TELL, Vol. 2, No. 1, 2007

A Test of English for Specific Purposes: Need Analysis, Design, and Evaluation

Parviz Birjandi
Allame Tabatabai University
Mona Khabiri
Azad University, Centeral Tehran Branch

Abstract

The current study aimed at constructing a test of Language for Specific Purposes (LSP) as a substitution for the English proficiency test of Postgraduate TEFL Admission examination (PTA) of Islamic Azad University. To this end, the study was instigated with a thorough needs analysis, by means of questionnaires, interviews, and observations, to determine the requirements of the postgraduate TEFL course. Overall, 294 subjects participated in this study. Seventy four postgraduate TEFL students and twenty university professors participated in the needs analysis phase of the study. Subsequent to the analysis and interpretation of the obtained data, the Test of English for Specific Purposes (TESP) was constructed based on the requirements of the program. TESP was put into trial by 200 subjects in three different pilot studies, each of which led to modification of the form and the content of the test. The scores obtained on the final edition of TESP proved to be highly reliable and both *a priori* and *a posteriori* evidence were gathered regarding the validity of TESP in addition to the convergent and divergent validity evidence that were provided for the questionnaires of the needs analysis. Following this process of standardization, TESP was compared with the English proficiency test of PTA in

predicting the Grade Point Average (GPA) scores of postgraduate TEFL students. Multiple regression models were drawn for the English proficiency test of PTA and TESP as the predictor variables and GPA as the predicted variable. The results indicated that TESP was a significant predictor of postgraduate TEFL students' GPA scores, whereas PTA was excluded from the regression model.
Key words: TEFL, TLU, PTA, GPA score, LSP.

1. Introduction

In the past twenty years, language testing research and practice have witnessed the refinement of a rich variety of approaches and tools for research and development, along with a broadening of philosophical perspectives and the kinds of research questions that are being investigated. Bachman (2000) maintains that in 1990s, the field of language testing witnessed expansion in a number of areas such as, research methodology, practical advances, factors that affect performance on language tests, and performance assessment. With respect to practical advances, he points to computer-based assessment and Testing Languages for Specific Purposes (LSP testing). The latter is the focus of this study.

2. Testing Languages for Specific Purposes

Douglas (2000) defines LSP testing (Testing Language for Specific Purposes) as referring to that branch of language testing in which test content and test methods are derived from an analysis of a specific language use situation. Douglas further identifies two major characteristics of LSP tests: *authenticity of task*, and *interaction between language knowledge* and *specific purpose content knowledge*. The former refers to the similarity of test tasks to the tasks in the target language use situation in order to increase the likelihood that the test taker will carry out the test task in the same way as the task would be carried out in the actual target situation. Douglas believes that the latter is the clearest defining feature of LSP testing. In fact, what makes

Douglas refer to the communicative language ability in LSP testing as 'specific purpose language ability' is the emphasis he puts on 'background' or 'topical' knowledge which he defines as "discourse domains; frames of reference based on past experience which we use to make sense of current input and make predictions about that which is to come" (p.35).

Consequently, the first step in LSP test construction is the identification, description, and analysis of target language use (TLU) situation. Bachman and Palmer (1996) define target language use domain as "a set of specific language use tasks that the test taker is likely to encounter outside the test itself, and to which we want our inferences about language ability to generalize" (p. 44). In identifying TLU domain, Douglas (2000) emphasizes the role of context, and maintains that in specific purpose language testing, one must be careful to ensure that the discourse domain of the target language use situation is well signaled in the test.

3. Language Performance Assessment

LSP tests can be considered as one type of performance assessment. Various types of performance assessments have been used in language testing for years. In fact, virtually all language tests have some degree of performance included. For McNamara (1996), a defining characteristic of performance assessment is that "actual performances of relevant tasks are required of candidates, rather than more abstract demonstration of knowledge often by means of pencil-and-paper tests" (p.6).

