Language for Specific Purposes: Current and future issues

Tineke Brunfaut

Abstract

This paper presents a number of issues on the topic of Language for Specific Purposes testing that were raised during a plenary discussion at the 30th annual Language Testing Forum. The comments particularly focused on (1) past and current conceptualizations and categorizations of LSP tests, (2) tensions between specificity and practicality in LSP test design, and (3) the role of locality in LSP testing. The views exchanged on each of these themes are reported and considered in light of current research and debates. Suggestions are made for future research in the area of LSP testing.

1 Introduction

As explained in the introduction of this special issue, one of the three themes of the first Language Testing Forum (LTF) in 1980 concerned 'Testing of English for Specific Purposes' (ESP). The discussion held at that time concentrated on a paper by Brendan J. Carroll to which others reacted. The full collection of 1980 papers can be found in Alderson and Hughes' 1981 *Issues in Language Testing* (available at http://www.ling.lancs.ac.uk/groups/ltrg/ltf2010.htm). In his paper, Carroll (1981) describes the changing needs of the English Proficiency Test Battery (EPTB; used for British study entry screening), due to increases in test-taker numbers and the move towards communicative approaches in language learning and teaching. His call for the diversification of the test battery was driven by a needs analysis which led to a set of draft specifications for the ‘successor’ of the EPTB, i.e. the English Language Testing Service (ELTS). Reactions to Carroll’s paper, amongst others, expressed concern over the generalizability of (small-scale) needs analysis results (Clapham, 1981), the practicality of discipline-specific tests (Criper, 1981), and score interpretation and extrapolation issues of specific purposes tests (Alderson, 1981). The need for discipline-specific tests in the EPTB academic entry context was even questioned (Alderson, 1981), and research on this topic was urged for (which was later conducted in studies such as Clapham, 1996).

Three decades after the first Language Testing Forum the ESP theme was revisited during a symposium at the 2010 edition of the LTF. The scope of this symposium was broadened from *English* for Specific Purposes testing to *Language* for Specific Purposes testing, to recognise and reflect the widened scale of testing for specific purposes. An academic review of the theme by Barry O’Sullivan (University of Roehampton) and an exam board view by Henry Emery (*emeryroberts* Aviation English Training; see Emery this issue) were followed by a plenary discussion on the topic of Language for Specific Purposes (LSP) testing.

The aim of the present paper is to record and discuss a number of issues that were raised during this plenary discussion; it does not intend to provide a comprehensive review of issues in LSP testing. The plenary comments particularly focused on (1) past and current conceptualizations and categorizations of LSP tests, (2) tensions between specificity and practicality in LSP test design, and (3) the role of locality in LSP testing. Each of these themes will be explored below, thereby reporting on the views exchanged during the 2010 LTF discussion and considering these in light of current research and debates. Suggestions for future research will also be made.

2 Conceptualizations of Language for Specific Purposes Testing

A discourse of boundaries

A common theme running through the LSP testing literature, which was echoed in the LTF discussion, concerns the struggle to conceptualise LSP. A number of observations can be made in this regard.

Often, the discourse tends to evolve around the terms *language* and *content*, and their separability. Earlier sources in particular suggested that background knowledge needs to be controlled for, and thereby seemed to imply that specific purposes language knowledge and content are clearly distinguishable (e.g. Criper, 1981). Also, although it appears to be generally accepted that specific purpose background knowledge is part-and-parcel of the construct underlying LSP tests, it has been argued that the test situation may determine the need to distinguish between language and content knowledge (Douglas, 2000). More recent research, however, has shown the (insurmountable) challenges in neatly separating language from specific purposes background knowledge, and has highlighted the fuzzy boundaries between the two (e.g. Elder, 2001).

