System, Vol. 20, No. 3, pp. 317–328, 1992
Printed in Great Britain

0346–251X/92 $5.00 + 0.00
Pergamon Press Ltd

ANALYZING ORAL PROFICIENCY TEST PERFORMANCE IN GENERAL AND SPECIFIC PURPOSE CONTEXTS*

DAN DOUGLAS and LARRY SELINKER

Iowa State University; and University of Michigan

To investigate whether a field-specific oral proficiency test, constructed by manipulating test method facets, would be a better predictor of field-specific performance than a general purpose oral proficiency test, 31 Chinese chemistry graduate students were given three English tests: the field specific test, the general purpose test and a chemistry teaching performance test. Results suggested that when raters of the performance test were asked to recommend specifically whether or not a subject should be allowed to actually teach chemistry in a lab or classroom, the field-specific test was a better predictor than the general purpose test. The paper contains a theoretical discussion of field-specific language testing and guidelines for the construction of oral proficiency tests in specific purpose contexts.

INTRODUCTION

This paper is another in our continuing series of studies where we investigate the interaction between context and language test performance, using both quantitative and qualitative approaches. In a recent study at Educational Testing Service, Henning and Cascallar (1992) analyzed videotapes of subjects engaging in a number of language tasks and found that the largest amount of language proficiency variance was explained by situational variables. It has been our view for some time that language testing research is much needed in the area of interaction between context and test performance (Douglas and Selinker, 1985; Wesche, 1987), and the study we are reporting on here deals with this interaction.

Specifically, we investigated the following research question: would a field-specific test of oral language proficiency be more useful than a general purpose test in predicting subsequent field-specific performance? It would seem almost axiomatic that the answer to such a question would be in the affirmative, yet studies conducted in the area of specific purpose language testing (for example, Alderson and Urquhart, 1985; Hale, 1988; Smith, 1989; Douglas and Selinker, 1990; Shoham *et al.*, 1987, and Clapham, 1990) have failed to show clearly that such tests are superior predictors of field-specific performance. The present study is also an attempt to investigate related questions of how a field-specific oral proficiency test can be produced by the manipulation of test method facets (Bachman, 1990), and whether raters, trained to score a general purpose oral proficiency test, can score a field-specific one using the same scales.

*This is a revised version of a paper presented at the 1991 Language Testing Research Colloquium, Princeton.

THEORETICAL BACKGROUND

The *Test of Spoken English* (TSE) and its offspring, the *Speaking Proficiency English Assessment Kit* (SPEAK), produced by Educational Testing Service, are currently used in numerous U.S. colleges and universities to evaluate the spoken English proficiency of prospective international teaching assistants (ITAs), and in the health professions to evaluate the spoken English of various medical practitioners. Although these prospective ITAs and health professionals will be working in a variety of technical fields, the TSE and SPEAK are tests of general, non-technical spoken English, with the content largely unrelated to any specialized area. The candidates and their supervisors have complained that if they were tested in speaking about their own fields, they would do better. There has been some research of the validity of the TSE/SPEAK as a predictor of field-specific language use, most notably that by Powers and Stansfield (1983), which examined the TSE as a measure of communicative ability in the health professions, and studies by Smith (1989), and Douglas and Selinker (1990), both of which considered the effects of giving field-specific versions of TSE/SPEAK to ITAs. Although the Powers and Stansfield study found that judges from the health professions could make consistent ratings of "acceptable/not acceptable" spoken English proficiency and that the relationship of their ratings to TSE scores were relatively stable, it has not been shown whether a field-specific measure might provide a better basis for making judgments. The two ITA studies cited above did attempt to investigate this question, but with quite mixed and tentative results. As in the field of specific purpose language testing in general, it is far from clear what effect, if any, changes in content may have on oral test performance.