McNamara (1996) maintains that performance assessment is essentially a 'methodological' issue (p.9). He introduces two approaches to second language performance assessment. The first one, he calls the *work sample* approach and maintains that in this approach, performance is the target of assessment. McNamara defines the second approach to be a more cognitive and distinctively linguistic approach, in which attention is focused less on the task but on the *qualities of execution in the performance*, and/or the evidence it provides about the candidates control of the underlying *linguistic system*.

> In the words of Messick (1994), the performance is the vehicle of assessment, the performance task itself is of less interest than what the performance reveals, the underlying knowledge and ability is the actual target of assessment. (1996, p. 25)

McNamara adopts a broad view of second language performance tests as characterized by "a relatively simple performance requirement, that is, that assessment will take place when the candidate is engaged in an act of communication" (p.26). McNamara, therefore, introduces a dichotomy for language performance assessment, by offering a distinction between a *strong* and a *weak* sense of the term second language performance test, depending on the extent to which assessment criteria reflect the non-linguistic aspects of task performance. Norris et al. (1998) maintain that virtually all language tests have some degree of performance included and that "it might be more appropriate to think of tests as more performance oriented or less performance oriented along a continuum from least direct and least real-world or authentic to most direct and most real-world or authentic" (p.3).

McNamara's distinction is proposed in terms of the *criteria* used in assessing performance on the tasks set. In the *strong* sense, tasks will represent real-world tasks and performance will primarily be judged on real-world criteria, that is, the fulfillment of the task set. He asserts that this type of test is not strictly a language test at all. However, second language performance tests in the *weak* sense focus on the language performance. The task may resemble or simulate real-world tasks, or be artificial in other ways (e.g. the Oral Proficiency Interview). He claims that the capacity to perform the task is not actually the focus of assessment; rather, the purpose of the assessment is to elicit a language sample so that second language proficiency, and perhaps additional qualities of the execution of the performance, may be assessed.

4. Authenticity in Language Testing

The characterization of authenticity is one of the most problematic concerns of language testing. However complex and difficult to be defined and specified, the importance and centrality of authenticity cannot be denied. Bachman describes the preoccupation with authenticity as reflecting "a sincere concern to somehow capture or recreate in language tests the essence of language use" (p. 300) and asserts that authenticity is important as a way of ensuring that language tests reflect language use in the target domain, and that their results are thus valid for application in that domain.

Bachman (1990) introduces two approaches toward authenticity; the 'real-life' (RL) approach and the 'interactional/ability' (IA) approach. What he calls real-life approach is concerned with the extent to which test performance replicates some specified non-test language performance. On the other hand, IA approach is focused on the interaction between the language user, the context, and the discourse; as such interaction is the distinguishing characteristic of communicative language use.

By pointing to the difficulty of simulating the real-life language use situation and the problem of construct validity, Bachman maintains that rather than insisting on replicating real-life language use in the test situation, test developers need to recognize that the test language is inherently different from the real-life language and try to define what constitutes 'authentic test language' (p.314). This entails two requirements, namely the test taker and the characteristics of the test method facets.

5. Method

Through its bimethodological design, this study investigated the opinion of postgraduate TEFL students and their professors qualitatively regarding the proficiency level of these students and the language needs of the postgraduate TEFL program. The groundwork was to evaluate the screening of postgraduate TEFL admission examination as well as illustrating the Target Language Use situation, based on which the researcher could

design and construct a language for specific purposes test (LSP test) and then quantitatively compare it with the proficiency test that is currently used in the postgraduate TEFL admission examination. Therefore, the LSP test was compared with PTA with respect to its predictability of GPA scores of the subjects.

6. Subjects

By and large, subjects who participated in this study comprised 274 students and 20 university instructors. These subjects took part in different phases of the study, that is, the needs analysis and pilot studies.

The first group of subjects who contributed to the needs analysis of this study comprised 74 postgraduate TEFL students studying at Islamic Azad University, Tehran Central branch, and Science and Research Campus. Out of the 74 subjects who completed the questionnaire, 20 also took part in an interview for needs analysis. The second group of subjects that participated in the needs analysis consisted of 20 Heads of Departments and university instructors, who teach postgraduate TEFL courses at Islamic Azad University, Allameh Tabatabaie University, and University of Tehran.