A similar comparative or contrastive thread runs through debates on the *S* in LSP (i.e., specificity), and mostly centres on the relationship between Language for Specific Purposes versus General Language Proficiency (GLP) testing. Often, tests are visually depicted as being positioned somewhere on a continuum with the opposing extremes being labelled *specific* and *general*. Hamp-Lyons and Lumley (2001) stated that a key distinguishing element between LSP and GLP tests is that the former requires a much more comprehensive, fine-grained and context-specific needs analysis. Davies (2001), on the other hand, has argued that it is not possible to draw a clear-cut line between LSP and GLP tests. The reason for this, as O’Sullivan (2012) explains, is that the language used in a specific domain does not operate as a universe detached from language use in the general domain, but interacts with it. Similarly, language use in one domain is never completely unconnected from language use in another domain. In fact, Davies (2001) has posited that content rather than language is the distinguishing feature of an LSP test task.

During the LTF discussion, the observation was made that issues surrounding LSP testing, such as the relationship between language and content, equally apply to GLP testing. O’Sullivan endorsed this viewpoint by putting forward Communicative Language Testing (CLT) as a form of LSP testing. He argued that, similar to LSP testing, communicative language testing is testing language for use in a particular domain, in this case communication. Douglas (2000), however, coined the relationship the other way around, and depicted LSP as a ‘special case’ of CLT (p.9). Leaving aside the exact nature of the relationship, both conceptions emphasise commonalities.

Interestingly, the ‘boundary discourses’ (i.e. discussions to conceptualise something which are characterised by and evolving around a discourse of separability and distinguishability) are reflected in the growing body of literature and studies on integrated forms of language assessment (which also happen to be adopted mostly in LSP tests). The discussion in this case focuses on the separability of different language skills, for example, reading and writing in a reading-to-write task. Remarkably, a considerable amount of empirical research on integrated tasks (e.g., Chan, 2011; Cumming et al., 2006; Delaney, 2008; Gebril, 2010; Plakans, 2008; Weigle, 2004) has specifically contrasted performances on integrated tasks with performances on independent tasks (for example, reading-to-write versus writing-only tasks), rather than investigating the integrated construct as such (for example, reading-to-write). The underlying assumption seems to be that integrated tasks start off from independent skills, and the discourse used to describe findings seems to suggest it is desirable (and possible) to compare and separate individual skills in performances on integrated tasks.

It thus seems that, conceptually, there is a struggle with integrated concepts (whether it is language and content, different language skills, or general and specific language use), and discourses tend to revert to contrastive approaches and boundary discussions. It has become clear from research such as Elder (2001), however, that these may not be the most productive discussions; as O'Sullivan (2012) pointed out: "there are no "exact limits" " (p.73).

Boundary conceptualizations as operationalized in scales

An important part of the theoretical and empirical literature on LSP testing focuses on issues related to one 'side' of tests, i.e. the device that elicits language knowledge and use in a particular domain. Some of the issues covered are task authenticity, needs analyses and their impact on task design, and the elicitation of language and content knowledge in a manner similar to their interaction in the target language use situation.

Far less research seems to be available on another crucial aspect of tests, i.e. the means by which the elicited domain language use and knowledge is evaluated. Despite Douglas' plea (2001) for 'indigenous' assessment criteria which are derived from a target language use analysis, it is not uncommon for scales used in LSP assessment to establish a split between a range of generic language criteria such as lexical and grammatical knowledge or organization, and a task achievement criterion which is set out to represent the domain aspect.

