Language for Specific Purposes Testing: A Historical Review

Scott E. Grapin[1]
New York University

ABSTRACT

Language for specific purposes (LSP) has a long history in the language testing literature. An outgrowth of the communicative language movement of the 1970s, LSP testing arose out of the practical need to assess individuals' abilities to perform specific tasks in academic and professional settings. This historical review traces the evolution of LSP testing in the language testing literature, focusing specifically on theory and research in two key areas: (a) authenticity and (b) the interaction between language knowledge and background knowledge. The review then turns to how Douglas (2000), in the most comprehensive treatment of LSP testing to date, incorporates insights from these two lines of work into his conceptualization. Throughout the review, tensions and debates emerging from the literature are discussed. The final section addresses the uncertain future of LSP in the language testing landscape.

INTRODUCTION

Language for specific purposes (LSP) is the branch of applied linguistics concerned with the teaching and learning of language for some practical purpose (Robinson, 1989), such as Spanish for business, English for academic study, or German for mechanical engineering. Coinciding with the advent of communicative language teaching in the 1970s, LSP testing arose out of the need to make a variety of decisions about individuals' abilities to perform specific tasks in academic and professional settings. One of the earliest examples of an LSP test was an exam designed by the British General Medical Council to assess the clinical competency and language proficiency of doctors from outside the UK applying to practice medicine in Britain (Rea-Dickins, 1987). Though definitions of LSP testing vary, there is general agreement that LSP tests are different from general purpose tests, which target broader language proficiency. In the most thorough treatment of LSP testing to date, Douglas (2000) identifies two features, in particular, that distinguish LSP testing from its general purpose counterpart: (a) authenticity of task and (b) the interaction between language knowledge and background knowledge. The first two sections of this review trace the evolution of these concepts as they are discussed in the language testing literature. Then, the third section provides an in-depth examination of how Douglas weaves these concepts into his conceptualization of specific purpose language ability (SPLA). Throughout the discussion, tensions and debates that have been influential in the field's ongoing development are

[1] Scott Grapin is a doctoral student in TESOL in the Department of Teaching and Learning at NYU's Steinhardt School of Culture, Education, and Human Development. His research interests center on second language writing and the intersection of language and content learning. Correspondence should be sent to Scott Grapin, 239 Greene Street, Room 208, New York, NY 10003. Email: sg4413@nyu.edu

highlighted. In light of the literature reviewed, the final section comments on the uncertain future of LSP in the language testing landscape.

AUTHENTICITY

The notion of authenticity has captured the attention of language testers for some time. Concerns over the authenticity of language tests and tasks can be traced back to early proponents of performance testing (Clark, 1975, 1979; Jones 1979, 1985), who set out to "duplicate as closely as possible the setting and operation of the real-life situations in which proficiency is normally demonstrated" (Clark, 1975, p. 10). Morrow (1979) was likewise concerned with assessing a candidate's ability to "read, write, speak or listen in ways and contexts which correspond to real life" (p. 149). Around the same time, Widdowson (1979) added a layer of nuance to the discussion around authenticity by proposing a distinction between genuineness, a characteristic of texts written for proficient users of the language (rather than for language learners), and authenticity, a social construction arising from the interaction between a text and a reader. From Widdowson's perspective, texts were not inherently authentic but became so when authenticated by readers. Widdowson's farsighted conceptualization, however, was not readily incorporated into the mainstream of language testing, where discussions of authenticity continued to revolve around the surface features of texts and tasks. It was not until Bachman (1990) and later Bachman and Palmer (1991, 1996) that Widdowson's insights into the nature of authenticity were recognized and further developed.

