A Needs-Based Approach to the Evaluation of the Spoken Language Ability of International Teaching Assistants

Shahrzad Saif
Université Laval

This study addresses the problem of appropriately assessing the spoken language ability of non-native graduate students functioning as international teaching assistants (ITAs) in English-speaking environments in general and that of a Canadian university in particular. It examines the problem with reference to the needs of ITAs in actual contexts of language use in the light of two validity standards of 'authenticity' and 'directness' (Messick, 1989) and the model of language testing proposed by Bachman and Palmer (1996). The paper summarizes the results of a needs assessment carried out among three major groups of participants at the University of Victoria: administrators and graduate advisors, undergraduate students and the ITAs themselves. Test constructs are then formulated based on the results of the needs analysis. It is also shown how test constructs are translated into the communicative task types that would involve ITAs in performances from which inferences can be made with respect to their language abilities. Finally, the resulting assessment device and its rating instrument together with an account of the pilot administration of the test are introduced. Conclusions have been drawn with respect to the reliability and practicality of the test.

Cette étude traite du problème de l'évaluation de la compétence orale des étudiants gradués étrangers travaillant comme assistants internationaux à l'enseignement dans le milieu anglophone canadien en général, et à l'Université de Victoria en particulier. Le problème a été étudié en considérant les besoins des assistants dans des contextes réels de compétence langagière. On s'est servi de deux normes de validité, « l'authenticité » et « l'absence d'ambiguité » (Messick, 1989), ainsi que du modèle d'évaluation conçu par Bachman et Palmer (1996). Cet article résume les résultats d'une analyse de besoins effectuée auprès de trois groupes de participants de l'Université de Victoria, à savoir un groupe d'administrateurs et de conseillers d'étudiants gradués, un groupe d'étudiants de premier cycle et un groupe d'assistants internationaux à l'enseignement. Les résultats de cette analyse ont permis d'établir des critères à inclure dans l'évaluation des compétences de communication orale. Cet article explique également la façon selon laquelle ces critères peuvent être utilisés pour créer des tâches communicatives dont l'accomplissement permet

Address for correspondence: Shahrzad Saif, Département des langues, linguistique et traduction, Université Laval, Québec, QC, G1K 7P4.
E-mail: shahrzad.saif@lli.ulaval.ca.

des inférences concernant la compétence linguistique des assistants. Finalement, on présente le test et l'instrument d'évaluation, avec un rapport du premier essai du système. Des conclusions ont pu être établies au sujet de la fiabilité et de l'aspect pratique de la situation d'évaluation dans son ensemble.

Introduction

During the past two decades, many studies have been conducted, mainly in American universities, with respect to the spoken language problems of international teaching assistants (ITAs). Having established that such standardized tests as the Test of Spoken English (TSE), the Speaking Proficiency English Assessment Kit (SPEAK) or the Oral Proficiency Interview (OPI) are not adequate for measuring the spoken language ability of ITAs (articles in Valdman, 1988; Smith, 1989; Ponder, 1991; Gokcora, 1992; Hoekje and Linnell, 1994 among others), researchers have focused on analysing the language used by ITAs in different academic settings (Rounds, 1987; Byrd and Constantinides, 1992; Myers and Douglas, 1991), the communicative frameworks and concepts that could potentially underlie ITA curriculum design and test development (Bailey, 1982, 1985; Hoekje and Williams, 1992) and interactional aspects of ITA communication (Madden and Myers, 1994; Williams, Inscoe and Tasker, 1997).

On the other hand, a survey of the ITA programs in 48 American universities (Bauer and Tanner, 1994) reveals that the majority of these institutions use the TSE as their testing device. Yet follow-up programs in the same institutions have a variety of objectives ranging from language proficiency and oral communication skills to teaching strategies and cultural issues — areas not addressed by such tests as the TSE/SPEAK or OPI.

This disagreement between the objectives of the tests and those of the programs justifies developing a testing device and a training course that are closely knit together and geared to the practical needs of ITAs in instructional settings. Furthermore, in light of the unified theory of validity which considers the consequences of test use as an aspect of test validity (Messick, 1989, 1996), it is difficult not to see the development of a valid test based on a clear understanding of ITAs' needs as a prominent aspect of ITA programs and something that should underlie ITA curricula.

Context of the Study

The context of this study is the University of Victoria (UVic), Canada, where there was a need for a testing mechanism to determine whether ITAs had an adequate level of spoken English proficiency to be able to communicate successfully in instructional contexts. This need was brought to the attention of the Faculty of Graduate Studies as a result of undergraduate students' evaluations of ITA-run classes and the recommendations of some departments and graduate advisors. There was also a need for a training program for those ITAs whose test scores indicated that they needed further training. The objective, as proposed by the Faculty of Graduate Studies, was the development of a testing instrument that:

1. would enable test users to make inferences with respect to the spoken language ability of the candidates;
2. could be administered to ITAs in different academic disciplines and with different language backgrounds;
3. could serve as an entry and exit criterion for an ITA preparation course; and
4. would influence the kind of teaching and learning that would bring about the language ability necessary for TAship. (Saif, 2000, p. 52)

It was expected that the development of such a test together with a training program would assist ITAs in the efficient use of the target language in their instructional duties. It is the former — the test development process — that will be reported on here.

