**THE JOURNAL OF TEACHING ENGLISH FOR SPECIFIC AND ACADEMIC PURPOSES**
Vol. 7, N$^{o}$ 2, 2019, pp. 159−169

UDC: 811.111'255.4:159.9 https://doi.org/10.22190/JTESAP1902159L

# PURSUING AUTHENTICITY IN ESP TESTING – THE NEED FOR INTERDISCIPLINARY COLLABORATION

Lan Luo

Zhejiang University, China
E-Mail: laura_luolan@yahoo.com

**Abstract**. *Authenticity is the primary factor affecting test validity in ESP (English for Specific Purposes) test. While ensuring the authenticity of test tasks, pursuing authenticity in assessment criteria has received more and more attention. Studies have shown that experts in the professional field use different assessment criteria when assessing candidates' communicative competence in a particular professional context. It is certainly the case that the construct of communicative competence informing practice in language testing is different from the views of communication informing the communication literature in the professional setting, and hence the views of educators in that field. Rapprochement between these two perspectives is clearly desirable. This paper reviews the history and development of ESP testing, emphasizes the necessity of balancing the different scoring views between linguists and professional experts from the perspective of EMP (English for Medical Purposes) oral test, and discusses the implication of pursuing authenticity in ESP testing as well as ESP teaching.*

**Key words**: *ESP testing, authenticity, assessment criteria*

## 1. INTRODUCTION

Recently, ESP education has received more and more attention. It is different from English for General Purposes (EGP) in many aspects, but the need for separate ESP testing is controversial. Supporters like Douglas (2000) believed that the main purpose of LSP tests is to evaluate the test taker's specialized language ability, but their background knowledge is a necessary and an inseparable part of their specialized language ability. Therefore, authenticity of tasks and interaction between language knowledge and specific purpose content knowledge are two important characteristics of ESP testing that may be applied to distinguish ESP from EGP testing. Authenticity is an important factor affecting the validity of a test, and has always been a hot issue in the language testing field. Defining authenticity is critical to rating, test development, test use, etc. While ensuring the authenticity of test tasks, pursuing authenticity in assessment criteria used to judge test-taker performance has also became an important consideration in evaluating test authenticity.

Submitted March 18th, 2019, accepted for publication April 21st, 2019

## 2. ESP TESTING

### 2.1. The definition and purpose of ESP testing

The area of ESP testing has aroused considerable controversy. Many scholars doubt the necessity to separate Language for Specific Purposes (LSP) test with Language for General Purposes test (Widdowson 1981; Cumming 2001; Davis 2001). ESP testing has been criticized on a number of grounds: 1) There is no obvious boundary between EGP and ESP, ESP is really just EGP with technical vocabulary thrown in; 2) ESP tests are unnecessary, ESP can be included in EGP tests; 3) ESP tests are unreliable and invalid since subject knowledge interferes with the measurement of English knowledge; 4) ESP testing has no theoretical justification; 5) ESP testing has no predictive ability, there's no large advantage compared with EGP testing (Douglas 2000).

But on the other hand, many scholars have a favorable attitude toward ESP testing, and one of the leading figures is Dan Douglas. He proposed a very precise definition of specific purpose language testing in his book *Assessing Languages for Specific Purposes*:

> *A specific purpose language test is one in which the test content and test methods are derived from an analysis of a specific language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purposes content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain* (Douglas, 2000:19).

He believed that language for specific purposes differs from language for general purposes in many aspects and it is necessary to test them separately. The two most important reasons he provided are as follow:

Reason 1: language performances vary with context

A test taker's language performances vary with both context and test task, therefore our interpretations of a test taker's language ability must vary from performance to performance. For example, if we give test takers a reading test based on a passage about medical science, followed by one based on a passage about computer science, the test takers will probably perform somewhat differently on the two tests, particularly if there are test takers majoring in one of the subjects. Therefore, in order to measure and interpret the test takers' language skills accurately, the material the test in based on must engage test takers in a task in which both language ability and knowledge of the field interact with the test content in a way which is similar to the target language use situation. And ESP testing requires the use of field specific content in tasks which might plausibly be carried out in those fields.