Building on Widdowson's genuineness-authenticity distinction, Bachman and Palmer (1991) conceptualized authenticity as a dual notion consisting of (a) situational authenticity and (b) interactional authenticity. Situational authenticity, according to Bachman and Palmer, referred to the level of correspondence between characteristics of the test task and features of the target language use (TLU) situation. As described earlier, this type of authenticity was the main preoccupation of early performance testers. Interactional authenticity, on the other hand, aligned more with Widdowson's (1979) definition, referring to the interaction between the characteristics of individuals or test-takers (e.g., language ability, topical knowledge, affective schemata) and the TLU situation or test task. Later, Bachman and Palmer (1996) used the term authenticity to refer only to situational authenticity, while interactional authenticity was renamed interactiveness. Bachman and Palmer emphasized that both authenticity and interactiveness are relative, rather than absolute, qualities, as tests can be relatively more or relatively less authentic and interactive. In their view, the desired levels of authenticity and interactiveness for a given test depended on the use for which the test was intended. Bachman and Palmer also suggested the possibility that test-takers' perceptions of authenticity may influence their test performance, a claim that would soon be challenged on empirical grounds.

Despite important contributions by Bachman (1990) and Bachman and Palmer (1991, 1996), the conversation around authenticity, to this point, remained relatively theoretical. In response to this shortage of empirical work, Lewkowicz (2000) "put the cat among the pigeons," to borrow Alderson and Banerjee's (2001) turn of phrase, by questioning the importance accorded to authenticity in the language testing literature. In particular, Lewkowicz raised a number of outstanding questions about authenticity, both conceptual and practical, that remained unresolved in the literature. These included questions regarding the identification of critical task characteristics (for both general and specific purpose tests), the constituent elements of

authenticity (e.g., authenticity of input and authenticity of purpose), and stakeholders' perceptions of authenticity. In her study, Lewkowicz addressed this final question by evaluating Bachman and Palmer's (1996) claim that perceptions of authenticity may affect test performance. Interestingly, Lewkowicz found that authenticity was not a priority for the majority of student participants, although its perceived importance varied as a function of students' levels of proficiency. She thus concluded that authenticity "may be of theoretical importance for language testers needing to ensure that they can generalize from test to non-test situations, but not so important for other stakeholders in the testing process" (p. 60).

While the notion of authenticity has permeated all areas of performance assessment, it has been particularly influential in discussions and debates around LSP testing, where tasks are carefully constructed to "share critical features of tasks in the target language use situation of interest to the test takers" (Douglas, 2000, p. 2). Skehan (1984) emerged as an early critic of what he perceived to be an undue emphasis on content and face validity (i.e., the types of validity most closely associated with authenticity) in LSP testing. In particular, he questioned whether "authentic" tasks provided a sufficient basis for making predictions about language behavior in new situations. He noted that the problems associated with sampling in LSP tests are further compounded by the fact that discourse domains (i.e., domains in which language is used for a specific purpose) are elusive to characterize and rarely fall into discrete categories. Skehan illustrated this complexity with the example of a hypothetical needs analysis to inform a test for waiters:

> Although at first sight 'waiter behaviour' might seem to be a straightforward affair, we soon need to ask questions like: what range of customers needs to be dealt with? What range of food is to be served? Once one probes a little, the well-defined and restricted language associated with any role is revealed to be variable. (p. 216)

In light of the multidimensional and complex nature of any language performance, Skehan recommended that language testers consider face validity (and, by association, authenticity) only to the extent that it does not "interfere with the deeper aims of the test" (p. 208), whatever those deeper aims may be.

More than a decade later, Fulcher (1999) extended Skehan's (1984) criticism by problematizing what he referred to as a "simplistic view of validity" (p. 223) adopted by LSP testers. Informed by Messick's (1989) framework for test validation, Fulcher argued that, by focusing almost exclusively on content relevance and representativeness, LSP testers largely ignored the substantive and structural aspects of construct validity (perhaps what Skehan meant by "the deeper aims of the test"). As a result, issues of validity and sampling had become almost synonymous in LSP testing. In Fulcher's view, reorienting the focus of LSP testing would require placing authenticity and content validity within a larger theoretical framework that considered these in relation to the broader objective of construct validity rather than as ends in themselves. Still, Fulcher was careful to note that, despite these limitations, constructing tests based on content analyses has its advantages, including beneficial washback on instruction and high face validity in the eyes of various stakeholders. These constitute important evidence of consequential validity that should also be taken into account within a broader framework.