As a first step in this direction, a needs assessment was conducted among stakeholders (administrators, graduate advisors, undergraduates and teaching assistants). The approach to the needs assessment and analysis was both inductive and deductive (Berwick, 1989). The techniques used for gathering subjective information were observation and interview while, on the other hand, relevant parts of Munby's (1981) framework were used in a questionnaire to gather objective information regarding ITAs' general background. The results of the needs assessment are reported in detail in Saif (2000) and will not be discussed here. In short, the information obtained from the stakeholders at UVic pointed to the fact that the problem was first and foremost a "language" problem. Administrators and graduate advisors agreed that all teaching assistants (including ITAs) were academically suitable for the type of work they were assigned to, and that past teaching experience was not a criterion for teaching assistant (TA) assignments. Observations and interviews with ITAs and administrators also revealed that familiarity with teaching techniques and strategies (or the lack thereof) is just as much an issue for ITAs as it is for native-speaking TAs. This is a point that has important implications for ITA test design and subsequent curriculum development in that it is directly related to the question of whether or not, and, if yes, to what extent, a test of ITAs' language ability should measure teaching skills.

An analysis of the language-use contexts, on the other hand, revealed that the type of discourse used during ITAs' interactions with undergraduates is by nature bilateral. It involves the speaker in both the production and comprehension of the spoken language, a complex activity which requires the employment of all aspects of communicative competence. This implies an ability beyond the basic general English level to form and interpret utterances in relation to different functions common to instructional settings, to organize them using common methods of organizing thoughts and to relate them to the characteristics of the context. Moreover, the nature of the discourse in the above contexts requires ITAs to be able to adapt their language use to unpredictable situations, such as questions that might arise in the course of communication, and thus avoid a breakdown in the transfer of meaning by making use of verbal and non-verbal communication strategies.

Test Design

The range of abilities described above could best be captured through a performance-based test that involves the test-takers in the completion of tasks that closely resemble those of the actual context(s) of language use and have the potential to elicit, as directly as possible, the behaviors revealing the abilities in question. These two qualities of test task have been respectively referred to in the literature as "authenticity" (Messick, 1994; Bachman and Palmer, 1996), and "directness" (Frederiksen and Collins, 1989; Messick, 1994), the two important validity considerations in the design of performance tests. The process of test design for ITAs should therefore include:

1. a detailed specification of the characteristics of the instructional tasks ITAs are involved in in their day-to-day encounters with undergraduate students;
2. a clear description of the constructs (that is, the complex of knowledge, skills, or component skills) to be assessed by the test; and
3. the development of test tasks that best elicit the behaviours revealing those constructs.

To that end, the model of language testing proposed by Bachman and Palmer (1996) was chosen as a guiding rationale that could specify theoretically the characteristics of the constructs and tasks that would be both direct and authentic. The broad range of components in their model of language ability — grammatical, textual, functional, sociolinguistic and strategic knowledge — can reasonably account for the presence or absence of the linguistic and communicative skills required by ITAs. At the same time, since the model specifies the characteristics of the input (the language the test-takers are exposed to in a given task) in addition to those of the expected response (the language produced by the test-takers in response to a given input), the test tasks generated based on the model would have the potential to account for the interactional nature of the discourse in instructional settings and address such areas as topical knowledge and teaching expertise without making them the primary focus of the ITA tests. The remaining part of this section will illustrate in detail how this model can be applied to the three different stages of test development mentioned above.

Task characteristics

Based on the results of the survey, “teaching undergraduate classes”, “conducting lab sessions” and “holding tutorials/office hours” were identified as the three most fundamental language-use tasks ITAs normally engage in while performing their TA duties. Therefore, in the first stage of the test development process, the characteristics of these three tasks were formalized (Appendix 1) based on the information obtained through the needs assessment and couched in terms of the theoretical model of task characteristics proposed by Bachman and Palmer (1996). The resulting specification of tasks reveals a considerable similarity in their various characteristics. The three tasks, however, differ in the physical characteristics of the setting as well as some characteristics of the input and expected response. In particular, Task One — teaching task — proved to be longer, more speeded and more textually complex than the other two. While all three of these real-life tasks are directly relevant to the purposes of the test, whether or not they are potential test tasks depends on the extent to which their characteristics represent the components of the construct definition underlying the test and contribute to the practicality of the test. I will return to this point below, in the section “Test task”.

Construct definition

In the second stage of the test design, the constructs to be measured by the test were defined. This was done with direct reference to the test’s objectives, the specific needs of the test-takers and the characteristics of the testing context. These are summarized in Table 1.

Strategic competence is included in the definition of the constructs since ITAs need to demonstrate their ability to set clear communicative goals. This is an aspect of language use that is relevant to all interactional communicative situations, including teaching contexts. Even in short private conversations, speakers need to express unambiguously the topic of the conversation and focus on the related information in order to avoid confusion and misunderstanding. Also important for ITAs is the ability to react to the communicative problems encountered in their various interactions with undergraduate students. Because their primary job is to transmit information to undergraduate students, it is very important that ITAs be able to keep communication going by resorting to the compensatory communication strategies available to them.

