of Clapham (1996), Davidson (1998: 292) comes to the conclusion that there are essentially two choices facing test developers. The first is to continue to produce subject specific tests or test modules 'and possibly come to the same inconclusive end result' that Clapham did, or to 'rely on general test-tasks and topics'. Douglas (1998: 117–18) envisages the same choices, and predicts that the future trend will be to use more general language tests unless there is unusually compelling evidence that specific content can be justified.

We have argued, however, that the face validity argument that emerges from the work of Alderson and Clapham needs to be taken seriously within a theoretical framework that values the consequential validity of EAP tests. On the one hand, test takers need to perceive the test as relevant to their subject and studies to achieve response validity. On the other, test content, title and labels of sub-tests may have significant washback effect upon what teachers do in classrooms. New tests may therefore continue to look very similar to their ancestors, but score meaning will be established in the light of construct validity studies rather than merely test content. However, unless future

research (such as that into performance testing) can provide new and measurable definitions of 'specific', it may no longer be appropriate to talk about tests of English *for* Academic Purposes, but rather of tests of English *through* Academic Contexts (EAC).

*(Revised version received September 1998)*

## REFERENCES


Alderson, J. C. 1981. 'Report of the discussion on communicative language testing' in J. C. Alderson and A. Hughes (eds.): *Issues in Language Testing.* London: The British Council 123–34.

Alderson, J. C. 1988a. 'Testing English for specific purposes—how specific can we get?' in A. Hughes (ed.): *Testing English for University Study. London.* London: Modern English Publications in association with the British Council. 16–28.

Alderson, J. C. 1988b. 'New procedures for validating proficiency tests of ESP? Theory and practice.' *Language Testing* 5: 220–32.

Alderson, J. C. 1993. 'The relationship between grammar and reading in an English for academic purposes test battery' in D. Douglas and C. Chapelle (eds.): *A New Decade of Language Testing Research.* Washington DC: TESOL Publications. 203–19.

Alderson, J. C. and A. H. Urquhart. 1983. 'The effect of student background discipline on comprehension: A pilot study' in A. Hughes and D. Porter (eds.): *Current Developments in Language Testing.* London: Academic Press. 121–8.

Alderson, J. C. and A. H. Urquhart. 1985a. 'The effect of students' academic discipline on their performance on ESP reading tests.' *Language Testing* 2: 192–204.

Alderson, J. C. and A. H. Urquhart. 1985b. 'This test is unfair: I'm not an economist' in P. C. Hauptman, R. Le Blanc and M. B. Wesche (eds.): *Second Language Performance Testing.* Ottawa: University of Ottawa Press.

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F. and A. S. Palmer. 1986. *Language Testing in Practice.* Oxford: Oxford University Press.

Carroll, B. 1980. 'Specifications for an English Language Testing Service' in J. C. Alderson and A. Hughes (eds.): *Issues in Language Testing.* ELT Documents 111. London: The British Council. 66–110.

Clapham, C. 1981. 'Reaction to the Carroll Paper 1' in J. C. Alderson and A. Hughes (eds.): *Issues in Language Testing.* ELT Documents 111. London: The British Council, 111–16.

Clapham, C. 1993. 'Is ESP testing justified?' in D. Douglas and C. Chapelle (eds.): *A New Decade of Language Testing Research.* Washington DC: Teachers of English to Speakers of Other Languages. 257–71.

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

Criper, C. 1981. 'Reaction to the Carroll paper 2' in J. C. Alderson and A. Hughes (eds.): *Issues in Language Testing.* ELT Documents 111. London: The British Council. 117–20.

Criper, C. and A. Davies. 1988. *English Language Testing Service Validation Project Report 1(i)* Cambridge: University of Cambridge Local Examinations Syndicate.

Davidson, F. 1998. Review of C. Clapham (1996) *The Development of IELTS. A Study of the Effect of Background Knowledge on Reading Comprehension.*' Cambridge: Cambridge University Press. *Language Testing* 15/2: 288–92.

Davies, A. 1965. *Proficiency in English as a Second Language.* Unpublished PhD thesis, University of Birmingham.

Douglas, D. 1998. 'Language for specific purposes testing' in C. Clapham, and D. Corson (eds.): *Encyclopedia of Language and Education, Volume 7: Language Testing and Assessment.* Amsterdam: Kluwer Academic Publishers. 111–19.

Ebel, R. L. 1983. 'The practical validation of tests of ability.' *Educational Measurement: Issues and Practice* 2/2: 7–10.

Fulcher, G. 1987. 'Tests of oral performance: The need for data-based criteria.' *English Language Teaching Journal* 41: 287–91.

**Henning, G.** 1987. *A Guide to Language Testing: Development—Evaluation—Research.* Cambridge, MA.: Newbury House.

**Henning, G.** 1992. 'Dimensionality and construct validity of language tests.' *Language Testing* 9/1: 1–11.

**Hughes, A., D. Porter, and C. Weir.** (eds.) 1988. 'English Language Testing Service ELTS Validation Project: Proceedings of a conference held to consider the ELTS validation project report.' Cambridge: University of Cambridge Local Examinations Syndicate.

**Hutchinson, T. and A. Waters.** 1987. *English for Specific Purposes: A Learning-centred Approach.* Cambridge: Cambridge University Press.

**Loevinger, J.** 1957. 'Objective tests as instruments of psychological theory.' *Psychological Reports* 3: 635–94.

**McDonough, J.** 1984. *ESP in Perspective: A Practical Guide.* London: Collins ELT.

**Messick, S.** 1989. 'Validity' in R. L. Linn (ed.): *Educational Measurement.* New York: American Council on Education/Macmillan. 13–103.

**Munby, J. L.** 1978. *Communicative Syllabus Design.* Cambridge: Cambridge University Press.

**Stevenson, D. K.** 1985. 'Pop validity and performance testing' in Y. P. Lee, A. C. Y. Y. Fok, R. Lord, and G. Low (eds.): *New Directions in Language Testing.* Oxford: Pergamon Institute of English.

**Tan, S. H.** 1990. 'The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates' in J. H. A. L. de Jong, and D. K. Stevenson. (eds.): *Individualizing the Assessment of Language Abilities.* Clevedon, Avon: Multilingual Matters, 214–24.

**Underhill. N.** 1987. *Testing Spoken Language: A Handbook of Oral Testing Techniques.* Cambridge: Cambridge University Press.

**Weir, C.** 1983. *Identifying Language Problems of Overseas Students in Tertiary Education in the United Kingdom.* Unpublished PhD thesis, University of London.

**Weir, C.** 1988. 'The specification, realization and validation of an English Language Proficiency Test' in A. Hughes (ed.): *Testing English for University Study.* ELT Documents 127. London: Modern English Publications in association with the British Council. 45–110.

**Widdowson, H.** 1983. *Learning Purpose and Language Use.* Oxford: Oxford University Press.

**Wiliam, D.** 1997. 'Construct-referenced standard setting of achievement and placement tests.' Paper delivered at the 1997 Language Testing Forum, University of Surrey, 20–21 November.

**Wood, R.** 1993. *Assessment and Testing.* Cambridge: Cambridge University Press.