It is our view that testing language for specific purposes involves more than just changing content; specific purpose testing requires a change in discourse domain (Douglas and Selinker, 1985), which involves the language user's assessment of a communicative situation and her subsequent planning of a linguistic response to the situation, following Bachman's view of strategic competence in communicative language ability (Bachman, 1990). This change in discourse domain is brought about in turn through the language user's attention to contextualization cues, culturally conventional, highly redundant signals, such as voice tone, pitch, tempo, rhythm, code, topic, style, posture, gaze and facial expression, that interactants attend to in assessing the communicative situation (Gumperz, 1976). These kinds of cues bring us finally to a consideration of test method facets (Bachman, 1990) such as environment (personnel and location), instructions (providing domain specific reasons for carrying out the test tasks), and language (subtechnical as well as technical language, context embedded discourse, field-specific topic, authentic genre, and pragmatic and sociolinguistic domain related features). These facets need to be manipulated in specific purpose tests in order to produce a sufficient number of contextualization cues to influence the testee's assessment of the linguistic situation and thus to engage her in the specific purpose domain. This begs the question, of course, of what a sufficient number of cues might be. This is an area in which research is needed. Wesche (1987) discussed the interaction between competence and context and its influence on test validity, and called for research to determine which contextual features have what influence on performance. We believe that this is the crux of specific purpose language testing: providing, through method facets, adequate contextualization cues to trigger domain engagement.

Our present study is an investigation of performance on a general test (SPEAK) with that on a

field-specific test (CHEMSPEAK) among prospective ITAs in chemistry at Iowa State University, comparing the results with those from a teaching performance test, called, appropriately enough, TEACH. In producing CHEMSPEAK, most features of organization (salience and sequence of parts, and time allocation) and format (channel, mode and form of presentation) of the SPEAK were kept constant. Changes were made, however, in the nature of the language, in terms of propositional content (including vocabulary, contextualization, distribution of information, level of abstraction, topic and genre), and in the instructions (especially in terms of pragmatic and sociolinguistic characteristics), attempting to engage the subjects in talking about chemistry as if they were in a classroom or laboratory.

METHOD

Materials

SPEAK is a tape-recorded, seven section oral test of general English, requiring the subjects to perform tasks ranging from highly restricted (e.g. reading a paragraph out loud) to minimally controlled (e.g. answering open ended questions). Section 1, consisting of four short questions, is a warm-up section and is not scored; Section 2 involves the reading of a 100–125 word paragraph in 1 min; Section 3 asks subjects to complete 10 partial sentences grammatically; Section 4 contains six pictures that tell a simple story which subjects have 1 min to relate; Section 5 asks four questions about a single picture; Section 6 asks three open ended questions and gives subjects 1 min to answer each; and Section 7 provides some type of printed schedule (e.g. class, bus, club) and requires subjects to give the information orally. The taped responses are scored by trained raters for pronunciation, grammaticality, fluency and overall comprehensibility, the scores ranging from 0 to 3 on the first three areas and from 0 to 300 on the fourth. Scores above 2 on pronunciation, grammar and fluency, and above 200 on comprehensibility are generally considered to be "acceptable", though many institutions impose higher standards (viz. 230+) for ITAs.

CHEMSPEAK was constructed to the same format as SPEAK, but, as noted above, language and instructions were changed to engage subjects in thinking and talking as chemists in an academic situation. The read aloud paragraph, the partial sentences and the pictures were all taken from actual freshman chemistry texts. The open ended questions were devised in consultation with chemists, and the class schedule is genuine. The instructions were modified to focus the subjects on the context of chemistry teaching and provide them with some plausible reason for performing the test task. CHEMSPEAK was scored in the same way as SPEAK, by the same pool of raters.