Topical knowledge, on the other hand, is not included in the construct definition primarily because ITAs come from different academic backgrounds and major in different areas. Furthermore, our survey showed that departments assign assistantships on the basis of a TA’s academic preparedness: TAs are assigned to courses for which they have the requisite academic knowledge. Furthermore, TAs would have the opportunity to prepare in advance to perform

Table 1: Test Constructs

Linguistic knowledge:	
Grammatical	Ability to draw upon syntactic, lexical and phonological knowledge in production of well-formed, comprehensible utterances: – knowledge of grammatical structures, accurate use of them for the purpose of communication; – knowledge of general and specialized vocabulary; – knowledge of phonological rules.
Textual	Ability to organize utterances to form an oral text: – knowledge of cohesive devices used to mark the relationships; – knowledge of common methods for organizing thoughts.
Functional	Ability to create and interpret spoken language in relation to different functions common to instructional settings: – how to use language for expressing information, ideas and knowledge (descriptions, classifications, explanations), making suggestions and comments, establishing relationships, and transmitting knowledge.
Sociolinguistic	Ability to relate utterances to the characteristics of the setting: – use of the standard dialect; – relatively formal register.
Strategic competence:	
	Ability to set goals for the communication of the intended meanings, assess alternative linguistic means (especially when there is a linguistic problem preventing the speaker/hearer from completing a default task) and to draw upon the areas of language knowledge for the successful implementation and completion of a chosen task.

all three real-life language-use tasks identified above. The test scores, therefore, are used to make inferences about the language ability of the TAs and not their knowledge of the subject matter.

A third factor considered in the definition of the constructs is the inclusion of pedagogical skills. McNamara (1996) distinguishes an individual's ability to use the target language in future jobs — the *Weak Performance Hypothesis* — from his/her ability to perform future job tasks successfully in that language — the *Strong Performance Hypothesis*. Such a distinction is of significance to the process of test development in this study since it directly affects the definition of the constructs. The results of the observations and needs assessment suggest that, unlike the area of language proficiency, native-speaking TAs do not outrank ITAs in teaching. Many native-speaking TAs are equally inexperienced in teaching, yet never tested for teaching skills. Also, if teaching skills were to be included in the test constructs, ITAs with previous teaching experience could compensate for their inadequate language skills by demonstrating teaching strategies that are not necessarily language related. More importantly, the test scores are to be used primarily to make inferences about the test-takers' ability to use *language* in a range of instructional settings in which speaking is necessary (that is, inferences based on the Weak Performance Hypothesis). Given this fact and its implications for the test's validity, acceptability and fairness, it is only reasonable that the test of ITAs' spoken language ability measure those language abilities that ITAs need to perform their TA tasks, not those abilities needed to perform teaching tasks that are unrelated to language ability. In other words, using this test, one does not evaluate ITAs for abilities that native-speaking TAs are not expected to possess. It should, however, be noted that teaching-related language skills, the lack of which would interfere with the communication of meaning in most instructional contexts, are adequately addressed by the different areas of the construct definition (strategic and textual knowledge, for example). The pedagogical nature of the communicative context is addressed further in the choice of the test task.

Test task

ITAs at UVic perform a number of activities, not all of which can be considered as possible test tasks. This is because some tasks, such as grading or materials preparation, are not directly related to the purpose of the test of speaking ability. On the other hand, some other tasks, such as those discussed in the section "Task characterisics", are relevant to the purpose of the test but for reasons of test practicality cannot all be included in the test. So, based on the specifications of the three representative language-use tasks and the existing overlap between them, the characteristics of Task One (teaching) are used as a basis for describing the test tasks. This is because the teaching task is challenging enough to measure ITAs' ability in the areas specified by the test construct. Task Three (holding office hours), on the other hand, is not long enough to tap areas of language ability (such as strategic competence). Likewise, the activity in Task Two (lab supervision) does not sufficiently cover certain areas of functional and textual knowledge. Consequently, test task specifications are summarized based on the characteristics of Task One and the definition of the constructs given above. Due to the requirements of reliability and practicality, a few characteristics of the test task (such as the presence of a video camera in the classroom, the participants and the length of the task) are different from those of the real-life setting. However, measures can be taken in order to assure that the test-taker's performance is not adversely affected by these factors. For example, the video camera can be removed for certain test-takers who express concern, or the test can be preceded by a short warm-up for the purpose of familiarizing test-takers with test administrators.

The test (Appendix 2) is, therefore, designed around a teaching task with two parts: a teaching part and a question/answer part. It not only closely simulates the natural situation of a classroom but also incorporates the basic properties of Task Three (holding tutorials and office hours). In addition, the inclusion of a question/answer part provides a better opportunity for the test-takers to demonstrate their language knowledge and strategic abilities. The scoring of the test is also affected by the construct definition. The rating instrument (Appendix 3), therefore, includes the same ability components as the construct definition. The performance of the students on each component is analyzed in terms of the levels of ability exhibited in fulfilling the test task. A five-point ability-based scale with explicit scoring guidelines (Appendix 4) is used for this purpose.

The time allotted to the whole task is 15 minutes, during which the test-taker presents a 10-minute lesson and answers questions for five minutes. Because topical knowledge is not part of the construct definition, the topic of the presentation is chosen by the test-taker. This enhances the authenticity of the task, since in real-life situations instructors determine the content of the syllabus and prepare for the class in advance.

