Teaching the Second-Language Testing Course through Test Development by Teachers-in-Training

by Ruth Johnson, Preston Becker, & Floyd Olive

Introduction

In an M.A. TESOL (Teaching English to Speakers of Other Languages) teacher education program, one of the courses usually included in the curriculum is a course on second-language testing. This is done for several important reasons. From the standpoint of classroom time, Andrew D. Cohen (1994) observes that "the assessment of students' language abilities is something on which teachers spend a fair amount of class time" (p. 1), while Alan Davies (1990) states that "language testing...provides goals for language teaching...it monitors...success in reaching those goals" (p. 1). Well-constructed tests can benefit ESL (English as a second language) students (1) by encouraging them in the creation of positive attitudes towards their classes and (2) by helping them to master English (Madsen, 1983).

On the other hand, Arthur Hughes (1989) asserts that "many language teachers harbour a deep mistrust of tests and of testers...too often language tests

Ruth Johnson and Floyd Olive are instructors in the Department of Linguistics, Southern Illinois University at Carbondale, Carbondale, Illinois. Preston Becker teaches in China.

have a harmful effect on teaching and learning...too often they fail to measure accurately whatever it is they are intended to measure" (p. 1). Davies (1990) agrees, noting that writings about language testing are very few, perhaps owing to the "suspicion of the quantitative methods of test analysis and the strong conviction that (language) knowledge does not reduce to mere numbers" (p. 1). Also, too often second-language teachers ignore the power of tests and of testing, especially in the area of test construction, thus giving almost exclusive attention to test scoring (Davies, 1990).

Further, Josef Rohrer (1981) states that pivotal questions about learning need to be asked by teachers/test-writers when preparing tests. These same questions pertain to the instructor of teachers-in-training enrolled in a language-testing course, questions such as "What will the teachers-in-training be doing when they leave?" (task analysis), "What do they need to learn about testing to do what they need to do?" (testing requirements), "How can the teachers-in-training find out and state clearly whether they have acquired what they need?" (criterion-referenced performance evaluation), and "How can instructors of language-testing courses help teachers-in-training to learn?" (p.32).

Many testing experts have argued for a more humanistic approach to second-language testing (Davies, 1990; Henning, 1987; Oller, 1979), for the development of a theory of language and of language testing that acknowledges second-language learning as comprised of many components and as communicative in nature (Bachman, 1991; Connor-Linton, in press). Helmut J. Vollmer (1981) argues that it is critical that test-givers "...really understand what our judgment and prediction is based upon, that the construct is as valid as are the tests designed to assess it" (p. 115). Because of the limiting nature of tests, they should not be done at a single point in time and they should be of more than one type (Vollmer, 1981). And Douglas K. Stevenson (1981) says that a language test is also a test of the linguistic theory underlying it.

John R. Ross (1979), too, notes that language testing must become more tolerant of language intricacies due to the demands of testing pragmatics and discourse, but measuring such linguistic concerns is not easy because of the difficulty in establishing norms (Davies, 1990). Both Elena Shohamy (1988) and Liz Hamp-Lyons (1991) have attempted innovative measures of speaking ability and writing ability, respectively, in a second language.

The number of textbooks written about second-language testing is large (Bachman, 1990; Brindley, 1989; Cohen, 1994; Davies, 1990; Henning, 1987; Hughes, 1989; Madsen, 1983; Shohamy, 1985; & Weir, 1990), and the literature is replete with articles written about second-language testing issues. One insightful contribution is that of Kathleen M. Bailey and James D. Brown (1996) which reported the results of a survey of instructors of language-testing courses. They found that the contents of the language-testing course were very diversified, and such diversity is a good indicator of interest in testing and should lead to more investigation and improvement within the field of language testing.

But while the content in regard to the availability of text and research resources for the students enrolled in testing courses is plentiful and diverse, the need to describe a suitable method for conducting a language-testing course still must be addressed. Mac E. Aogain (1987) suggests using simulated tasks or actual projects coupled with action research (Elliot, 1976) and data analysis (Thissen, Baker, & Wainer 1981), leading to item response theory. A responsible course on testing must include steps to illustrate how test items perform in real settings (Aogain, 1987). Also, Ulrich Raatz (1981) states that every measuring instrument must possess basic qualities, including being objective, reliable, valid, economic, useful, and quantitative.