The lack of domain embedding of scales and their usage was illustrated in the LTF discussion by means of the International Civil Aviation Organization (ICAO) rating scale. This scale distinguishes 6 rating criteria (pronunciation, structure, vocabulary, fluency, comprehension, and interactions) and describes 6 language proficiency levels (ranging from pre-elementary to expert). John de Jong argued that the ICAO rating scale, which was designed in a period in which many official bodies (e.g. the Council of Europe, NATO) were developing language proficiency scales, copied the type of scale that was being constructed by others at the time. In this process, the differences in testing purposes had been overlooked. He claimed that, as a consequence, ICAO adopted the wrong model. Unlike general proficiency tests which aim at establishing a test taker's language proficiency level, ICAO's primary goal for assessing pilots' and air traffic controllers' language ability is to determine whether they have sufficient language skills to operate safely. Therefore, what matters is the cut-off, the minimum requirement in terms of language proficiency (in practice, this is defined as level 4 of the ICAO scale, called 'Operational'). He pointed out that whether or not a pilot or air traffic controller is more proficient than that minimum (i.e., levels 5 and 6 of the ICAO scale – 'Extended' and 'Expert') is irrelevant in this context. Members of the LTF audience active in aviation language testing agreed that mastery or non-mastery of the minimum language requirement is the crucial information. However, they explained that the extent to which a pilot's language proficiency exceeds the minimum determines the time frame for re-assessment and thus further proficiency specification does meet the aviation world's practice of recurrent testing. That is, ICAO recommends that pilots who demonstrate level 4 (Operational) or level 5 (Extended) language proficiency are re-tested every 3 and 6 years respectively. Nevertheless, attention was drawn to the fact that the minimum language requirement and re-testing policy did not apply to native speaker pilots. This was questioned for the field of domain-specific aviation English testing, whereby aviation subject-matter knowledge is an integral part of the construct. It was argued that it should not be taken for granted that native speakers meet the minimum requirement as related to this LSP construct, and they should thus also be assessed. In fact, it was pointed out that in a few cases in which native speakers had not used the prescribed, formulaic language, but reverted to more colloquial language, this had played a role in aviation incidents. Through its test and scale usage, it thus seems that ICAO treats content and language as two separate constructs.[1] This in turn raises questions regarding the LSP nature of aviation English testing. As O'Sullivan stressed during the LTF plenary discussion, the domain is entangled in the construct: if an LSP test does not include aspects of the domain such as subject-matter knowledge, it cannot be considered a s*pecific* test. He stated that a domain-specific test essentially is *of* the domain. As a consequence, a solid link with the domain is a prerequisite for the validity of an LSP test, and such links need to be made explicit.

Conceptualizations of the link between a specific domain and an LSP test

Emery's exam board view (see Emery, this issue) demonstrated the link with the domain *in practice*. What seem to be missing are comprehensive *theoretical* justifications for such links. In fact, as was observed by the LTF academic review of LSP testing, the focus of attention has to a large extent been on documenting and addressing practical issues (e.g., Green & Wall, 2005; Lockwood, 2012a), and less on developing a coherent theory of LSP assessment (but see Douglas' (2000) contributions). This prompted the LTF audience to turn to the need for a bridge between language testing theory and practice. Reference was made to language testing history to make clear how weak that bridge had been at the time of the first Language Testing Forum in 1981. Only one of the attendants at the time was employed outside of academia. Many of the tests (e.g. ELTS and ELBA) were 'one-man tests', conceptualised and

[1] Note that as opposed to the ICAO Rating Scale descriptors which do not explicitly mention the domain, reference is made to the aviation domain in the ICAO Holistic Descriptors. For example, it is stipulated that "proficient speakers shall use a dialect or accent which is intelligible to the aeronautical community" (*ICAO Doc 9835 Appendix A ICAO SARPs referring to Annex 1 Personnel Licensing 1.2.9*).

constructed by very small teams (or one main developer; often linguists employed in academia). Since then, however, a significant body of responsibilities for testing have moved to large organizations. It was argued that the link between theory and practice is now stronger than it was 30 years ago. However, it was also noted that practicalities related to the take-over of many smaller examination boards by large organizations may have put validity somewhat to the side, whereas test validity had been the basic concern of those who attended the first Language Testing Forum. Therefore, it was thought to be important to emphasise the centrality of validity, in particular predictive validity, at the 30th anniversary of the Language Testing Forum. O'Sullivan's symposium contribution, which initiated a theoretical model of LSP testing, was therefore welcomed. Its development responds to the call for a rigorous theoretical underpinning of a more practicality-driven focus in LSP test validation. It also pursues a validation approach which forefronts the *S* of LSP by putting the specific target language domain at the heart of the approach. In this regard, it was noted that an LSP validation model that captures the link with the domain does not only require a definition of the domain, but also incorporates test purpose and test population as fundamental elements. Adopting this view, theoretical justifications for links with the domain necessitate specifications of the domain but also of the people within the domain. These core features put forward at the LTF have remained central to further developments of a theoretical model of LSP testing, presented in O'Sullivan (2012).