Reason 2: specific purpose language is precise

Technical languages that are used in any academic, professional or vocational field, have specific characteristics that people who work in the field must control. There are lexical, semantic, syntactic, and even phonological characteristics of language peculiar to any field, and these characteristics allow for people in that field to speak and write more precisely about aspects of the field that outsiders sometimes find impenetrable. And this is also one of the most important reasons why LSP (Language for Specific Purposes) and LSP testing exist. A classic example of the need for precise, specific purpose language comes from the field of law. If we wanted to measure a lawyer's control of English to conduct the business of law, it would not seem to be sufficient to use texts and tasks which were not specific to the legal profession. Thus, if our goal is to measure a test taker's ability to use language within a specific vocation, profession or academic field, then specific purpose texts and tasks will be needed.

### 2.2. The characteristics of ESP testing

According to Douglas (2000), authenticity of tasks and interaction between language knowledge and specific purpose content knowledge are two important characteristics of ESP testing that may be applied to distinguish ESP from EGP testing.

a. Authenticity

In the 1970s, with the emergence of communicative approach, Widdowson (1978) raised the question of authenticity in language teaching. While communicative language testing became mature and the main stream by the end of the 20th century, authenticity has also become one of the focuses in language testing (Bachman, 1990; Morrow, 1991; Bachman & Palmer, 1996).

Authenticity can be defined from the following aspects:

(1) real-life authenticity

Real-life authenticity means that tests should be designed to reflect real life situations, or in other words, the task characteristics should be consistent with the features of the real-life situations in which the target language is used (Bachman, 1990).

(2) interactional authenticity

Communicative authenticity emphasizes the interaction between the candidate and the test task. The greater the degree of interaction between the test taker and the test task, the higher the degree of authenticity of the test. Therefore, the test authenticity can be defined as interactive authenticity, that is, the degree to which it invokes the test takers' language ability (Bachman, 1991).

(3) correspondence approach (CA)

The concept of CA was put forth by Bachman and Palmer (1996). They defined authenticity as the degree of consistency between the characteristics of a particular test task and the features of a TLU (target language use) task. They believed that this approach would provide a more practical way of taking authenticity into account in the design and the development process of a language test.

LSP test takers are generally the target users in the actual TLU domain. The main purpose of the LSP test is to determine and predict the test takers' specialized language skills in target language scenarios. The validity of the test depends largely on the authenticity of the test, and the degree of consistency between the content of the test and the content of the TLU domain. Therefore, the test content and the test method should be based on the analysis of the LSP scenarios, so that the tasks and content can truly reflect the actual language used in the target situation.

b. Specificity

In LSP tests, the test takers' performance will depend on both their professional knowledge and language skills. In other words, the test takers' language skills interact with their professional knowledge and the test tasks. But in EGP test, the test takers' background knowledge and expertise are part of their personal characteristics. These individual characteristics (including cultural background, background knowledge, cognitive ability, gender, age, etc.) are all factors that may affect a test taker's performance regularly, and therefore, affect the fairness of the test (Bachman, 1990).

The main purpose of LSP tests is to evaluate the test taker's specialized language ability, but their background knowledge is a necessary and an inseparable part of their specialized language ability. Their background knowledge, in this case, refers to the specialized knowledge associated with their profession or occupation. The interaction of language skills and background knowledge is a key feature of LSP testing. LSP is closely related to a specific profession. It differs with EGP in the sense that the vocabulary and discourse used, the subject matter discussed, all relate to a specific field. Languages may differ from vocabulary to syntax among different disciplines, and people in a particular field of expertise may say something that layman wouldn't understand.