The Pilot Study

As part of a larger impact study (Saif, 1999), the test was administered at different occasions to ITAs at UVic. The details of that study are beyond the scope of this paper. Here, I will focus on the data gathered from the initial administration of the test, on the basis of which the test's practicality and reliability were examined. Forty-seven entry-level male and female ITAs participated in the study. These students were referred to the English Language Centre by the graduate advisors of the corresponding departments. They came from different language backgrounds and specialized in different academic areas. They were at an advanced level of proficiency with a minimum TOEFL score of 550 (required by the Faculty of Graduate Studies at UVic) and an average age of 32.

Because UVic admits ITAs on the basis of their TOEFL scores without requiring proof of their spoken language abilities, in order to have a homogeneous sample, the first step was to determine that the subjects had the general oral English language proficiency required for the TA program. For this reason, the SPEAK (Spoken Proficiency English Assessment Kit),[1] a context-free standardized test, was administered to all 47 participants about two weeks before the start of the TA program. It was used as a screening device with a passing score of 220 out of 300.[2] In the next stage, about a week after the administration of the SPEAK, those ITAs who had passed the SPEAK (N = 26) took the oral performance ITA test.

Table 2: Correlations and Reliability Coefficients for the Raters and Items

Correlations	Mean	Minimum	Maximum	Reliability Coefficients
Inter-rater	.8072	.6965	.8957	α = .9505 (No. of raters = 5)
Inter-item	.8674	.7426	.9442	α = .9738 (No. of item categories = 6)

No. of cases = 26

The test was administered over a period of one week and was rated by a panel of raters comprised of two ESL instructors and three native-speaking undergraduate students suggested by the graduate advisors in the departments to which the ITAs belonged. Altogether, 15 student raters from different departments took part in the study. The undergraduate students' participation in the testing sessions is an important consideration in the design of this test, adding to the authenticity of the test task and the testing context by providing an atmosphere similar to the classroom situation. It was expected that their involvement during the presentation and question-answer phases of the test would generate a lot of spontaneous speech on the part of the ITAs from which their level of comprehensibility and success in communication could be assessed. To ensure that the raters understood the rating procedure and the areas of ability included in the rating instrument, the researcher met with the ESL instructors and potential undergraduate participants from each department, fully explained the rating procedures and provided them with copies of the rating instrument, the rating scale and a description of the ability components included in the rating instrument (Appendix 5) several days before the administration of the test. The performance of each ITA was rated either during his/her performance or shortly thereafter. Nevertheless, due to the transient nature of oral production, the entire testing session was videotaped in case the raters missed some parts of the production or major disagreements were later found in their ratings of the same individual.

Reliability Analyses

Based on the performance of the test-takers on the test, reliability analyses were conducted to determine the sources of inconsistencies, if any, among the raters and the six categories of items within the test. The results, summarized in Table 2, indicate a high level of reliability both for the categories within the test and among the raters.

However, to estimate the relative effects of multiple sources of error and the dependability of the data, a generalizability analysis was conducted. The

Table 3: G-Study Results

Source	df	SS	MS	Estimated Variance Component	Percentage of Total Variance
Persons (P)	25	186.46421	7.45857	0.2385408	63.93%
Raters (R)	4	2.37659	0.59415	0.0023978	0.64%
Items (I)	5	10.43610	2.08722	0.0142149	3.81%
PR	100	17.86874	0.17869	0.0174112	4.67%
PI	125	24.73456	0.19788	0.0247312	6.63%
RI	20	2.31249	0.11562	0.0015925	0.43%
PRI	500	37.11018	0.07422	0.0742204	19.89%

G-study design was a random effects model with two facets: raters and items with five and six conditions respectively. Table 3 shows the estimated variance components indicating the effects of different sources of error on the test score variability.

The variance component for persons is the largest while those of the raters and items are low. This means that ITAs systematically differed in their individual abilities and that the rater and item facets and their conditions have only a small effect on the test score variability. There is also a low interaction effect for raters by items indicating that the raters used the scale consistently for all items. Despite this, a small person-by-rater interaction effect tells us that the raters disagreed somewhat with respect to the ranking of the ITAs. Moreover, a slightly larger person-by-item (6.63%) interaction effect indicates that certain individuals performed better on certain items affecting the relative standing of individuals. Finally, the variance component for the residual shows that a large proportion of the variance (20%) is due to the three-way interaction between the persons, raters and items and/or other sources of variation not measured here.

Having estimated the effects of the rater and item facets, the reliability/dependability coefficients were computed. Further decision studies were also conducted with a different number of raters (Table 4).

These computations show that the degree of dependability (phi coefficient) for the D-study based on the original sample size for raters ($N = 5$) is quite high. The results of the subsequent D-studies with different numbers of raters (three and four), however, show that reducing the number of raters affects the variability of the scores very little. Therefore, if, under certain practical circumstances, the administration of the test with five raters were not possible, the test scores arrived at by using three or four raters would still be highly dependable. This option, however, is not recommended since, as mentioned earlier, a higher number of raters adds to the authenticity of the test task.

On the whole, the results of the G-study summarized above indicate that the differences among individual scores are mainly due to the differences in

Table 4: D-Study Results

	Raters = 5[a]	Raters = 4	Raters = 3
G-Coefficient	0.95946	0.95375	0.94438
Phi-Coefficient[b]	0.94839	0.94232	0.93236

[a] The first column gives the reliability/dependability coefficients for the original G-study sample size.