3 Specificity and practicality

Exam board representatives at the LTF (e.g. from Cambridge ESOL) acknowledged and fully supported the centrality of validity. Nevertheless, it was pointed out that, from an exam board’s perspective, practicality is an important factor which needs to be taken into account at all times.

Emery illustrated the impact of local practicality with an example from his workplace. Aiming for authenticity, the test developers had considered conducting an aviation English test in a flight simulator. This was also thought to have higher face validity and would potentially be more motivating for the test candidates, since pilots’ main goal is to fly an aircraft. However, the high operational costs of simulators made this unaffordable, and thus the test had to be 'taken back into the traditional language test room'. This means that a test administrator and a pilot spend a considerable amount of time in a classroom for test-taking purposes, away from flight technology and from what Emery termed 'perhaps the rightful place for language behaviour'. An implication is that the domain may only be partly present in the tests. On the other hand, in other respects the local practicality of the aviation world could facilitate test administration. For example, close monitoring (e.g. health check-ups) and recurrent testing and training of pilots’ professional skills (e.g. by means of six-monthly flight simulator exercises) is commonplace – regardless of years of experience. Aviation English testing could therefore be slotted into the established programme of monitoring (in terms of its timing) and of recurrent training and testing in the professional domain (timewise and contentwise).

Practicality is often set off against specificity; especially when a particular type of LSP is conceptualised as being on a general-specific continuum. When moving towards the specific end on the continuum, tests become more and more individualised, to the point that language testing is about the individualization of the test experience for one person. In his book *Assessing Languages for Specific Purposes*, published in 2000, Douglas observed that a number of LSP tests seemed to have shifted to or are situated at the general end of the spectrum (for examples, see Davies (2008) on the IELTS solution to assessing academic

English, or Taylor and Angelis (2008) on a pre-2000 history of TOEFL). Hamp-Lyons and Lumley (2001) explained this strategy as exam boards' response to score consumers who want "simple 'information' about the language performance of a wide range of people" (p.130). They questioned the validity of such decisions, particularly because test use and consequences are an integral part of more recent views on validity (e.g. Chapelle, 1999). Hamp-Lyons and Lumley felt, however, that they lacked a comprehensive theory to back this up.

A view expressed during the LTF discussion was that exam boards are still faced with the problem of having to develop a generalizable test out of an individual test experience, and thus they need to find a balance between the generalizable and the specific. Therefore, it was argued that, rather than viewing practicality as an aside or an after-thought, it should be central to language test development from the start of the process and at the same time be integrated in theory. It was mentioned that in the mid-1980s language testers at the University of Reading started thinking about theorising practicality and building-in practicality as a sensible issue in language testing. Reference was also made to the Cambridge VRIP approach to test validation (see also Taylor, this issue), which evaluates tests in terms of validity, reliability, impact *and* practicality. It was argued that such an approach – with practicality as one of the central concepts – is particularly relevant to LSP testing. The importance of practicality is also marked in Bachman and Palmer's (1996) definition of test usefulness in which practicality is described as a quality of usefulness, in addition to reliability, validity, authenticity, interactiveness, and impact.

4 Specificity and localization

The discussion on practicality and generalizability prompted O'Sullivan to raise the issue of localization in language testing. One of his observations on the period 1980-2010 has been the increasing fragmentation of language testing. At the time of the first Language Testing Forum, language testing was mainly governed by two monoliths, the US and the UK. However, the international audience present at the 30th anniversary of the Forum exemplifies the spread of language testing expertise. Similarly, the focus of the 2010 symposium on *Language* for Specific Purposes, instead of the *English* for Specific Purposes theme at the first Language Testing Forum, testifies to the broadening of the field. O'Sullivan acknowledged that organizations such as the Association of Language Testers in Europe (ALTE), the European Association for Language Testing (EALTA) and the International Language Testing Association (ILTA) have assisted in the professionalization and proliferation of expertise in language testing, and he argued that at the same time as increasing fragmentation and professionalization, the field has become more aware of the importance of the localization of language tests. Localization, he clarified, should not be understood as local to a particular geographical space. Instead, localization concerns putting the test taker at the heart of the test development process; it is about "tak[ing] into account characteristics (individual, linguistic, cultural and social) of the learners from a particular population when developing tests for use with that population" (O'Sullivan, 2011, p.6). Also embedded in this concept is the centrality of text context to test development (O'Sullivan, 2011). Usually, needs analyses documenting individuals' language operations within a particular domain will feed into LSP test development. Therefore, not taking into account the domain and the individual within the domain will mean that the test taker will not fully partake in the specific domain – i.e., language and context (O'Sullivan, 2012).