This article describes a course in second-language testing in one such program. The M.A. TESOL program at Southern Illinois University (SIU) is a teacher education program but is housed in the Department of Linguistics in the College of Liberal Arts. Many graduates of this program are international students who return to their home countries to teach English as a second language. The majority of native-English-speaking students (American students) also leave to teach English in foreign countries. A few pursue a Ph.D. degree and even fewer find employment in ESL in the United States.

In the second-language-testing course at SIU, student teachers were taught to construct tests with a view to understanding how their test items performed in real settings. Such a test-construction exercise can be effective in illustrating for teachers-in-training the important of doing trial runs of test material because it is difficult for even experienced writers of tests to predict how particular test items will function operationally.

The Course

Seven teachers-in-training were enrolled recently in a language-testing course taught in the M.A. TESOL program at SIU. They included three women and four men, six American students and one student from St. Lucia in the Caribbean, and two students with more than ten years each of teaching experience (although none of the students had experience in teaching ESL) and five students with no prior teaching experience.

As part of the course requirements, they learned not only about test-taking strategies and test wiseness on the part of the ESL students they will teach and of the use of test results, but also about test-construction issues and test-wiseness issues that come from an informed concept of the interpretation of test results. Students in the course extensively read about and discussed issues related to language testing and then wrote items in several different formats and for varying purposes. They examined existing second-language tests, including standardized tests such as the TWE (Test of Written English), the TSE (Test of Spoken English), and the TOEFL (Test of English as a Foreign Language), and read expert reviews of these tests.

In addition, the teachers-in-training studied issues of authenticity, reliability, and validity; of individual and situational factors; and of issues surrounding item format, item elicitation, and item response, all with an aim at understanding how tests and test-taking contexts interact with the test-taker and how tests relate to each other, to the goals of a course, and to the curriculum as a whole.

While the theory of second-language testing was covered extensively, the most beneficial experience of the course, according to the students, was their writing, administering, and analyzing a test of their own making. As a class, the teachers-in-training wrote a test; different teams worked on different sections with varying formats. The test was then administered to students in two sections of an intermediate-level ESL class.[1] (Having an IEP program available in which the language-testing students could administer a test was crucial in order to make the exercise as realistic as possible.)

Description of the Test-Writing Exercise

At the outset of this exercise, the instructor took a "hands-off" approach; in other words, the concepts had been taught during the earlier parts of the course, and it was now time to simulate a test-writing experience. The instructor observed the class members as they worked on their test construction and took notes.

First, class members decided to write a test to measure reading proficiency at the intermediate level. This parallels Stage 1 of test development, as outlined by Brian K. Lynch and Fred Davidson (1994), in which a skill to be tested is selected by the test writer. In completing this stage, the students made no effort to operationalize reading proficiency; instead, they haphazardly located test items in already-existing tests, without regard for their validity or for how they fit together as an instrument.

Then, the test-writers had two considerations in writing test items: the writing of the test prompts, which were to be presented (given) to the test-taker, and the writing of the test responses, which were what the test-taker would be expected to provide as results of the prompts.

It was observed that the students gave little thought selecting/writing items for the test. For example, the first two items on the test asked test-takers to look at two sets of pictures, read sentences below the pictures, and in each set circle the picture that most closely matches the sentence. The sentences were only four words long for the first set of pictures and five words long for the second set. The instructor noted that these two items looked more appropriate for testing at the beginning level.

By contrast, Part III of the test was a modified cloze exercise. Modified clozes differ from the traditional ones in that they can be adapted to provide multiple choice responses to the cued word or they can (as in one of our cases) provide help in eliciting the correct response by providing the test-taker with the initial letter or letters of the desired answer(s). It was noted that the test-writer for this part did have the foresight to pilot his section on the instructor and his fellow students. He noticed that (1) the instructor and colleagues could not guess the "right" answer in one out of ten cases and (2) we wrote an "appropriate but wrong" answer (synonym) in five out of ten cases. Based upon these observations, the test-writer decided to supply the first letter of the "right" answer for each of the ten items.

The test items were written (or assembled from previously-written test items) and organized into a test. At this point, the class members divided themselves into teams to construct different sections of the test. The test was divided into seven subtests based upon the use of four different item-elicitation formats: matching text to a picture, true/false, cloze, and same/different comparison.