TEACH (Abraham *et al.*, 1986) was devised at Iowa State University in order to get a better evaluation of a prospective ITA's ability to convey information to a live audience and respond to questions. Subjects register a day before taking the test and are given a topic from a list provided by the subject area departments, and a textbook in which the assigned topic appears. The following day, the subject appears at an assigned time for the test. Testing takes place in a typical university classroom and consists of three parts: (1) a minute or two for the subject to meet the "class" made up of three undergraduate questioners, two or three raters, a test proctor and a video camera technician; the subject may write terms, formulae, diagrams on the blackboard during this time, if desired; (2) 5 min to explain the assigned topic clearly and in words that an undergraduate class could understand; and (3) 3 min of questions on the topic

asked by the student questioners. Each presentation is videotaped. The raters score the presentation as it takes place, but, should they disagree, the videotape serves as a backup for additional raters. Scores are given for overall language comprehensibility (pronunciation, grammar and fluency), cultural ability (familiarity with the cultural code, appropriate nonverbal behaviour and rapport with class), communication skills (development, clarity, use of supporting evidence, eye contact, use of blackboard and teacher presence), interaction (listening and responding to questions) and overall impression. The scoring scale is similar to that on SPEAK, with scores above 2 being considered "adequate". The overall impression score is on a 10-point scale, from 0 to 9.

In this study, the TEACH results were used as criteria to compare performance on SPEAK and CHEMSPEAK.

Subjects

There were 31 prospective ITAs in chemistry in this study. All of them were Chinese, either from the PRC or Taiwan. They had been in the United States for an average of 14 months at the time of testing, though the time ranged from 3 days to 6 years. Twenty-two had an M.S. in chemistry, while the other nine had a B.S.

Procedure

The subjects took all three tests within 24 h, with the SPEAK and CHEMSPEAK back to back in alternating order, in July 1990 (20 subjects) and August 1991 (11 subjects). The ratings were completed within a week, with at least two raters scoring each protocol. The raters were asked whether they had any chemistry background and whether they were "frightened" by the prospect of scoring CHEMSPEAK. Those who indicated uneasiness were not assigned the CHEMSPEAK tapes. Fifteen different raters scored the SPEAK tapes, while 12 different raters from the same pool scored the CHEMSPEAK tapes. Where the two raters disagreed by more than 20 points on the overall comprehensibility score, a third rater scored the tape. The scores were averaged to obtain the final score. The scores were entered into a computer file and analyzed with the SPSS (1990) statistical package.

QUANTITATIVE RESULTS

First, a word about the ratings themselves. The raters had a more difficult time agreeing on scoring CHEMSPEAK than they did SPEAK; they varied from each other by an average of 25 points on CHEMSPEAK compared with an average of 16 points on SPEAK.

Concerning performance on the three tests, Table 1 shows the means and standard deviations of the totals and subscores. The overall scores are out of a possible 300, while the subscores are out of a possible 3, with the exception of the TEACH overall impression which is out of a possible 9.

Note that the mean comprehensibility score on CHEMSPEAK is lower than those on both SPEAK and TEACH, indicating that CHEMSPEAK was more difficult than the other two tests. The CHEMSPEAK and SPEAK comprehensibility scores were significantly different ($t = 2.49$, $df = 30$, $p = 0.019$), while the differences between the CHEMSPEAK and TEACH comprehensibility scores and the SPEAK and TEACH comprehensibility scores were not

Table 1. Descriptive statistics ($n = 31$)

	Mean	Standard deviation
CHEMSPEAK pronunciation	1.58	0.31
CHEMSPEAK grammar	1.79	0.33
CHEMSPEAK fluency	1.74	0.39
CHEMSPEAK comprehensibility	168.71	29.30
SPEAK pronunciation	1.65	0.33
SPEAK grammar	1.90	0.32
SPEAK fluency	1.77	0.29
SPEAK comprehensibility	180.97	27.37
TEACH pronunciation	1.62	0.35
TEACH grammar	1.84	0.33
TEACH fluency	1.89	0.35
TEACH comprehensibility	178.71	31.38
TEACH cultural ability	2.03	0.30
TEACH communication skills	2.04	0.31
TEACH interactional skills	1.97	0.43
TEACH overall impression	4.69	1.03
TEACH rater recommendation	1.52	1.03
Graduate College recommendation	1.19	0.95

statistically significant. The mean scores on all three tests are below 200, suggesting that the sample population is of generally low English proficiency. This is borne out by the recommendations made by the raters and by the Graduate College on whether the candidates should be offered teaching assignments or not: the mean ratings of 1.52 and 1.19 suggest that the majority were recommended either for no TA assignment or for lab supervision only.