### 2.3. The history and development of ESP testing

The history of ESP testing is relatively short. As early as 1913, the *Certificate of Proficiency in English* could be regarded as the beginning of ESP test. It was instituted by the University of Cambridge Local Examinations Syndicate (UCLES) to be the world's first public English examination designed for foreign students who desire a satisfactory proof of their English proficiency with an aim of teaching English in foreign schools (Cambridge English Language Assessment, 2016). Another competitor for the first ESP test title might be the *English Competence* examination established by the College Entrance Examination Board (CEEB), a test designed for international applicants who wish to study in US colleges and universities in 1930 (Spolsky, 1995). Although these two tests have clear and definite purposes related to vocational and academic English, it is still hard to define them as truly qualified ESP tests based on the two important features of LSP testing.

Taking the two criteria into consideration, the strongest candidate for the title of first ESP test is the *Temporary Registration Assessment Board* (TRAB) examination, introduced in 1975 by the British General Medical Council to assess and evaluate the professional and language abilities of international physicians for temporary registration to practice medicine in Britain (TRAB was later changed to *Professional and Linguistic Assessment Board* (PLAB) in 1978 when Temporary Registration was replaced by Limited Registration) (Rea-Dickins, 1987). Before the introduction of TRAB/PLAB, there was no individual equivalent test that can assess and recognize an overseas qualification, and thereby providing a sufficient guarantee of medical knowledge and skills for the purpose of temporary or full registration. The examination consisted of an assessment of both professional competence and English language competence. It consisted all the critical features of an ESP test: the authenticity of the test was guaranteed since the test content and method were actually based on an analysis of the language used by the medical workers and patients in British hospitals; the test was constructed and developed in a joint effort of both the language testing specialists and the medical experts as the specific domain was beyond the linguists' area of expertise; and the test was made in attempt to promote the engagement of both the test takers' language ability and their background knowledge by presenting rich and authentic test materials (Douglas, 2000).

Although the interest in LSP tests dates from the late 1970s with relevant preliminary test developments, the development of highly occupation-specific language tests did not start until the late 1980s (Stansfield, 2008). The most representative research at that moment was possibly the project undertaken by the Center for Applied Linguistics (CAL) to develop and validate a listening summary translation examination for the Federal Bureau of Investigation (FBI) in 1988 (Stansfield, Kenyon & Scott, 1990; 1992). The Listening

Summary Translation Exam (LSTE-Spanish version) was developed to assess the examinees' sufficient proficiency in both listening comprehension in Spanish and summary writing ability in English, two critical skills that were related to the job performance of language specialists in the FBI. The topics and language adapted in the test were intended to be representative of the conversations which the FBI routinely monitors, and the test itself manifested a high degree of authenticity, reliability and validity (Stansfield, Kenyon & Scott, 1996).

The development of LSP tests spread rapidly in the 1990s, extending to more countries and occupational fields. For example, the Proficiency Test in English for Air Traffic Controllers (Institute of Air Navigation Services, 1994), in which examinees had to communicate effectively and appropriately in voice-only work-related situations; the Japanese Test for Tour Guides (Brown, 1995), in which the test takers were required to play the role of a tour guide in six phases in order to show their performance in both linguistic skill and task fulfilment; and the Occupational English Test for Health Professionals (McNamara, 1996), designed to assess the language proficiency of healthcare professionals who seek to practice medicine in English speaking countries.

LSP testing benefited from the theories and frameworks of language testing, but it was not until 2000 that the nature of LSP testing was well documented by Douglas. Since then, the development of LSP testing has maintained an upward trend. Many research studies have been conducted towards some of the issues identified by Douglas (2000) as problematic for LSP testing. For instance, Elder (2001) illustrated the problems of specificity, authenticity and inseparability with reference to three specific performance-based instruments designed to assess the language proficiency of teachers;

Wu and Stansfield (2001) focused on the authenticity of task and described a working model used to determine the Target Language Use (TLU) (Bachman & Palmer, 1996) in developing the Listening Summary Translation Exam in Taiwanese (LSTE/T) that was designed in the purpose of evaluating the summary translation ability of employees who had the intention of working as linguists in the US Law Enforcement Agencies; Douglas (2001) further discussed the issue of separability of language and content by arguing that the criteria by which the performances are judged should be derived from an analysis of the same TLU situation, using the concept made by Jacoby and McNamara (1999) on 'indigenous' assessment criteria.