Over the course of its evolution in the language testing literature, the notion of authenticity has been the subject of considerable discussion and debate. More than a decade and a half ago, Lewkowicz (2000) called for more focused inquiry into the nature of authenticity, its

impact on test performance, and its importance to various stakeholders. Today, many of her questions remain unanswered and open to investigation. It is hoped that the coming years bring about renewed interest in this area but, in the meantime, the notion of authenticity will continue to occupy a special role in LSP testing. The next section traces the evolution of a second concept central to the LSP testing enterprise: the interaction between language knowledge and background knowledge.

INTERACTION BETWEEN LANGUAGE KNOWLEDGE AND BACKGROUND KNOWLEDGE

While authenticity, on the one hand, and the interaction between language knowledge and background knowledge, on the other, are addressed separately in this review, they are perhaps best understood as two sides of the same coin. A distinguishing feature of LSP contexts (e.g., using Spanish for business) is that they require individuals to make use of both their language knowledge and background knowledge. For instance, in pitching a new product to potential clients, a business professional may draw on her knowledge of Spanish as well as her knowledge of marketing, sales, and statistics. While this specific purpose background knowledge has traditionally been viewed as a source of construct-irrelevant variance in language test performance (Llosa, 2016), it is considered an integral aspect of the LSP construct (Douglas, 2000). Thus, if LSP test tasks are to be considered authentic (particularly in Bachman's interactional sense), they too should engage both language and specific purpose background knowledge. A task that failed to meet these criteria would be of questionable authenticity.

Because of the potential for language and background knowledge to be confounded in LSP test performance and interpretation, LSP testers have long been interested in disentangling the relationship between the two. Alderson and Urquhart (1985) were among the earliest testers to attempt to investigate this relationship in the context of second language reading performance. Specifically, they examined the effect of background knowledge on reading comprehension in the subject-specific modules of the English Language Testing System (ELTS), a precursor to the well-known International English Language Testing System (IELTS) exam and an important site for early LSP research. They found that background knowledge seemed to be "operating" in test performance but in rather inconsistent ways. For example, engineering students scored significantly lower than business and economics students on modules pertaining to business and economics but scored the same as business and economics students on modules pertaining to engineering. Alderson and Urquhart pointed to differences in linguistic proficiency as a possible explanation for these inconsistent results, as high-proficiency test-takers may have compensated for their lack of background knowledge with "generalized text-handling ability" (p. 202). They also raised questions about the level of specificity appropriate for constructing discipline-specific texts and tasks as well as the feasibility of distinguishing between "general and particular knowledge" (p. 202). In a follow-up study appropriately titled "This Test is Unfair: I'm not an Economist," Alderson and Urquhart (1988) turned up similar findings. Faced with such inconclusive results, they recognized the need for "an explanation which combines the effects of linguistic proficiency and of background knowledge" (p. 41).

Nearly a decade later, Clapham (1996) answered this call by providing the most comprehensive study to date on the relationship between language knowledge and background knowledge in LSP test performance. Like Alderson and Urquhart (1985), Clapham focused on

test-takers' performance on the subject-specific reading modules of the IELTS. This work yielded a number of interesting findings with important implications for LSP testing: First, students generally scored higher on reading sub-tests in their own field than in other fields. Second, for students with scores of less than 60% on the grammar sub-test (i.e., the measure of domain-general language proficiency used in the study), there was no significant effect of background knowledge on reading test performance. Conversely, for students who scored 60% or above, the effect was significant. Thus, it appeared that test-takers needed to reach a threshold of language proficiency before they were able to make use of their background knowledge. In addition, students who scored particularly high on the grammar sub-test (i.e., above 80%) were less affected by their level of background knowledge, presumably because they were able to compensate for any deficiencies in background knowledge by leveraging their language proficiency. Finally, Clapham found that background knowledge exerted a stronger effect on test scores as the IELTS reading passages became more discipline-specific. Taken together, these findings suggest that, for students with intermediate language proficiency, background knowledge interacts with language knowledge in the completion of LSP reading tasks. Moreover, the specificity of test input may play a role in mediating this relationship (i.e., more specific tests may require more background knowledge).