Of primary interest in this study is how SPEAK and CHEMSPEAK compared in predicting performance on TEACH. Table 2 shows the intercorrelations of CHEMSPEAK, SPEAK and TEACH.

Note, first, that the intercorrelations within the three tests are fairly high, suggesting a high degree of internal consistency. Second, note that all CHEMSPEAK subscores correlate significantly with their counterparts on SPEAK (pronunciation, 0.73; grammar, 0.44; fluency, 0.69; and comprehensibility, 0.53), but in TEACH only with the subscores on pronunciation (0.50) and comprehensibility (0.43). However, the correlations of the subscores in both CHEMSPEAK and SPEAK on grammar with their counterpart in TEACH are nonsignificant (0.32 and 0.29, respectively). The same holds for the correlations for fluency (0.17 and 0.27). All the CHEMSPEAK scores show significant correlations with the TEACH raters' recommendations, while, on SPEAK, only fluency correlates significantly with the raters' recommendations. Finally, all the test scores correlate significantly with the Graduate College final recommendation, except for CHEMSPEAK fluency. Among the comprehensibility correlations, SPEAK comprehensibility and TEACH comprehensibility have stronger relationships with the final recommendation than does CHEMSPEAK comprehensibility.

Table 2. Intercorrelations ($n = 31$)

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
(1) CHEMSPEAK pronunciation	0.62**	0.49**	0.84**	0.73**	0.29	0.56**	0.63**	0.50**	0.57**	0.35	0.44**	0.40*	0.40*	0.36*	0.42	0.49**	0.69**
(2) CHEMSPEAK grammar	1.00	0.54**	0.74**	0.40*	0.44*	0.64**	0.44*	0.36*	0.32	0.20	0.47**	0.33	0.38*	0.34	0.45*	0.53**	0.58**
(3) CHEMSPEAK fluency		1.00	0.67**	0.24	0.16	0.69**	0.31	0.25	0.06	0.17	0.47**	0.27	0.22	0.36*	0.26	0.36*	0.31
(4) CHEMSPEAK comprehensibility			1.00	0.51**	0.28	0.60**	0.53**	0.33	0.30	0.28	0.43*	0.32	0.34	0.41*	0.34	0.50**	0.50**
(5) SPEAK pronunciation				1.00	0.51**	0.58**	0.78**	0.55**	0.66**	0.29	0.45*	0.31	0.25	0.16	0.26	0.25	0.72**
(6) SPEAK grammar					1.00	0.51**	0.72**	0.39*	0.29	0.17	0.35	0.28	0.23	0.19	0.27	0.22	0.47**
(7) SPEAK fluency						1.00	0.68**	0.51**	0.35	0.27	0.67**	0.48**	0.34	0.37*	0.44*	0.40*	0.63**
(8) SPEAK comprehensibility							1.00	0.44*	0.45*	0.23	0.41*	0.36*	0.31	0.23	0.31	0.34	0.68**
(9) TEACH pronunciation								1.00	0.82**	0.78**	0.85**	0.59**	0.59**	0.41*	0.59**	0.61**	0.78**
(10) TEACH grammar									1.00	0.66**	0.62**	0.30	0.34	0.10	0.33	0.37*	0.75**
(11) TEACH fluency										1.00	0.64**	0.40*	0.43*	0.33	0.40*	0.54**	0.51**
(12) TEACH comprehensibility											1.00	0.63**	0.61**	0.53**	0.64**	0.63**	0.77**
(13) TEACH cultural ability												1.00	0.57**	0.69**	0.88**	0.67**	0.57**
(14) TEACH communication skills													1.00	0.49**	0.72**	0.78**	0.56**
(15) TEACH interactional skills														1.00	0.80**	0.60**	0.49**
(16) TEACH overall impression															1.00	0.78**	0.65**
(17) TEACH rater recommendation																1.00	0.65**
(18) Graduate College recommendation																	1.00

*$p < 0.05$, **$p < 0.01$.