[b] Phi-coefficients are relevant here since the ITA test is a criterion-referenced test.

Table 5: Raters' Reaction to the Performance Test

	yes	no
Raters understood all the ability components of the rating instrument	76%	24%
Raters regarded the performance categories as adequate for measuring ITAs' spoken language ability	59%	41%
Raters believed that the test was a practical one	88%	12%
Raters believed that the 0–4 rating scale was reasonable and clear	71%	29%
Raters regarded the test task as closely related to real-life tasks	88%	12%
Raters believed that the test content would motivate ITAs to improve their spoken English	94%	6%
Raters thought that on-the-spot scoring was practical	76%	24%
Raters needed to go over the video-tape again	0	100%

individual abilities and that the rater and item effects are minimal. This implies that test's detailed rating scale and the description of the components included in the rating instrument were simple, clear and specific enough to prevent raters from subjectively scoring the test-takers' performances, and that the training procedures of the raters were effective.

Reactions to the Test and Its Practicality

During the first administration of the test, the two ESL teachers and 15 student raters were observed for their reaction towards the administration and scoring of the ITA test. A preliminary survey was also conducted among the raters and the test-takers after the administration of the ITA test and about a month before the training program began. Table 5 shows the raters' original reaction to the different aspects of the performance test and the testing process.

As can be seen from Table 5, the majority of raters believed that the test and its scoring system were practical. This result was backed by the researcher's observation of the whole rating process, during which it was noticed that all raters managed to score the test confidently during or immediately after the examinee's presentation. Despite the fact that the testing sessions were videotaped, none of the raters felt the need to review the test-takers' performances on video. It should be noted that all raters were paid on an hourly basis for the administration, scoring and, if necessary, reviewing the videotapes, so time allotment was not an issue here. In fact, once the raters became more adept at the process, they were able to go back and forth between different ability components and thus complete the rating instrument more quickly and simultaneously with the test-taker's performance.

As for the authenticity of the tasks, 88% of the raters considered the test tasks and the testing environment as closely related to the actual settings in which ITAs have to function. Observations further revealed that the undergraduate raters became involved in genuine discussions with the ITAs about the topics presented, particularly in the question-and-answer part of the test. This often provided excellent opportunities for other raters, especially the ESL teachers, to better evaluate the various ability areas in the spontaneous speech of the test-takers, as can be seen in their written comments on individual examinees' performances:

> . . . [the test-taker] mostly read from the text, but then there was a dramatic change when answering the questions . . . textual knowledge and pronunciation need some work . . .
>
> . . . didn't quite answer the questions, . . . didn't understand what they were asking . . . several attempts, . . . definitely has comprehension problems.
>
> . . . got carried away with the subject while answering the question.
>
> . . . wrote too much during the presentation, very little actual speaking until he had to answer the questions.

At this stage, the ESL teachers, who had been trained for and scored the performance of the same population on the SPEAK, were also questioned for their reaction to the SPEAK in comparison to the ITA test. In their verbatim answers to the questionnaire, they commented on different qualities of the SPEAK:

> This job has been extremely tedious and time-consuming. . . . rating takes forever because of the numerous pauses in the tape and long introductions to each section. . . . and you have to listen to each answer three times before you can decide it is incomprehensible.
>
> How can one score short correct answers relative to more complex ones?
>
> . . . sometimes the students avoided production because of their unfamiliarity with the topic being asked about, not their inability to speak in English . . . this is not fair.

Given the time they had invested on the scoring of the SPEAK, both ESL teachers believed that despite the ease of administration, scoring the SPEAK was much more time-consuming than the administration and scoring of the ITA test.

The test-takers who had participated in the preliminary administration of both the SPEAK and the ITA test before the start of the program were also surveyed for their reaction to the test. Table 6 summarizes the results.

Table 6: Test-Takers' Reaction to the Performance Test

	yes	no
Students regarded the test as more challenging than the SPEAK	86%	14%
Students regarded the test as directly related to their real-life TA tasks	63%	37%
Students regarded the performance categories as adequate for measuring their spoken language abilities	71%	29%
Students thought that the test was fair and acceptable	69%	31%
Students felt uncomfortable being videotaped and speaking in front of the panel	14%	86%
Students felt uncomfortable with the tape-recorded format of SPEAK	46%	54%
Students believed that preparation for the test would require them to improve their spoken language abilities	91%	9%

When asked about their reaction to the performance test as opposed to the SPEAK, 86% of the test-takers responded that they found it more challenging in terms of the spoken language abilities than the SPEAK. In their comments, they also added that the performance test, because of its interactive nature, provided them with a better chance to demonstrate their knowledge of English. They also mentioned that as opposed to the artificial situation created by the SPEAK, the tasks and topics in the performance test were all meaningfully connected, creating a sense of purpose during the test. Thirty-seven percent of the students, however, thought that the tasks in the ITA test were not directly related to their real-life TA responsibilities since, according to them, newly appointed TAs in their departments only did marking. Still, this group of ITAs was motivated to participate in the course to improve their spoken language abilities. A majority of test-takers (69%) also believed that the format of the test and what it measured was acceptable and fair. Interestingly, most of the learners (86%), including those with a lower proficiency level, did not express any concern about their performance being videotaped while, on the other hand, 46% of them expressed their dislike for the tape-recorded format of the SPEAK. In their comments, they described it as a "dry", "controlled", "confusing" and "unrealistic" form of measuring speaking.