Though much work remains to be done in this area, Clapham's (1996) study represented a significant advance in our understanding of the relationship between language and background knowledge in LSP test performance. At the same time, it raised important questions about the level of specificity needed to engage test-takers' SPLA, or as Fulcher (1999) succinctly put it, "How specific is specific?" (p. 230). This persistent question facing LSP testers would soon be taken up by Douglas (2000) in the most comprehensive conceptualization of LSP to date.

SPECIFIC PURPOSE LANGUAGE ABILITY

Building on his previous work in LSP testing (e.g., Douglas & Selinker, 1992), Douglas (2000) sets out to bring LSP "more in line with the theoretical underpinnings of communicative language testing" (p. 11). Specifically, he considers LSP to be a special case of communicative language testing in which:

> test content and methods are derived from an analysis of a specific purpose target 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 purpose content knowledge, on the one hand, and the test tasks on the other. (p. 19)

In Douglas's definition, we see clearly the influence of the two lines of research traced above: (a) authenticity and (b) the interaction between language knowledge and background knowledge. In his conceptualization of SPLA, Douglas modifies Bachman and Palmer's (1996) model of communicative language ability by including background knowledge, in addition to language knowledge and strategic competence, as part of the SPLA construct. Strategic competence, then, is seen as a mediator between the external situational context and internal language and background knowledge.

A priority for LSP testers, according to Douglas (2000), is to engage test-takers in specific purpose contexts of language use so as to make valid inferences about their SPLA. Douglas is critical of previous conceptualizations of context (e.g., Hymes, 1974), which failed to take into account how participants themselves interpret the communicative situation. In the case of LSP testing, if test-takers do not recognize the situational factors of test tasks (setting, participants, etc.) as indicative of particular contexts of language use, regardless of the transparency of these cues to test developers, it is unlikely their SPLA will be engaged. Thus, Douglas prefers to think of context in terms of discourse domains, a cognitive construct that is activated when sufficient contextualization clues are provided to engage test-takers in specific purpose language use. These contextualization clues are operationalized by Douglas in the form of task characteristics using Bachman and Palmer's (1996) framework. In Douglas's view, the use of task characteristics that reflect the essential features of the TLU situation allows LSP test developers to capitalize on method effects and offers a way out of the "dilemma of never-ending specificity on the one hand and non-generalizability on the other" (p. 19). Importantly, the criteria for judging LSP performance are also derived from an analysis of the TLU situation, although Douglas cautions that some aspects of these "indigenous assessment criteria" (Jacoby, 1998) may not be relevant to the construct of interest (e.g., the appearance of test-takers). In line with McNamara's (1996) weak form of performance assessment, Douglas asserts that LSP tests do not attempt to measure communicative success, but rather, the knowledge and abilities underlying communicative performance.

Though Douglas (2000) has been instrumental in establishing a conceptual foundation for LSP research and practice, questions linger regarding the theoretical soundness and practical value of LSP testing. Motivated by his earlier research on the ELTS/IELTS exam (e.g., Criper & Davies, 1988), Davies (2001) asks the critical question of whether the LSP testing enterprise is a logical one. He answers this question in two ways. First, he notes that LSP testing faces serious problems related to the still unclear role of background knowledge in test performance (e.g., Alderson & Urquhart, 1988) and the difficulty of delineating language varieties in any principled way (echoing earlier observations by Skehan, 1984). As a result of these persistent problems, Davies claims that LSP tests have not proven to be more valid than tests of general proficiency. His second answer, however, is decidedly more optimistic. Davies observes that, from a pragmatic perspective, if LSP tests have a positive impact on teachers and learners and do not predict any less well than general proficiency tests, their value can be justified. Like Fulcher (1999), Davies recognizes the need to consider the merits and drawbacks of LSP testing within a broader theoretical framework of test use (e.g., Messick, 1996).