DISCUSSION OF QUANTITATIVE RESULTS

First, it is clear that the raters had more difficulty giving consistent ratings on CHEMSPEAK than on SPEAK. This may be because they were bothered by not knowing whether an answer was factually accurate, although none of them indicated that this was the case in their comments. There seemed to be less apprehension about scoring the CHEMSPEAK tapes than there had been about scoring MATHSPEAK in our previous study. It may be that the inconsistency was simply the result of unfamiliarity with the content of the test *qua* test, and that as the raters became more familiar with it, they would become more consistent.

The raters were invited to comment in writing on the scoring sheets. The comments on SPEAK were relatively few and dealt with pronoun gender and pronunciation. One rater commented on a subject's "good organization' and "sense of humor". Two others commented on problems with the test itself: ". . . answered the school bus question instead of the grocery bag question . . ." (i.e. the "hot house special" problem). Comments on CHEMSPEAK were more numerous and dealt more often with the content of the responses that those on SPEAK did: ". . . talked about water rather than bromine", "didn't know his chemistry here", "answers not appropriate to questions", "didn't know how to answer but kept talking anyway", "excellent command of English as it relates to chemistry!". There seemed to be some perception of difficulty in understanding the questions: "unable to explain chemistry problems in English, probably due to inability to comprehend questions", "obviously didn't understand basic terminology in English". It seems clear that raters perceived themselves to have more difficulty rating CHEMSPEAK than SPEAK and that this was related to an uncertainty about content accuracy.

Regarding the subjects' performance on the three tests, it is of interest to note that the CHEMSPEAK test was more difficult for the Chemistry graduate students than was SPEAK. Thus, at least for this group, it is *not* the case that "if they were tested in speaking about their own fields" they would automatically and universally do better. In fact, 15 of the 31 subjects scored higher on SPEAK than on CHEMSPEAK, 10 scored higher on CHEMSPEAK, and six scored the same on both tests. Four who were assigned teaching duties on the basis of SPEAK would not have been so assigned on the basis of CHEMSPEAK, while three would have been assigned on the basis of CHEMSPEAK, but failed to make the requisite score on SPEAK. Why was it apparently harder for them to talk about chemistry than about general topics? Perhaps they felt under more pressure to be factually accurate when discussing chemistry than when discussing bicycles or photography club schedules. Perhaps they were focussing more on the content and less on form in the field-specific test. Perhaps the tasks required on CHEMSPEAK were more complex than those on SPEAK.

Another factor contributing to the lower CHEMSPEAK scores may be that the raters found it more difficult to judge performance on CHEMSPEAK, since they could not be certain whether responses were factually accurate or not, and so were more conservative in their ratings. One rater has commented that there is a difference in this regard between TEACH (a field-specific test, after all) and CHEMSPEAK: raters not knowing whether a response was factually accurate or not seemed much more troublesome for them on CHEMSPEAK, perhaps because raters could ascertain the accuracy of subjects' responses from the behavior of the "student" questioners in the TEACH test. This is worth investigating, and it would be of interest to

consult both the subjects and the raters to obtain a sense of the importance of field-specific accuracy in both production and rating. Also, it would be good to consult experienced chemists about the chemical accuracy of the responses to the field-specific tests.

The correlational analysis suggests first that all three measures are internally consistent and are reasonably accurate. The patterns of intercorrelations among the tests are reasonable, though one would have liked to see somewhat stronger correlations between some of the subscores (e.g. between CHEMSPEAK comprehensibility and TEACH comprehensibility). Part of the explanation for relatively low correlations is to be found in the attenuated variance, owing to the fact that the test population had generally low scores. Certainly it would be advisable to test a group of higher proficiency subjects.