The subjects who took part in the first pilot study included 70 senior undergraduate students from Islamic Azad University, Tehran Central branch, majoring in English translation. For the second pilot study, the test was administered to 50 undergraduates of English translation who were attending preparatory courses for postgraduate admission examination and intended to take part in this examination within few months. Ultimately, for the final pilot study the test was administered to 80 postgraduate TEFL students at Islamic Azad University, Tehran Central Branch, and Science and Research Campus, Tarbiyat Moallem University, and University of Tehran.

7. Instrumentation

As suggested by Graves (2001), three instruments were used for needs analysis: class observation, interview, and questionnaire. The class observation included 10 hours of audio-taped

observation of postgraduate TEFL classes. There were 4 questionnaires used in this study. The first and the preliminary questionnaire had the purpose of determining the questions and possible alternatives of the subsequent questionnaires. Interviews accompanied the questionnaires. The purpose of the questionnaires and the interview was twofold; first to carry out needs analysis and second to make an opinion survey.

In addition to the aforementioned instruments, a Test of English for Specific Purposes was designed and constructed based on the needs analysis. Since TESP was put into trial in three pilot studies, it can be said that it included three editions. In addition to the answer sheets, each edition was accompanied by a questionnaire that asked about the examinees' opinion regarding different method facets such as difficulty of texts or tasks, sufficiency of allocated time, the clarity of the instructions, and salience of parts. As suggested by Bachman (1990) and Bachman and Palmer (1996) the results of these questionnaires were used both for modification of the test and as evidence for the validity of the test. There were also questions on the comparison between TESP and the proficiency test of PTA.

TESP included three modules: Reading, Writing, and Speaking Modules. The speaking module was only used in the final edition since only objective and close-ended items were to be included in the pilot studies for further revision and modification. In the final edition, the reading module of TESP consisted of 39 tasks, and the writing module consisted of two tasks; a gap-fill summary based on the information and comparison delineated in a chart, and an argumentative essay. The speaking module included an oral interview which consisted of two parts: *Introduction* and *Extended Discourse*. The introduction part of the interview was designed to put the candidate on familiar grounds, the second part asked the candidates to produce some extended discourse on a familiar topic involving description, comparison and contrast, and argumentation.

In addition to the three modules of TESP, three rating profiles were used as the instruments of the study. Two profiles

were used for ratings of the speech of interviewees during and after the interview and one for the rating of the writing module. The '*Interview Scoring Profile*' was designed by the researcher based on the analytic assessment criteria for Cambridge Speaking Test and included four analytic scales. The '*UCLES Common Scale for Speaking*' (UCLES 1999f) (UCLES stands for University of Cambridge Local Examination Syndicate), which is used for the assessment of the Main Suite Cambridge EFL Examination (Lazaraton, 2002), was utilized in this study both for the global rating of the candidates during the interview and as a criterion for validating the *Interview Scoring Profile* which was designed and constructed by the researcher for this study.

The Writing Assessment Scoring Profile, which consisted of six analytic scales, was designed by the researcher. The profile was used for rating the second task of the writing module, which was an argumentative essay. The researcher designed and constructed the profile based on the requirements determined by the needs analysis study. However, some ideas were extracted from the *Michigan Writing Assessment Scoring Guide* (Hamp-Lyons, 1990, as resented in Weigle, 2002), which is used for grading an entry-level university writing examination.

8. Procedure

As mentioned before the study benefited both from qualitative and quantitative methods of research. The qualitative part of the study consisted of an opinion survey and a needs analysis study which was the foundation for the design and construction of TESP. After consummating the analysis and interpretation of the data obtained through needs analysis, the researcher embarked on designing and constructing TESP. Subsequently, TESP was put into trial in three different pilot studies, each of which led to certain modifications and amendments in the form and content of the test. Through the pilot studies, the researcher attempted to remove any malfunctioning element in the test to standardize it for the target population that are usually graduates who are candidates of postgraduate TEFL admission examination.