Language tests, whether they are large-scale international tests or small-scale tests, have been designed for a particular purpose and a particular population. That is, they have a particular usage that is local to a community of users (in the broad sense of the word; for example including test takers, target domain language users and score users). Increasingly, testing is done on a smaller scale, at a local level, and thus with different communities. As a consequence, the importance of understanding the needs of different communities of users has become crucial. Testing in a specific domain now means that one takes into account the domain, the context, the culture – the locality. It means that one does not merely pick any existing test for use in another context, for a different purpose, or with another population. At the LTF, O'Sullivan identified the bearing in mind of locality as one of the underarticulated, but key changes over the last three decades. He emphasised that this localization is a prerequisite for language testing, and later called it “the essential expectation of any LSP test” (O'Sullivan, 2012, p.79).

Aspects of the concept of localization have also been voiced by others, without necessarily using this terminology, being as comprehensive, or emphasizing the centrality of the test taker to the same extent. For example, Douglas (2001) discusses the virtue of grounded ethnography, i.e. “an approach to describing and understanding a TLU [Target Language Use] situation from the perspective of language users in that situation” (p.181), and Jacoby and McNamara (1999) stated that “special purpose performance is by definition task-related, context-related, specific, and local” (p.234).

At the LTF, the concept was further discussed through an example. In the weeks preceding the LTF gathering, someone had asked in an applied linguistics forum what was meant by ‘academic’ in Language for Academic Purposes (LAP) teaching and testing. The forum reply had been that one needs to look at the countries where the students or test-takers come from and identify what is meant by ‘academic’ in those contexts, in order to bridge the gap between academic practice in the country of origin and the target country. A member of the LTF audience was concerned about the implications of such a view. Does it imply that a language test is supposed to tell everything one would like to know – ranging from whether students can write from sources, over plagiarism issues, to whether they can give an academic presentation? Is too much being asked of language testers?[2] In response to this concern, the concept of localization was referred to, and it was stated that where test takers come from cannot be the language tester's problem. Where they come from also does not matter. It was argued that it is where test takers are going to that counts: an LSP test has to "[engage] the test taker in both the language and the context of [the] specific domain" (O'Sullivan, 2012, p. 79). So, in the case of LAP testing for study in the UK, it is the UK notion of 'academic' that counts. What matters is the locality of the community of users.

The practical implications of such a view call for a user-oriented test development process, with user input at all stages; for example, during needs analyses, task development or scale development (for examples of the latter, see Abdul Raof, 2011; Abdul Raof, Hamzah, Aziz, Omar & Atan, 2011). It also suggests user involvement at the operationalization stage to fully ensure test-taker domain engagement; for example, during test administration or scoring (e.g. in a (co-)interlocutor or rater role). Currently, this does not appear to be common practice (but see Lockwood, 2012b). In terms of theory, an emphasis on localization implies a prominent role in LSP models for the test taker and for test usage, all in relation to the specific domain.

5 Conclusions – A Research Agenda for the Future

[2] For similar concerns, see Diane Schmitt's opinion piece in The Guardian Weekly of 13 November 2012, entitled 'UK universities failing to bridge culture gap for foreign students', http://www.guardian.co.uk/education/2012/nov/13/international-student-testing-culture-gap

It is clear from the above report and discussion that a current concern is the need for a coherent theory on LSP assessment. In addition, the need for a link between LSP theory and practice has been acknowledged. During the LTF symposium, the impractical nature of most validation models was pointed out (for a similar view see O'Sullivan & Weir, 2011), and, as O'Sullivan (2012) phrased it, there is a need for a theory for LSP test validation "which could then be used to drive both the development and validation process as well as forming the basis for a meaningful research agenda" (p.80).