The most interesting result in the correlational analysis is that CHEMSPEAK had uniformly significant correlations with the TEACH raters' recommendations (recommended for unsupervised teaching, closely supervised recitation teaching, supervised lab teaching only, or no teaching at all) while SPEAK had a significant correlation with this rater recommendation only in its fluency score. More tantalizingly, the comprehensibility scores on CHEMSPEAK correlated with the raters' recommendations at 0.50 ($p < 0.01$), while SPEAK comprehensibility correlated at 0.34 (not significant). That is, when the raters were asked to focus finally on whether or not a subject should be allowed to go into a chemistry classroom or laboratory and teach, it was CHEMSPEAK that was the better predictor, not SPEAK. The difference in the two correlations, however, was not statistically significant, because the sample size in this study ($n = 31$) was too small ($t = 0.944$, $df = 28$, $p > 0.05$). A sample size of around 100 is needed for a difference in correlations of the present size (0.16) to reach the requisite level of significance. Given this result, the future of the present line of research is clear: continue giving CHEMSPEAK to more subjects to see whether the difference in the predictive value of the field-specific measure holds up.

THEORETICAL CONCERNS AND GUIDELINES FOR TESTING ORAL LANGUAGE PROFICIENCY

In this section of the paper we attempt to integrate language testing theory with our experience of studying interlanguage in context. To begin, we refer to Bachman (1990: p. 244) where he presents conclusions on the method effect in testing theory when he states that:

> While it is generally recognized that [specification of the task or test domain] involves the specification of the ability domain, what is often ignored is that examining content relevance also requires the specification of the test method facets.

This seems to us another point of view on what we have called in several places, the discourse domains approach to interlanguage. Namely, any factor one changes in the test environment — personnel, physical conditions, time, organization, instructions, level of precision, propositional content, etc. — can lead to changes in learner perceptions and assessment of communicative situation, and thus to changes in interlanguage performance on a test. Elsewhere (Douglas and Selinker, 1991), we have related the concept of discourse domains in SLA (Douglas and Selinker, 1985) to that of contextualization cues in ethnomethodology

(Gumperz, 1976) and hence to the concept of method facets in language testing (Bachman, 1990). We feel, too, that this addition to our conception of test methodology is important to second language acquisition research because "tests" [as defined by Carroll (1968: p. 6) as "... procedures(s) designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual"] of one sort or another are commonly used in SLA research, regardless of the particular theory being explored.

With regard to the focus on test method as a distinguishing feature of performance, we refer to a notion discussed by Bachman (1990) as "the dilemma of language testing", that language is both the object and the instrument of measurement. For Bachman, a way out of the dilemma is to understand more explicitly both the nature of language ability and the nature of test methods, so that we can "minimize the effect of test method" in the interpretation of results as indicators of language abilities. Our perspective is focussed differently: rather than attempting to *minimize* the method effects, we propose to *capitalize* on them to produce tests, useful in both language testing concerns and SLA empirical work, that would provide information interpretable as evidence of language ability in specific purpose contexts.

In the past 4 years, in interlanguage work extended into the language testing field, we have been concerned with the problem of performance on general versus field-specific tests of speaking proficiency. In an earlier study (Douglas and Selinker, 1990), we created a field-specific mathematics version of SPEAK which we called MATHSPEAK. In terms of the discourse domains hypothesis in contextually-based interlanguage research, we were trying to set up a situation where the field-specific test was conducted in a way to give the testees oral interlanguage experience at talking about mathematics during the testing session. Our methodological goal was that the only difference between the two tests would be one of domain of discourse, as defined by test method facets. In our experiment with MATHSPEAK, we found statistically significant differences between SPEAK and MATHSPEAK in grammar and fluency subscores, and a number of rhetorical differences in the responses. Where the raters claimed the differences in talk about math were grammatical, we did not find such evidence. We looked through the transcripts ourselves, and gave them to expert informants to search through. None found consistent IL grammatical differences. This is important from a testing point of view, for we need to explain the fact that the raters gave the math testees higher ratings on grammar in MATHSPEAK. We concluded that the raters were responding to something else and calling it "grammar". Thus we became wary of the subjective gross ratings on MATHSPEAK (and on SPEAK, for that matter) and carried out a rhetorical/grammatical interlanguage analysis of the test protocols. This led us to a validation principle, namely that rhetorical/grammatical interlanguage analysis may be necessary to disambiguate subjective gross ratings on tests. In applying this principle to the technical case of MATHSPEAK, we found that the knowledge of technical math allowed the IL speakers to use metaphors where the IL speakers without the technical knowledge seemingly could not produce them. Finally, on MATHSPEAK, we saw more rhetorical complexity, i.e. more embedding of content information in larger rhetorical structures by the technically competent.