The speaking module of TESP included an oral interview which consisted of two parts: *Introduction* and *Extended Discourse*. For each task set, a task card was prepared and offered to the candidates and the interviewers also possessed an Interlocutor's Guide Script. Paired-format was used in the interview, where two examinees evaluated two candidates. One of the examinees served the role of interlocutor and managed the interaction and provided a global assessment using the *UCLES Common Scale for Speaking* and the other was the assessor, a passive observer who applied detailed analytic criteria to the candidates' performance utilizing the *Interview Scoring Profile* which was designed and constructed by the researcher for this study.

For the second writing task, as mentioned in instrumentation section, the researcher had prepared assessment criteria, that is, *Writing Assessment Scoring Profile*, which entailed six analytic scales. Subsequent to the final administration, the researcher started training other raters for rating the scripts based on the scoring profile. To this end, two raters were invited who held MA in TEFL and the training took place during two consecutive sessions in which the researcher explained in detail all the scales within the profile, and a few scripts were analyzed and rated according to the profile. Moreover, as the researcher was reading the scripts, she wrote some descriptive notes regarding the features of each script prior to the rating of the scripts according to the scoring profile. This was carried out for 42 scripts. Then after rating the 42 scripts, for each script the rating was compared with the descriptive notes. The comparison revealed a high correlation or match between the descriptive notes and the features determined for the selected level of the rating scale. This was considered as evidence of the validity of the Writing Assessment Scoring Profile. Another evidential basis for the validity of the profile was the fact that the profile had been read and proved by the experts in the field. Thus, the researcher was assured that the profile assessed what it claimed to assess. The reliability estimates of the profile are reported in the 'Results' section.

Finally, after standardization of TESP, the researcher embarked on comparing the results of TESP with the English proficiency test of Postgraduate TEFL Admission examination (PTA) of Islamic Azad University. In order to evaluate and compare the predictability of TESP and PTA, GPA (Grade Point Average) scores of all 54 examinees were requested from the responsible authorities in the respective universities. Moreover, the raw scores of these examinees on the English proficiency test of PTA were requested from the Examination Syndicate of Azad University. Finally, the examinees' scores on TESP and PTA were compared with their GPA scores.

9. Design

The discussion of the design of this study requires a reference to the quantitative and qualitative approaches to research, as the design of this study merits both approaches. Lazaraton (1995) maintains that "very few (too few, perhaps) researchers design studies that employ both qualitative and quantitative approaches, despite the fact that today, bimethodologicalism may be a true mark of scholarly sophistication" (p. 463).

The qualitative aspect of the current study, were the initial needs analysis and opinion survey carried out through class observations, interviews, and questionnaires. The quantification carried out on this qualitative domain of the study was in the form of descriptive statistics as mentioned by Lazaraton (1995). Furthermore, the descriptive notes that the researcher wrote about the writing scripts of the examinees and the oral performance of the interviewees during the interview were more of a qualitative nature than quantitative. Finally, the questionnaires that were attached to the answer sheets in each of the three pilot studies and gathered the opinion of the examinees regarding different aspects of each module of the test can be encompassed by the qualitative domain of the study.

Regarding the quantitative aspect of this research study, the design adopted was an *Ex post facto* design. The researcher had no control over what had already happened to the subjects. Moreover, in comparing the predictability of TESP with

entrance examination, the researcher sought to look at the type or degree of relationship between two variables rather than at a cause-and-effect relationship.

10. Analysis and Interpretation of Results

As discussed above, subsequent to the design and construction of TESP, three pilot studies were carried out which resulted in the standardization of this test. The results of the final pilot study will be reported hereunder.

TESP included reading, writing, and speaking modules. The mean, standard deviation, variance, and standard error of mean are reported for each module as well as the entire TESP package in Table 1.

The reliability estimate for the 39 items of the reading module, using Cronbach alpha, came out to be 0.83. The Cronbach alpha reliability was estimated 0.76 for the first task of the writing module, which contained 10 items. For the second task of this module, inter-rater reliability estimates were computed for the three raters and they all came out to be significant as reported in Table 2. The results indicated that the ratings of the raters based on the Writing Assessment Scoring Profile, were all consistent and reliable. However, the highest correlation is observed between R1 and R2.