Douglas' extensive work has been of great significance to theorizing LSP assessment (e.g. Douglas, 2000), and recent work by O'Sullivan (2012) forms a promising contribution. O'Sullivan's theoretical model of a Languages for Specific Purposes System (2012, p.80), which has its roots in Weir (2005), draws attention to the test system (performance parameters and linguistic demands of the task), the test taker (individual and cognitive characteristics of the test taker), and the scoring system (theoretical fit, and accuracy and value of decisions). The distinguishing features of this model, as viewed by O'Sullivan (2012), are (1) the prominent role assigned to the test taker, and (2) the impact of the target language domain on all three elements (i.e., test system, test taker, and scoring system). In this manner, it aims to integrate the concept of localization in LSP modelling.

A logical suggestion for future research would therefore be to look into the success with which the concept of localization has been adopted by this particular model. At the same time, it remains to be investigated how well the model as a whole works in real terms. Although O'Sullivan (2012) points out that Weir's original model (2005) has shown to be of some use in LSP validation projects even though it was not specifically developed for the context of LSP testing, to the author's knowledge the model has not yet been applied to an LSP test validation study in its current form. Furthermore, while O'Sullivan claims to acknowledge the issue of consequences and their impact on all aspects of his model, this understanding does not appear to be visually implemented in the schematic representation of the model (O'Sullivan, 2012, p.80). As a consequence, it may risk being (partly) overlooked by users of the model.

Given the foregrounding of the test taker in O'Sullivan's LSP System (2012), a more detailed research agenda could focus on the cognitive processes activated during LSP assessment (and the way in which the target language domain is embodied in these). A particular issue in this respect concerns the methods with which to study the cognitive processes. Although techniques such as expert judgements or verbal protocols have their merits, there are also several problems associated with these methods (see e.g. Alderson (1993) and Alderson, Brunfaut, McCray & Nieminen (2012) for issues on expert agreement; and see e.g. Afflerbach & Johnston (1984) and Green (1998) for drawbacks of verbal protocols). It has recently been suggested that eye-tracking could increase insights into cognitive processing during test completion, and that the technique may be particularly valuable in combination with other methods such as interviews (McCray, Alderson & Brunfaut, 2012). However, explorations of this technique in the field of language testing are still limited (e.g., Bax & Weir, 2012; Gorin, 2006; McCray, Alderson & Brunfaut, 2012; Taylor, 2011), and would definitely benefit from a larger body of studies using the method for the testing of a range of skills in order to validate its use.[3]

A different area in need of further investigation relates to the point made in the discussion above; namely, the relative scarcity of studies on scoring-related issues in LSP testing. The importance of scoring validity has not only been recognised in traditional validation approaches (e.g. Alderson, Clapham & Wall, 1995), but the scoring system also constitutes an important component of O'Sullivan's (2012) LSP System. Therefore, in

[3] So far, most research has focused on cognitive processing during reading test completion.

addition to more extensive research on domain representations in scoring criteria, studies on user-oriented LSP test operationalization (e.g. domain professionals as raters) may help ensure domain representations in LSP assessment, in general, and the scoring system, in particular.

References

Abdul Raof, A. H. (2011). An alternative approach to rating scale development. In B. O'Sullivan (Ed.), *Language testing theories and practices* (pp. 151-163). Basingstoke: Palgrave Macmillan.

Abdul Raof, A. H., Hamzah, M., Aziz, A. A., Omar, N. A. M., & Atan, A. (2011). Profiling graduating students' workplace oral communicative competence. In Powell-Davies, P. (Ed.), *New directions: Assessment and evaluation*. British Council. Retrieved from http://www.britishcouncil.org/download-accessenglish-publications-ebe-proceedings-2012.pdf

Afflerbach, P., & Johnston, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behaviour*, 16, 307-321.

Alderson, J. C. (1981). Report on the discussion on testing English for Specific Purposes. In J. C. Alderson & A. Hughes (Eds.), *Issues in language testing* (pp. 123-134). London: The British Council.

Alderson, J. C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing* (pp. 46-57). Washington DC: TESOL.