## 3. Pursuing Authenticity in ESP Testing

### 3.1. Concerns for authenticity in assessment criteria

This notion of authenticity as involving an interaction between the language users and the text or task was elaborated for testing purposes by Bachman (1990). Authenticity is, however, only one of several (potentially competing) components in Bachman's framework of test usefulness and may not necessarily be the prime consideration in all testing situations. Concerns for authenticity inform not only domain description and task design in LSP testing, but also the nature of interaction during the test encounter. For instance, test takers and interlocutors (in the case of speaking assessment) may frame tasks very differently, producing language that is not necessarily in accord with the test designer's intentions or representative of language behaviour in the target situation (Spence-Brown, 2001).

The assessment criteria used to judge test-taker performance are also an important consideration in evaluating test authenticity. Research (e.g., Brown, 1995; Elder, 1993; Plough, Briggs, & Van Bonn, 2010) has shown that applied linguists and content experts may not always be oriented to the same criteria in evaluations of oral performance in specific purpose contexts. Thus, it is important to determine these criteria, which represent an articulation of the test construct and should therefore reflect what is germane to the particular professional or academic context rather than general language-focused criteria familiar from other language tests. Assessment criteria, as McNamara (1996) points out, embody the test construct and should ideally reflect what domain experts consider important for effective functioning in the target setting.

Lewkowicz (2000) also reminds us, questions of authenticity extend to how tests and test scores are perceived and used. Do test takers and other test users consider the test, or the test scores, as providing an adequate representation of the ability to use language in the target domain? How do they construe the notion of authenticity? Are the expectations of test users appropriately aligned with the purpose of the test in question? And how do these perceptions impact on test preparation practices? Such questions highlight the complexities of LSP testing, and the slipperiness of the authenticity concept.

### 3.2. 'Indigenous assessment criteria'

The concept of indigenous assessment criteria was first introduced by Sally Jacoby (1998) in her dissertation, a study of conference presentation rehearsals among physicists. The criteria found through analysis were activity specific and tacitly known to insiders in the group. She defines such criteria as those used by subject specialists in assessing the communicative performances of apprentices in academic and vocational fields. Jacoby and McNamara (1999) found some significant disparities between the criteria the physicists used to judge each others' language performances and those employed in the OET, thus proposing the use of criteria indigenous in LSP tests.

Although the idea of exploring such indigenous assessment criteria has been addressed by a number of different researchers using various methods and frameworks, its actualization is still very complex and challenging (Elder & McNamara, 2016).

Erdosy (2005, 2009) established the fundamental principles behind the professor's scoring criteria for test answers through a case study of how two in-class tests were assessed in an undergraduate course on modern Chinese history taught at a Canadian university. The results suggested that indigenous assessment criteria are cloaked in substantive content and embedded in the discourse generated within communities of practice.

Fulcher, Davidson, and Kemp (2011) sought to develop a rating scale for judging the successfulness of service encounters based on current theories and empirical descriptions which capture the richness of actual performance in such contexts. However, the data are largely based on native speaker interactions and the result cannot strictly be defined as indigenous, since they do not directly capture the perspectives of insiders on what matters for effective performance.

Adbul Raof (2011) investigated the criteria oriented to by Malaysian civil engineers in evaluating conference presentations by means of semi-structured interviews taking place after the informants had viewed such presentations. The criteria raised doubts as to whether what the informants said in the presence of the interviewer reflected the actual basis for the decisions made as they watched the presentations.

Kim (2013) attempted to identify domain experts' perspectives on what mattered for effective communication in the aviation airspace, convened focus groups of aviation personnel to comment on audio-recorded episodes of actual radiotelephony discourse between pilots and air-traffic controllers. Her findings were used to interrogate the construct validity of the International Civil Aviation Organization (ICAO) policy and associated test of aviation English in Korea.

Elder and McNamara (2016) offered a qualitative comparison of domain experts' feedback using three varying degrees of authenticity in the physiotherapy workplace. The study revealed that the feedback given from the authentic workplace setting was rather scant and vague, while the less authentic workshop setting yielded richer insights to communication skills and therefore provide more material for the development of relevant LSP test criteria.