Table 1: Descriptive statistics for final pilot study of TESP

	Total Reading	Total task 1&2 writing	Total reading & writing	Speaking	Total Score
No.	81 6	81 6	87 0	30 57	27 60
Mean	21.84	10.73	30.33	15.93	48.69
Std Erro of mean	.598	.424	1.04	.893	2.18
Std. Deviation	5.38	3.81	9.74	4.89	11.30
Variance	28.99	14.53	94.92	23.91	127.75

Table 2: Inter-rater consistency of WASP

		Total score Rater1	Total score Rater2	Total score Rater 3
Total score (R1)	Pearson Corr.	1.000	.963**	.956**
	Sig. (2-tailed)		.000	.000
	N	66	66	66

** Correlation is significant at the 0.01 level (2-tailed)

The Cronbach alpha reliability for the two modules, that is, the reading and writing modules (62 items) came out to be 0.87, which was higher that the reliability estimate obtained for the two modules in the second pilot study, that is, 0.75.

As a piece of evidence for the internal validity of Writing Assessment Scoring Profile (WASP), the correlation between each of the six analytic scales of WASP and the total writing score was estimated. The results indicated that all the six analytic scales of WASP were highly correlated with the total writing score in the case of all three raters, as all the correlations obtained were significant at 0.01.

Regarding the speaking module, which was in the form of an interview, the inter-rater reliability estimate for the analytic ratings of rater 1 (R1) and rater 2 (R2) based on the ISP came out to be 0.97, which was quite significant, showing a high consistency between the ratings of the two raters. Moreover, the inter-rater reliability estimate for the global ratings of R1 and R2, the intra-rater reliability for global ratings of R1, and the intra-rater reliability for the first and second analytic ratings of R1 are reported in Table 3.

Table3: Inter/intra-rater consistencies for analytic and global ratings of the interview

		Analytic rating R1	Analytic rating R2	Global rating R1	Global rating R1	Global rating R2
Analytic rating						
R1	Pearson corr.	1.000	.969**	.940**	.951**	.963**
	Sig. (2-tailed)	.	.000	.000	.000	.000
	N	24	24	24	24	24
Global rating						
R1	Pearson corr.	.940**	.919**	1.000	.975**	.867**
	Sig. (2-tailed)	.000	.000	.	.000	.000
	N	24	24	24	24	24

** Correlation is significant at the 0.01 level (2-tailed).

As a piece of evidence for the validity of the ISP, the correlation between the analytic ratings based on the *Interview Scoring Profile* (ISP) and the global ratings based on UCLES Common Scale for Speaking was estimated. For the first rater (R1), the correlation coefficient between these two rating scales came out to be 0.94. However, for the second rater (R2), the correlation between the analytic ratings based on ISP and the global ratings based on UCLES came out to be 0.97.

As an indicator for the validity of TESP, some of the results of the questionnaire offered to the examinees will be reported. Regarding the difficulty of the test, 60% and 50% of the examinees reported that the text and task difficulty of the reading passages 1 and 2, respectively, were moderate. Regarding the instructions of each module as a feature of test method facet, 81% of the examinees maintained that the instruction for the tasks of reading module were comprehensible and clear. This figure changed to 92%, for the instructions of the writing tasks. Finally, concerning the effect of test method facet and test wiseness, 67% of the examinees reported that they were not familiar with the test rubric, but that according to the instructions they were able to perform the tasks of the reading module. This figure changed to 56% for the tasks of the writing module.

11. The Predictability of TESP and PTA

As was mentioned in Method section, out of 80 subjects of the third and final administration of TESP, 53 were postgraduate students of Islamic Azad University, Tehran Central branch and Science and Research Campus. These 53 examinees were the entire population of postgraduate TEFL students at Azad University in Tehran at the time of the study. Multiple regressions was used as the statistical procedure to evaluate the predictability of TESP and postgraduate TEFL Admission Examination (PTA) regarding the Grade Point Average (GPA) scores of the postgraduate TEFL students of Azad University. The two predictors or independent variables, therefore, were examinees' scores on TESP and PTA.