Alderson, J. C., & Hughes, A. (1981). *Issues in language testing.* London: The British Council.

Alderson, J. C., Brunfaut, T., McCray, G., & Nieminen, L. (2012). *Component-skills approach to L2 reading: Findings, challenges, and innovations*. Paper presented at the American Association for Applied Linguistics (AAAL) conference, Boston, USA.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bax, S., & Weir, C. J. (2012). Investigating learners' cognitive processes during a computer-based CAE reading test. *Cambridge ESOL: Research Notes, 47*, 3-14.

Carroll, J. B. (1981). Specifications for an English Language Testing Service. In J. C. Alderson and A. Hughes (Eds.), *Issues in language testing* (pp. 66-110). London: The British Council.

Chan, S. (2011). *Demonstrating cognitive validity and face validity of PTE Academic writing items: Summarize written text and write essay*. Retrieved from http://www.pearsonpte.com/research/Documents/RN_DemonstratingCognitiveAndFaceValidityOfPTEAcademicWritingItems_2011.pdf

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272.

Clapham, C. (1981). Reaction to the Carroll paper (1). In J. C. Alderson & A. Hughes (Eds.), *Issues in language testing* (pp. 111-116). London: The British Council.

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 & A. Hughes (Eds.), *Issues in language testing* (pp. 117-120). London: The British Council.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL test.* TOEFL Monograph No. MS-30. Princeton, NJ: Educational Testing Service.

Davies, A. (2001). The logic of testing Languages for Specific Purposes. *Language Testing, 18*, 133-147.

Davies, A. (2008). *Assessing academic English: Testing English proficiency, 1950–1989 — the IELTS solution.* Cambridge: Cambridge University Press.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7*, 140-150.

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

Douglas, D. (2001). Language for Specific Purposes assessment criteria: Where do they come from? *Language Testing, 18*, 171-185.

Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing, 18*, 149-170.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15,* 100-117.

Gorin, J. S. (2006). *Using alternative data sources to inform item difficulty modelling*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge: Cambridge University Press.

Green, R., & Wall, D. (2005). Language testing in the military: problems, politics and progress. *Language Testing, 22,* 379-398.

Hamp-Lyons, L., & Lumley, T. (2001). Editorial. Assessing Language for Specific Purposes. *Language Testing, 18*, 127-132.

Jacoby, S., & Mcnamara, T. F. (1999). Locating competence. *English for Specific Purposes Journal, 18*, 213-241.

Lockwood, J. (2012a). Are we getting the right people for the job? A study of English language recruitment assessment practices in the business processing outsourcing sector: India and the Philippines. *Journal of Business Communication, 49*, 107-127.

Lockwood, J. (2012b). English language assessment for the Business Processing Outsourcing (BPO) industry: business needs meet communication needs. *School of Doctoral Studies (European Union) Journal, 4*, 187-198. Retrieved from http://www.iiuedu.eu/press/journals/sds/SDS-2012/SSc_Article4.pdf

McCray, G., Alderson, J. C., & Brunfaut, T. (2012). *Combining eye-tracking with post test interview data to examine gapfill items: triangulation or tribulation?* Paper presented at the annual conference of the European Association for Language Testing and Assessment, Innsbruck, Austria. Retrieved from http://www.ealta.eu.org/conference/2012/presentations/McCray.pdf

O’Sullivan, B. (2011). Introduction. In B. O’Sullivan (Ed.), *Language testing theories and practices* (pp. 1-12). Basingstoke: Palgrave Macmillan.

O’Sullivan, B. (2012). Assessment issues in Languages for Specific Purposes. *The Modern Language Journal, 96*, 71-88.

O’Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O’Sullivan (Ed.), *Language testing theories and practices* (pp. 13-32). Basingstoke: Palgrave Macmillan.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13,* 111-129.

Taylor, C. A., & Angelis, A. (2008). The evolution of the TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 27-54). New York: Taylor and Francis.

Taylor, L. (2011). *Using eye-tracking technology to research L2 reading.* Paper presented at the BAAL Testing, Assessment and Evaluation SIG, University of Warwick, UK.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*, 27-55.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.