## 4. Identifying Indigenous Criteria in Assessing Communication in a Health-specific English Test

### 4.1. OET speaking test

As early as 1975, TRAB, the strongest candidate for the title of the first LSP test, was designed to assess and evaluate the professional and language abilities of international physicians to practice medicine in Britain. Thus, health-specific language test has become one of the most representative LSP test in the history and development of LSP testing.

Over the years, the establishment and development of health-specific language tests, like the Occupational English Test (OET), has attracted increasing attention as native or non-native health professionals are in need to fulfill the shortage in the healthcare workforce all around the world. For the overseas-trained health professionals, their language and communication skills are a major issue as their native language may not be suitable or may not be the principal language in the new workplace area, thus the related evaluation and assessment process became the key to this problem. Furthermore, in countries like Australia, it is required by law that overseas-trained health professionals need to pass the IELTS/ OET language test first before they take their professional competence test in order to receive license to practice (McNamara, 1996). The present OET test is recognized by 12 health-related occupations for ensuring adequate language proficiency for the non-native candidates. It consists of four skill components: Listening, Reading, Speaking and Writing. All the test materials are devised and developed collaboratively with professional experts. While Listening and Reading are the same for all professions, Speaking and Writing differ between the occupations.

Speaking is generally regarded as the most critical skill among all in healthcare communication. The OET Speaking task takes about 20 minutes and consists of two five-minute-role-play that are based on typical workplace situations. After a brief warm-up conversation with the interlocutor, the test takers are introduced to the role-play situation with a card, and are allowed to prepare for about 2-3 minutes. They will then take their professional role while the test interlocutor plays a patient/client, or sometimes a relative/carer. The whole test will be recorded and rated by the recording.

The Speaking component became the main focus of many research (e.g. Elder 2016; Elder & McNamara 2016;) as the previous criteria (Overall communicative Effectiveness, Intelligibility, Fluency, Appropriateness, and Recourses of Grammar and expression) used to rate the candidate's performance were devised without the professional experts' consultation, they only represent a common set of basic linguistic criteria. The limitation and appropriateness of these criteria have raised ongoing debates in terms of test authenticity and such a high-stake test like OET should be responsible for the decisions affecting the candidates and the healthcare stakeholders as well. There are cases that those who have passed the test were still unable to communicate effectively or still struggle to make interactions in the workplace. Thus, this has raised people's concerns about whether the candidates' performance in the present OET test is genuinely effective in a healthcare communication from a clinical perspective.

Pill (2016) drew on the feedback given by educators and clinical supervisors on trainee health professionals' performances with the patients. After decoding the data from each profession, clear similarities emerged. The features were then translated into two additional assessment criteria that can be used and expand on the more traditional linguistic criteria for the OET speaking sub-test – Clinician engagement and Management of interaction. Following Pill's research, O'Hagan, Pill and Zhang (2016) examined the results made by the seven OET assessors who were trained to apply these two newly developed professionally relevant criteria in re-assessing 300 samples from previous OET speaking test. After statistical analyses, the ratings suggested that the new criteria were consistent and aligned in terms of the speaking construct. Although it is unclear whether the language assessors adopted a professional perspective, they showed confidence in the new rating process and felt comfortable the whole time. These studies provided empirical evidence for the updated new criteria (intelligibility, fluency, appropriateness of language, resources of grammar and expression, relationship-building, understanding and incorporating the patient's perspective, providing structure, information-gathering, information-giving) which were launched in September 2018 (OET, 2018).