Conclusion

The performance test introduced here was systematically developed based on the practical needs of ITAs in academic contexts. The primary focus of the test task is to engage the test-takers in performances similar to those of the actual instructional settings. The constructs have been defined so that the test can be administered to test-takers with different language backgrounds and different areas of specialization. The rating instrument and the detailed rating scale have proved to be relatively practical and have generated reliable test scores.

The next stage in the process will address the final two objectives listed in the section “Context of the Study”:

> 3. could serve as an entry and exit criterion for an ITA preparation course; and
> 4. would influence the kind of teaching and learning that would bring about the language ability necessary for TA-ship. (Saif, 2000, p. 52)

Thus, it is anticipated the test will indicate its potential for positively influencing the content, activities and learning outcomes of an ITA training course.

Notes

This article was presented at the 2nd Annual Conference of the Midwest Association of Language Testers held at the University of Iowa in May 2000. I wish to thank the participants at the conference, as well as three anonymous *CJAL* reviewers for their comments. This research was funded by the Faculty of Graduate Studies, University of Victoria. I am grateful to the teaching and administrative staff in the English Language Centre for their help in carrying out the project.

[1] The SPEAK is the institutional version of the TSE (Test of Spoken English) and is usually rated by trained raters at the institution administering the test. The TSE is the most common measure of spoken ability used by universities that have TA programs (Bauer and Tanner, 1994). However, both the TSE and the SPEAK are considered as indirect measures of communicative ability since they are tape-recorded tests in which the examinee’s responses are also tape-recorded. Educational Testing Service (ETS) recommends that TSE scores should not be considered the only measure for evaluating ITAs and that other relevant information should also be taken into consideration (1990).

[2] There are no passing/failing scores on the TSE. Institutions using the TSE set their own standards depending on their purposes. In this study, to eliminate candidates with lower proficiency levels, the cut-off score was set at a 60% acceptance level, which, according to the ETS Manual for Score Users (1982, 1992), is equivalent to 220 on the TSE.

References

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

Bailey, K. M. 1982. "Teaching in a second language: The communicative competence of non-native speaking teaching assistants." Ph.D. Dissertation. Ann Arbor, MI: University Microfilms.

Bailey, K. M. 1985. "If I had known then what I know now: Performance testing of foreign teaching assistants." In *Second Language Performance Testing*. P. Hauptman, R. Leblanc and M. Wesche (eds.). Ottawa: University of Ottawa Press, pp. 153–180.

Bauer, G. and M. Tanner (eds.). 1994. *Current Approaches to International ITA Preparation in Higher Education: A Collection of Program Descriptions*. Seattle: Washington University.

Berwick, R. 1989. "Needs assessment in language programming: From theory to practice." In *The Second Language Curriculum*. R.K. Johnson (ed.). Cambridge: Cambridge University Press, pp. 48–62.

Byrd, P., and J. Constantinides. 1992. "The language of teaching mathematics: Implications for training ITAs." *TESOL Quarterly*, 26, pp. 163–167.

Educational Testing Service. 1982. *Test of Spoken English: Manual for Score Users*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. 1990. *Test of Spoken English: Manual for Score Users*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. 1992. *Test of Spoken English: Manual for Score Users*. Princeton, NJ: Educational Testing Service.

Frederiksen, J. R., and A. Collins. 1989. "A systems approach to educational testing." *Educational Researcher*, 18(9), pp. 27–32.

Gokcora, D. 1992. "The SPEAK test: International teaching assistants' and instructors' affective reactions." Paper presented at the 26th Annual TESOL Convention. San Francisco.

Hoekje, B., and K. Linnell. 1994. "Authenticity in language testing: Evaluating spoken language tests for international teaching assistants." *TESOL Quarterly*, 28, pp. 103–126.

Hoekje, B., and J. Williams. 1992. "Communicative competence and dilemma of international teaching assistant education." *TESOL Quarterly*, 26, pp. 243–269.

Madden, C., and C. Myers (eds.). 1994. *Discourse and Performance of International Teaching Assistants*. Washington, DC: TESOL.

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

Messick, S. 1996. "Validity and washback in language testing." *Language Testing*, 13, pp. 241–56.

Messick, S. 1994. "The interplay of evidence and consequences in the validation of performance assessment." *Educational Researcher*, 23(2), pp. 13–23.

Messick, S. 1989. "Validity." In *Educational Measurement*. R.L. Linn (ed.). New York: Macmillan, pp. 13–103.

Munby, J. 1981. *Communicative Syllabus Design: A Sociolinguistic Model for Defining the Content of Purpose-Specific Language Programmes*. Cambridge: Cambridge University Press.

Myers, C., and D. Douglas. 1991. "The ITA lab assistant: Strategies for success." Paper presented at the Annual NAFSA Convention. Boston.

Ponder, R. 1991. "The TSE: A language teacher's critique." Paper presented at the 25th Annual TESOL Convention. New York.