The first step is to analyze the correlation between the independent and dependent variables. As demonstrated in Table 4 and 5, the correlation between TESP and GPA scores came out to be 0.38, and the model for this regression came out to be significant ($F_{(1,29)}$= 4.368, p= 0.047).

Table 4: Model summary for Regression – TESP & GPA

Model		Sum of Squares	Df	Mean Square	F	Sig
1	Regression	10.977	1	10.977	4.368	.47[a]
	Residual	65.342	29	2.513		
	Total	76.319	30			

a. Predictors: (constant), TESP
b. Dependent Variable: GPA

Table 5: Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coeffiecients	t	Sig
		B	Std. Error			
1	(constant)	12.737	1.510		8.437	.000
	TESP	6.533E-02	.031	.379	2.090	.047

a. Dependent Variable: GPA

Since the model also came out to be significant for PTA and GPA ($F_{(1, 29)} = 0.901$, p = 0.352), backward method was used in the multiple regression to evaluate the contribution of the predictor variables. The results indicated that the *t* value for TESP (*t* = 2.058) was higher than the *t* value for PTA (*t* = 0.387), and through the backward method PTA was excluded as the insignificant predictor variable for GPA. That is the model has identified TESP as a better predictor for the GPA scores of the examinees. The results are illustrated in Table 6.

Table 6: Results on Multiple Regression (Predictor Variables = TESP, PTA)

Adjusted R square = 0.111; $F_{(1,29)}$= 4.23, *p* = 0.050 (using the backward method). Significant variable is shown below.			
Predictor Variable	**Beta**	***t***	***P***
TESP	0.381	2.058	*P*= 0.050
PTA was not a significant predictor in this model.			

12. Conclusion

The conclusion from the two regression models discussed above is that the scores on the English proficiency test of PTA do not have predictability about the future performance of the examinees in postgraduate program. However, the scores on TESP proved to have significant prediction regarding the actual performance of the examinees in TLU situation. Consequently, it is proved by the results of this study that an appropriate and valid selection tool is not being used in the Postgraduate Admission Examination. A comprehensive and multi-dimensional reform is, thus, required to be made to the English proficiency test of postgraduate TEFL admission examination.

13. Pedagogical Implications

The nature and the results of this research study conduce to an assortment of appealing enquiries, and thus implications that only few of them are mentioned hereunder:

1. Substitution of the English proficiency test of PTA examination with TESP which leads to the improvement of the quality of screening procedure as well as the overall quality of the postgraduate TEFL program.
2. Realization of the necessity of pursuing a similar study on the knowledge test of the PTA examination according to the results of the needs analysis of this study in which majority maintained that the knowledge test of PTA did not screen candidates' background knowledge appropriately.
3. Proposition of a first-degree program evaluation which can lead to the improvement of the English language ability of the graduates who are the candidates of PTA examination.
4. Suggestion for a supplementary research dealing with the appropriacy of designing a common English proficiency test for both candidates of postgraduate TEFL and English Literature program.

References

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

Bachman, L, F. (1990). Constructing measures and measuring constructs. In Harley, B., Allen, P., Cummins, J., & Swain, M., (Eds.), *The Development of Second Language Proficiency*. New York: Cambridge University Press.

Bachman, L, F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17(4), 1-42.

Bachman, L, F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Dauglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing*, 18(2), 171-185

Gipps, C. V. (1994). *Beyond Testing: Towards a Ttheory of Educational Assessment*. London: The Falmer Press.

Graves, K. (2001). A framework of course development processes. In Hewings, A., & Hall, R. D. (Eds.), *Innovations in English Language Teaching*. London: Routledge.

Lazaraton, A. (1995). Qualitative research in applied linguistics: A progress report. *TESOL Quarterly*, 29(3), 455-472.

Lazaraton, A. (2002). A qualitative approach to the validation of oral language tests. *Studies in Language Testing 14*, Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring Second Language Performance*. New York: Longman.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In Wainer, H., & Braun, H. I. (Eds.), *Test Validity*. New Jersey: Lawrence Erlbaum Associates Publishers.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing Second Language Performance Assessment*. HI: University of Hawai'i Press.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.