### 4.2. Gaining insights from professional medical licensing examination USMLE

In the United States Medical Licensing Examination (USMLE) (administered to medical students who wish to become licensed physicians in the U.S.), candidates' language proficiency is also being assessed during the Step 2 Clinical Skills examination, which uses standardized patients to evaluate the examinees' ability to engage in a conversation that allows them to gather the information needed, perform physical examinations, communicate their findings to patients or colleagues, and develop an effective physician patient relationship. The test is scored in three separate subcomponents: Communication and Interpersonal Skills (CIS), Spoken English Proficiency (SEP), and Integrated Clinical Encounter (ICE). The CIS subcomponent assesses the patient-centered communication skills of gathering information, providing information, fostering relationship, helping the patient in decision making and supporting emotions. The SEP subcomponent assesses the clarity of candidates' spoken English within the context of the doctor-patient encounter (e.g. pronunciation, word choice, and minimizing the need to repeat questions or statements). The ICE subcomponent assesses candidates' data gathering and data interpretation skills. Candidates will need to collect and write down all the information listed on a patient note during their interaction with the standardized patients.

It is surprising to find that the above three subcomponents seem to be very similar in meaning with the OET speaking criteria. While the subcomponents CIS correspond with the OET communication criteria (information gathering, information giving, relationship building, understanding and incorporating the patient's perspective) and SEP seems to correspond with the OET linguistic criteria (intelligibility, fluency, appropriateness of language and recourses of grammar and expression). This also proves, from another perspective, that clinician engagement and management of interaction are truly two important criteria in assessing the candidates' performance in a clinical setting by the professional experts.

## 5. Conclusion and Implications

This paper highlights the necessity of ESP testing and the need for interdisciplinary collaboration in the fields of ESP and ESP testing. ESP tests that base entirely on linguistic criteria may fail to satisfy the purpose of the test users in the professional domain. For instance, the construct of communicative competence informing practice in language testing is different from the views of communication informing the communication literature in the clinical setting. Gaining a better understanding of the insights of the content specialists into what constitutes effective interaction in the workplace contexts is critical in validating the claims of ESP tests like the OET to mirror the demands of real-world communication. Therefore, the use of 'indigenous' assessment criteria identified by carefully chosen professionals may provide very useful supplements to the theory and practice of ESP testing in the range of workplace and classroom settings.

Through ESP testing, language testing has expanded its scope from academia to every occupation and purpose. This has been possible because language testers started to work with practitioners and subject matter experts in many different fields in order to develop work-related or work relevant language skills tests. It is not enough to consult only the language specialists when designing the task, the criteria in the specific professional domain are equally worthy of attention in ESP test development. It is also very important to balance the expertise of language specialists and domain experts in ESP test construction, rating and score interpretation, and invites consideration of the commensurability of each party's perspective on language and communication. The problem of reconciling the different perspectives of language and non-language professionals is a perennial concern for those who work in ESP test development and validation. But by this way, the outlook for the development of more ESP tests will be excellent.

## 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, UK: Palgrave Macmillan.

Bachman, L. F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. 1991. *What does language testing have to offer?* TESOL Quarterly 25: 671-704.

Bachman, L. and Palmer, A. 1996. *Language testing in practice*. Oxford: Oxford University Press.

Brown, A. 1995. The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12: 1–15.

Carter, D. 1983. Some propositions about ESP. *The ESP Journal* 2:131-137.

Cumming, A. 2001. ESL/EFL writing instructors' practices for assessment: General or specific purposes. *Language Testing* 18: 207–224.

Davies, A. 2001. The Logic of Testing Language for Specific Purposes. *Language Testing* 18(2): 133-147.

Douglas, D. 2000. *Assessing Languages for Specific Purposes*. Cambridge Language Assessment.

Dudley-Evans, T. 1998. *Developments in English for Specific Purposes: A multi-disciplinary approach*. Cambridge: Cambridge University Press, 4-5.

Erdősy, M. U. 2005. Responding to native and non-native writers of English: A history professor's indigenous criteria for grading and feedback in an undergraduate Sinology course. Unpublished doctoral dissertation. University of Toronto, Toronto.

Erdősy, M. U. 2009. Chasing Proteus: Identifying indigenous assessment criteria in academic settings. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment* (pp. 111–133). Frankfurt am Main: Peter Lang.

Elder, C. 1993. How do subject specialists construe classroom language proficiency? *Language Testing* 10(3), 235-354.

Elder, C. 2001. *Assessing the language proficiency of teachers: are there any border controls*? *Language Testing* 18 (2): 149–170.