This finding led us in the present study, in the domain of chemistry talk, to look at, among other features, the use of discourse markers and discourse complexity. Overall, the discourse in CHEMSPEAK was much more complex than that in SPEAK. In the SPEAK we found mostly a simple narrative joined with a succession of connectives. For example, a subject produced the

following discourse structure on SPEAK:

> declarative statement—so—and then—and—and—and.

The same subject produced the following structure on an analogous item on CHEMSPEAK:

> at x-condition—statement—so-result; but when change condition—statement—so-result; when continue to change condition—statement—so-result.

It would be noted that it was not the case that IL-talk in CHEMSPEAK was always more complex. One important misunderstanding of the discourse domains idea is the view that "talk about work" will always be more complex or target-like. What we found in this study, and what we hypothesize will be found in SLA generally, was IL difference by domain. However, in one significant case, there was very target-like talk about general things, and less target-like talk about chemistry things by a chemistry major:

SPEAK: A perfect vacation to me is to go to Miami with my family. I can play with my lovely daughter on the beach, and can also go swimming and fishing with my husband. I will pitch a tent on camping site and stay there with my husband and daughter listening to the country music. We can also go downtown in Miami. This will make me a perfect vacation.

CHEMSPEAK: The covalent bond is useful in chemistry. It can be used to express a molecule structure. In organic chemistry or inorganic chemistry we use it to express my some ideas or to talk about with teachers and so on.

This very much relates to a point made elsewhere (Douglas and Selinker, 1991) that the notion of "target-like" is independent in IL of the notion of "precision" of the IL for a particular purpose in a particular context. We feel that this is a most important and unnoticed part of the relationship between specific purpose testing and so-called "general" language testing: we make no claims about the *direction* of difference. That is, a learner may show more target-like IL features in a "general" domain and less target-like (though in some sense more "precise") features in another. We have just begun to explore the interesting question of "domain transfer" and will have more to say about this issue in subsequent studies.

As a result of our investigations, we offer the following suggestions and guidelines for field-specific oral proficiency testing:

1. Any factor test constructors change in the test domain can lead to changes in a testee's perceptions and assessment of the communicative situation, and thus to changes in interlanguage performance on the test.
2. Rather than attempting to minimize the effects of test method facets on the interpretation of results, specific purpose oral tests should capitalize on them.
3. Field-specific tests can be constructed by manipulating a number of test method facets, such as test environment (administration personnel and location), instructions (domain specific reasons for carrying out the test tasks) and language (subtechnical as well as technical language, context embedded discourse, level of precision, field-specific topic, authentic genre, and pragmatic and sociolinguistic domain related features).

4. Subjects taking field-specific tests will be more likely, as a result of domain engagement, to focus on content than on form.
5. Field-specific test tasks may be inherently more complex than those on general purpose tests.
6. The talk produced by subjects taking field-specific tests will not necessarily be more complex or even more target-like.
7. Raters of field-specific oral proficiency tests may need to have some sense of the content accuracy of responses.
8. Raters may rate field-specific oral test responses more conservatively as a function of not knowing how accurate the responses may be.
9. Rhetorical/grammatical interlanguage analysis may be necessary to disambiguate gross subjective ratings on field-specific oral tests.
10. Field-specific oral proficiency tests may provide more useful information than general purpose tests when the goal is to make field-specific judgments of subjects' oral language proficiency.