CONCLUSION

On the surface, the case for LSP tests is compelling. As Alderson and Urquhart (1988) put it, it is unlikely "the performance of a would-be post-graduate student of dentistry on a text about piracy in the seventeenth-century Caribbean can be used to predict his or her ability to read research material in dentistry" (p. 26). But despite its intuitive appeal, the LSP testing enterprise is fraught with issues, tensions, and debates. Language testers have long grappled with the elusive nature of authenticity, its role in test performance, and its importance to various stakeholders. In the same way, disentangling the relationship between language knowledge and background knowledge in LSP test performance has proven to be a complex endeavor. This is

not to mention the multitude of practical constraints associated with LSP testing which are beyond the scope of this review but loom large nonetheless. Summarizing a plenary discussion on LSP testing at the 2010 meeting of the Language Testing Forum, Brunfaut (2014) shows how current conversations in the field continue to revolve around this same set of core issues. Adding to the complexity of this picture, Purpura (2017) has recently argued that background (topical) knowledge plays a role in any language performance, an assertion that calls into question the utility of a distinction between general and specific language tests. Perhaps what this implies is a need for, in Davies' (2001) words, "richer proficiency tests" (p. 143) with more precise specifications. Given these considerations, the future of LSP testing remains uncertain. In the years to come, continued inquiry into the nature of SPLA, LSP performance, and the use of LSP tests will work to address these issues.

REFERENCES

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part 1). *Language Teaching, 34*(4), 213-236.

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2*, 192-204.

Alderson, J. C., & Urquhart, A. H. (1988). "This test is unfair: I'm not an economist." In P. C. Carrell, J. Devine, & D.E. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 168-182). New York, NY: Cambridge University Press.

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

Bachman, L. F., & Palmer, A. S. (1991). What does language testing have to offer? *TESOL Quarterly, 25*, 671-704.

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

Brunfaut, T. (2014). Language for specific purposes: Current and future issues. *Language Assessment Quarterly, 11*(2), 216-225.

Clapham, C. (1996). *The development IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge, England: Cambridge University Press.

Clark, J. (1975). Theoretical and technical considerations in oral proficiency testing. In S. Jones & B. Spolsky (Eds.), *Language testing proficiency* (pp. 10-24). Arlington, VA: Center for Applied Linguistics.

Clark, J. (1979). Direct vs. semi-direct tests of speaking ability. In E. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 35-39). Washington, D.C.: TESOL.

Criper, C., & Davies, A. (1988). *English language testing service validation project report 1(i)*. Cambridge, England: University of Cambridge Local Examinations Syndicate.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18*(2), 133-147.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, England: Cambridge University Press.

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific-purpose contexts. *System, 20*(3), 317-328.

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics, 20*(2), 221-236.

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach.* Philadelphia, PA: University of Pennsylvania Press.

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

Jones, R. (1979). Performance testing of second language proficiency. In E. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 50-57). Washington, D.C.: Teachers of English to Speakers of Other Languages.

Jones, R. (1985). Second language performance testing: An interview. In P. C. Haupman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 104-115). Ottawa, Canada: University of Ottawa Press.

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

Llosa, L. (2016). Assessing students' content knowledge and language proficiency. In E. Shohamy & I. Or (Eds.), *Encyclopedia of language and education, Volume 7* (pp. 3-14). New York, NY: Springer International Publishing.

McNamara, T. (1996). *Measuring second language performance.* London, England: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*. New York, NY: American.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143-157). Oxford, England: Oxford University Press.

Purpura, J. E. (2017). Assessing meaning. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment*. New York, NY: Springer.

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

Robinson, P. C. (1989). An overview of English for specific purposes. In H. Coleman (Ed.), *Working with language: A multidisciplinary consideration of language use in work contexts* (pp. 395-427). Berlin, Germany: Mouton de Gruyter.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing, 1*, 202-220.

Widdowson, H. (1979). *Explorations in applied linguistics*. Oxford, England: Oxford University Press.