Elder, C. 2016. Exploring the limits of authenticity in LSP testing: The case of a specific-purpose language test for health professionals. *Language Testing* 33(2), 147-152.

Elder, C. & McNamara, T. 2016. The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. *Language Testing* 33(2), 153-174.

Fulcher, G., Davidson, F., & Kemp, J. 2011. Effective rating scale development for speaking tests: Performance decision trees. *Language Testing* 28(1), 5–29.

Halliday, M. A. K., A. McIntosh & P. Strevens. 1964. *The Linguistic Science and Language Teaching*. London: Longman.

Hutchinson, T. and Waters, A. 1987. *English for Specific Purposes: A learner-centered approach*. Cambridge University Press, 19.

Institute of Air Navigation Services. 1994. PELA: A test in the proficiency in English language for air traffic control. Luxembourg: Institute of Air Navigation Services.

Jacoby, S. 1998. Science as performance: socializing scientific discourse through conference talk rehearsals. Unpublished doctoral dissertation, University of California, Los Angeles.

Jacoby, S, & McNamara, T. 1999. Locating competence. *English for Specific Purposes 18*(3): 203-241.

Jordan, R. 1997. *English for Academic Purposes: A guide and resources book for teachers*. Cambridge: Cambridge University Press.

Kim, H. 2013. Exploring the construct of radiotelephony communication: A critique of the ICAO English testing policy from the perspective of Korean aviation experts. *Papers in Language Testing and Assessment* 2(2), 103–119.

Lewkowicz, J. A. 2000. Authenticity in language testing: Some outstanding questions. *Language testing* 17(1): 43-64.

McNamara, T. 1996. *Measuring Second Language Performance.* London: Longman.

Morrow, K. 1991. Evaluating communicative tests. In Anivan, S., editor, *Current developments in language testing*. Singapore: SEAMEO Regional Language Centre, 111-18.

O'Hagan, S., Pill, J. & Zhang, Y. 2016. Extending the scope of speaking assessment criteria in a specific-purpose language test: Operationalizing a health professional perspective. *Language Testing* 33(2), 195-216.

Pill, J. 2016. Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing* 33(2), 175-193.

Plough, I., Briggs, S., &Van Boonn, S. 2010. A multimethod analysis of criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing* 27: 235-260.

Rea-Dickins, P. 1987. Testing doctors' written communicative competence: an experimental technique in English for specialist purposes. *Quantitative Linguistics* 34: 185-218.

Robinson，P. 1991. *ESP today: a practitioner's guide*. Hemel Hempstead: Prentice Hall International.

Stansfield, C.W. 2008. Lecture 'Where we have been and where we should go'. *Language testing* 25(3): 311-326.

Stansfield, C. W., Scott, M. L., & Kenyon, D.M. 1990. Spanish - English Verbatim Translation Exam. Final Report. Ashington: CAL.

Stansfield, C. W., Scott, M. L., & Kenyon, D.M. 1992. The measurement of translation ability. *The Modern Language Journal* 76(4): 455-467.

Stansfield, C. W., Scott, M. L., & Kenyon, D.M. 1996. Examining validity in a performance test: The Listening Summary Translation Exam (LSTE) - Spanish version. *Language Testing 13*(1): 83-110.

Spence-brown, R. 2001. The eye of beholder: Authenticity in an embedded assessment task. *Language Testing* 18(4), 463-481.

Spolsky, B. 1995. *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Strevens, P. 1988. ESP after twenty years: a re-appraisal. In M. Tickoo (ed.), *ESP: State of the Art*, 1-13.

Widdowson, H.G. 1978. *Teaching Language as Communication*. Oxford: Oxford University Press.

Widdowson, H.G. 1981. English for Specific Purposes: Criteria for Course Design. In Selinker, L. (ed.). *English for Academic and Technical Purposes: Studies in honor of Louis Trimble*. Rowley, MA. Newbury House.

Wu, W., & Stansfield, C.W. 2001. Towards authenticity of task in test development. *Language Testing 18*(2), 187-206.