FURTHER RESEARCH AND CONCLUSIONS

Beyond the obvious requirement that future studies of this type must involve more subjects and other field-specific areas, a number of avenues might be explored. The testees need to be interviewed as they listen to their own tapes to gain some understanding of how they interpreted the input and to what extent they felt engaged as chemists rather than as learners of English. We need to have a clearer picture of what contextualization cues the testees attended to in assessing the situation and planning their response to it. Method facets, such as environment, instructions and language, need to be manipulated to discover whether there is a threshold of facets necessary to engage subjects in the field-specific area. Field-specific informants should be interviewed as they listen to or watch tapes of the responses to provide feedback on the correctness and appropriateness of information and discourse strategies; and interrater reliability needs to be studied, especially on the CHEMSPEAK, where raters expressed discomfort at not knowing whether responses were chemically accurate or not.

To conclude, in our view, there is justification in this data to continue studying this approach to field-specific oral proficiency testing. Subjects' performance on CHEMSPEAK does seem to be different from that on the general test of oral English proficiency, and the specific purpose test may inform field-specific judgments better than does the more general test. We need to obtain a clearer understanding of what those differences are and what they might suggest to us about the nature of language competence. It seems clear, too, that field-specific tests can be constructed by manipulating and capitalizing on test method facets, and that non-expert raters who are not experts in the specific field can score performance on such tests. Much remains to be learned, however, about the measurement of oral language proficiency in specific purpose IL contexts.

REFERENCES

ABRAHAM, R., KLEIN, C. and PLAKANS, B. (1986) Beyond SPEAK: testing non-native teaching assistants under classroom conditions. Paper presented to TESOL Convention, March, 1986, Anaheim.

ALDERSON, J. C. and URQUHART, A. H. (1985) The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* **2**, 192–204.

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

CARROLL, J. B. (1968) The psychology of language testing. In Davies, A. (ed.), *Language Testing Symposium: a Psycholinguistic Perspective*. Oxford: Oxford University Press.

CLAPHAM, C. (1990) Is ESP testing justified? Paper presented at Language Testing Research Colloquium, March 1990, San Francisco.

DOUGLAS, D. and SELINKER, L. (1985) Principles for language tests within the "discourse domains" theory of interlanguage: research, test construction and interpretation. *Language Testing* **2**, 205–226.

DOUGLAS, D. and SELINKER, L. (1990) Performance on general versus field-specific tests of speaking proficiency. Paper presented at Language Testing Research Colloquium, March 1990, San Francisco.

DOUGLAS, D. and SELINKER, L. (1991) Research methodology in context based second language research II. Paper presented at Conference on Theory Construction and Methodology in Second Language Research, October 1991, East Lansing, Michigan.

GUMPERZ, J. (1976) Language, communication and public negotiation. In Sanday, P. R. (ed.), *Anthropology and the Public Interest*. New York: Academic Press.

HALE, G. A. (1988) Student major field and text content: interaction effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing* **5**, 49–61.

HENNING, G. and CASCALLAR, E. (1992) *A Preliminary Study of the Nature of Communicative Competence*. TOEFL Research Report 36. Princeton, NJ: ETS.

POWERS, D. and STANSFIELD, C. (1983) *The Test of Spoken English as a Measure of Communicative Ability in the Health Professions*. TOEFL Research Report 13. Princeton, NJ: ETS.

SHOHAM, M., PERETZ, A. S. and VORHAUS, R. (1987) Reading comprehension tests: general or subject specific? *System* **15**, 81–88.

SMITH, J. (1989) Topic and variation in ITA oral proficiency: SPEAK and field-specific tests. *English for Specific Purposes* **8**(2), 155–168.

WESCHE, M. (1987) Second language performance testing: the Ontario Test of ESL as an example. *Language Testing* **4**, 28–47.