Rounds, P. 1987. "Characterizing successful classroom discourse for NNS teaching assistants." *TESOL Quarterly*, 21, pp. 643–671.

Saif, S. 2000. "ITAs' spoken language needs: Implications for test development." *Speak Out! Newsletter of the IATEFL Pronunciation Special Interest Group*, Issue 25, pp. 50–58.

Saif, S. 1999. "Theoretical and empirical considerations in investigating washback: A study of ESL and EFL learners." Ph.D. dissertation, University of Victoria.

Smith, H.J. 1989. ELT Project Success and the Management of Innovation. Unpublished Manuscript. University of Reading: Centre for Applied Language Studies.

Valdman, A. (ed.). 1988. *Studies in Second Language Acquisition* 10: *The assessment of foreign language oral proficiency*.

Williams, J., R. Inscoe and Thomas Tasker. 1997. "Communication strategies in an interactional context: The mutual achievement of comprehension." In *Communication Strategies: Psycholinguistic and Sociolinguistic Perspectives*. G. Kasper and E. Kellerman (eds.). New York: Addison Wesley, pp. 304–323.

Appendix 1: Characteristics of Target Language Use Tasks

	Task 1 Teaching undergraduate courses	Task 2 Supervising laboratory sessions	Task 3 Holding tutorials/office hours
Characteristics of the setting			
Physical characteristics	Location: on-campus, well-lit classroom Noise level: normal Temperature and humidity: comfortable Materials and equipment and degree of familiarity: books, notes, handouts, blackboard, overhead projector, etc., all familiar to the test-takers.	Location: mostly science/ engineering labs, well-lit Noise level: varied, including quiet or relatively noisy Temperature and humidity: comfortable Material and equipment: varied, including lab equipment, familiar.	Location: on-campus classroom/office, well-lit Noise level: quiet Temperature and humidity: comfortable Materials and equipment: same as Task 1
Participants	Undergraduate students	Same as Task 1	Same as Task 1
Time of task	Monday–Friday, daytime, evenings	Same as Task 1	Same as Task 1
Characteristics of the input			
Format			
Channel	Oral/aural and visual	Same as Task 1	Same as Task 1
Form	Language/non-language (tables, pictures, equations, graphs)	Same as Task 1	Same as Task 1
Language	Target (English)	Same as Task 1	Same as Task 1
Length	Varied including short or long oral or written prompts and tasks	Same as Task 1	Mostly short prompts (questions)
Type	Prompt and task	Same as Task 1	Same as Task 1
Speededness	Unspeeded	Same as Task 1	Same as Task 1
Vehicle	Live and reproduced	Same as Task 1	Live
Language of input			
Organizational characteristics			
Grammatical	Both technical and general vocabulary, widely varied grammatical structures, generally comprehensible phonology	Same as Task 1	Same as Task 1
Textual	All sorts of linking devices and mostly conversational organization patterns	Same as Task 1	Same as Task 1
Pragmatic characteristics			
Functional	Ideational, manipulative (including instrumental and interpersonal)	Same as Task 1	Same as Task 1

. . .

Sociolinguistic	Variety of dialects, mostly standard Canadian English Register: formal and informal, natural language	Same as Task 1	Same as Task 1
Topical characteristics	Varied, mostly academic technical topics	Same as Task 1	Same as Task 1
Characteristics of the expected response			
Format			
Channel	Oral	Same as Task 1	Same as Task 1
Form	Language and non-language (tables, graphs, pictures, etc.)	Same as Task 1	Same as Task 1
Language	Target (English)	Same as Task 1	Same as Task 1
Length	Relatively long (50–100 minutes)	Same as Task 1	Variable (depending on the number and nature of the problem areas)
Type	Extended production response	Same as Task 1	Same as Task 1
Speededness	Speeded (certain amount of material has to be covered during the class time)	Same as Task 1	Relatively speeded
Language of expected response			
Organizational characteristics			
Grammatical	General and technical vocabulary varied grammatical structures, intelligible pronunciation	Same as Task 1	Same as Task 1
Textual	Cohesive oral text presenting well-organized pieces of information all contributing to a topic, use of common methods of development	Cohesive presentation involving a topic stated at the beginning, common rhetorical methods involve description, explanation, step-by-step analysis, etc.	Same as Task 2
Pragmatic characteristics			
Functional	Ideational, manipulative (including instrumental and interpersonal), heuristic	Same as Task 1	Same as Task 1
Sociolinguistic	Standard dialect, both formal and informal register, natural language	Same as Task 1	Same as Task 1
Topical characteristic	Academic, technical topics	Same as Task 1	Same as Task 1
Relationship between input and response			
Reactivity	Reciprocal	Same as Task 1	Same as Task 1
Scope of relationship	Broad	Same as Task 1	Same as Task 1
Directness of relationship	Mainly indirect	Same as Task 1	Same as Task 1

Appendix 2:
Test of Spoken Language Ability for International Teaching Assistants (ITAs)

General Directions: In this test, the test-takers will be required to demonstrate how well they can use English language to talk about themes and topics in their own field of specialization. The approximate time for the entire test is between 15 and 20 minutes. The whole process in sections two and three will be videotaped for the purpose of review and precision. The test will be scored by a panel of observers including three undergraduate students from the test-taker's department and two ESL instructors.

I. Introduction phase

In this section of the test, the test-takers will be required to answer some questions about themselves. The purpose of this phase, which should not last more than five minutes, is to allow the candidates to establish themselves in readiness for the main part of the test. The questions will be asked by ESL instructors and depending on the time allocated to this part, test-takers can give shorter answers to two or more questions or a longer answer to only one question.

Questions in this phase can be related to the test-takers themselves, their educational background, their home country, their interests, their hopes and future plans, the relevance of what they are doing here to their life in their country, their reasons for choosing Canada in general and UVic in particular for studying, and so forth. Test-takers will not be scored for what they say in this section since it is a quick warm-up before the main phase. This phase might be waived for those candidates who have taken the test at least once in the past or are familiar enough with the panel members and test format.

II. Presentation

In this section the test-takers will be required to present a maximum 10-minute talk related to their major field of specialization as if they were talking in front of their undergraduate students in one of the course sessions to which they are or expect to be assigned as a TA.

The subject will be a topic of test-takers' choice for which they will prepare in advance. The setting will be a classroom setting including necessary accessories such as blackboard, over-head, etc. Test-takers should be informed about all of these at least 24 hours before the test. They should also be instructed that they will be graded both on the accuracy and the appropriateness of their English as well as on how well they plan and present the idea. They should also expect questions or requests for clarification in the middle of their talk.

III. Question/Answers

In this phase, the panelists will ask questions based on the presentation in section two. The questions might require the test-takers to elaborate the original topic or to be involved in a new unprepared but related topic. The time allocated to this phase is at most 5 minutes.

Appendix 3:
Rating Instrument

Based on the test-taker's performance during phases 2 and 3, the raters will use the following rating instrument. Judgment may be based on the notes that the raters have taken during the presentation or by viewing the videotapes after the test is over. Raters should review and completely understand the ability components listed here and the rating scale before administering the test.

Name: Date: Rater:

Directions: Please circle only one number for each category.

Ability Levels	None	Limited	Moderate	Extensive	Complete
	0	1	2	3	4
Ability Areas					
A. Grammatical knowledge					
1. Vocabulary					
2. Grammar					
3. Pronunciation					
B. Textual knowledge					
4. Cohesion					
5. Conversational organization					
C. Functional knowledge					
6. Use of ideational, manipulative and heuristic functions					
D. Sociolinguistic knowledge					
7. Dialect					
8. Register					
E. Strategic competence					
9. Goal-setting					
10. Use of verbal strategies					
11. Use of non-verbal strategies					
12. Achievement of communicative goal through production					
13. Achievement of communicative goal through comprehension					
F. Overall performance					

Appendix 4:
Samples of Rating Scales

Grammar:

0 No control of grammatical rules

1 Few basic grammatical structures accurately used, limited knowledge of grammar interferes with intelligibility

2 Knowledge of a medium range of grammatical structures used with good accuracy

3 Vast knowledge of grammatical structures, few errors

4 Complete knowledge of grammar, evidence of accurate use of all structures with no limitation

Achievement of communicative goal through comprehension:

0 No evidence of understanding the language of input

1 Limited ability to relate to the audience resulting in insufficient and/or irrelevant response

2 Moderate comprehension of the language of input, occasional request for clarification or repetition

3 Extensive comprehension and interpretation of the language of input, few errors

4 Complete ability to understand the language of input, no repetition or elaboration required

Appendix 5: Description of the Ability Components in the Rating Instrument

A. Grammatical Knowledge

1. Vocabulary: control of general and field specific vocabulary, choice of semantically appropriate words
2. Grammar: control of syntactic structures and morphological rules
3. Pronunciation: including vowel and consonant sounds, and syllable stress to the extent that they interfere with the communication of meaning

B. Textual Knowledge

4. Cohesion: the use of overt linking devices and appropriate transitions which add to the clarity of expression and thus help the communication run more smoothly
5. Conversational organization: including the techniques the examinees use to open, develop, and terminate the discussion; use of common methods of organization

C. Functional knowledge

6. Use of ideational, manipulative, and heuristic functions: whether or not the utterances are appropriate for performing specific functions such as the expression and exchange of ideas and knowledge, making suggestions and comments, establishing relationships and so forth

D. Sociolinguistic knowledge: the extent to which utterances are appropriately related to the characteristics of the setting

7. Dialect: standard/non-standard English; standard English is the kind of English that educated people use in public and accept as appropriate for almost any situation. It includes formal and informal levels of language but not slang
8. Register: appropriate use of formal/informal register depending on the context of language use

E. Strategic competence

9. Goal-setting: ability to relate to the audience by using appropriate communicative goals
10. Use of verbal strategies: the extent to which the examinee makes use of verbal communication strategies either to make his/her point more forcefully or to overcome possible linguistic gaps (e.g., paraphrase, circumlocution, exemplification, . . . etc.)
11. Use of non-verbal strategies: the extent to which the examinee supplements his verbal language by non-verbal communicative strategies (e.g., gestures, pauses)
12. Achievement of communicative goal through production: examinee's ability in matching his/her communicative goals and the linguistic devices at his/her disposal to the purpose of production

13. Achievement of communicative goal through comprehension: examinee's success in understanding the verbal/non-verbal language of input (questions, comments, requests for clarification, gestures, . . . etc.)