Each sub-test had a different item-response format. Section I (two items) asked the test-taker to match a sentence to a picture. Section II (four items) was a traditional true/false format keyed to an illustration. Section III (ten items) was a modified cloze. Section IV (five items) required the test-taker to read two sentences and circle if the sentences were the same or were different. Section V (five items) was similar to section II, but was keyed to a newspaper advertisement. Sections VI (two items) and VII (five items) were cloze tests with four choices given for each blank.

The test was titled "Reading Comprehension Quiz"; however, the teachers-in-training observed that it tested more than just reading. They noted that the issue of a test testing more than what it purports to be testing was an issue of construct validity: a test always tests more than it appears to test. They did not, however, debate how valid their test was for the testing of reading.

The true/false, matching, and cloze sections of the test were de facto tests of reading because the questions had to be read. On the other hand, some items may have tested cognitive functions beyond mere language ability, and often a test result could be a measure of test-taking skill as much as it could be a measure of language ability. For example, Part III, which was a cloze test based upon a paragraph about fishing in Alaska, probably tested a cultural schema as much as any language ability.

Further, the teachers-in-training did nothing to try to "match" their test to the reading section of a TOEFL test. The TOEFL reading section is comprised of several short (one to five paragraphs long) reading passages followed by questions regarding, for example, the main topic of the passage, inferring the meanings of words from context, recognizing paraphrases of facts from the passage, making cause-and-effect implications, and drawing conclusions.

Description of the Analysis of the Test

Once the test was written, two versions of the test were created using the split-half method. Sections III, VI, and VII were identical on both tests, while the other sections used similar item-elicitation formats and tested, as much as possible, language knowledge at a similar level of complexity. The two versions of the test had high reliability. When administered to two groups of the same level in an ESL program (by a teacher not directly associated with the testing class, so as to maintain the anonymity of the students), the scores were similar (t=-0.0495; probability>t=0.9610).

Group A:	91, 84, 81, 79, 79, 78, 77, 75, 74, 72, 60
Group B:	93, 87, 84, 82, 80, 79, 79, 79, 77, 72, 71, 69, 67, 65.

The class compared the results of this test to the scores the students in section B had gotten on an already established test, the TOEFL (see Table 1). This was an institutional TOEFL, administered as a regular, end-of-term assessment instrument, in the students' program at the Center for English as a Second Language (CESL). Again, the anonymity of the test-takers was maintained.

The test-taker with the highest of the fourteen scores on the class test had an overall TOEFL score higher than only three of the others; that is, the student who got the highest score on the class test ranked eleventh out of fourteen in the TOEFL scoring. The teachers-in-training were intrigued by this anecdotal evidence; they had assumed that their test would correlate closely with the TOEFL reading section.

Comparing scores on the class test to the TOEFL reading section, revealed a low level of concurrent validity. The students were asked to explain this. One reason they offered was that their test items had not been piloted before being made into a test and administered. They said they saw little value to such an exercise.

Also, the TOEFL listening scores showed a negative correlation with the class

Table 1
Section B Test Results and TOEFL Scores

Our Test	TOEFL	Reading	Grammar	Listening
93	433			
87	493	54	50	44
84	463			
82	447	40	47	47
80	463			
79	420	43	47	46
79	457			
79	470			
77	443	43	43	47
72	423			
71	487	51	45	51
69	433			
67	457	38	44	55
64	423			

test scores, with the lower scores on the class test receiving the highest score on the TOEFL listening section. The TOEFL grammar section, however, was closer in ranking to the ranking given by the test created in the class, although there were still differences. Whether or not the students' test had been tested beforehand for construct validity, it would not be expected that it would correlate with the TOEFL listening section or the TOEFL grammar section. The fact that it correlated negatively with the TOEFL listening section and did correlate with the TOEFL grammar section is not surprising.

After the tests were scored, the class as a whole examined each item to see how it had fared. The number of the item was written on the board and then the number of students that had missed it was recorded, as were the overall test scores of those students. This exercise proved revealing to the graduate students because their expectations were not met.

As was expected, some of the items on the test were good discriminators, and some were not. Two types of items did not discriminate at all: (1) Some very easy items that were answered incorrectly by everyone, such as those in section I, which asked the test-taker to look at pictures and circle the one that corresponded to the sentence prompt: "The dress is dirty" and "The tree is behind the house." (2) Some very difficult items that were missed by all of the students, such as one of the cloze items in section III: "...while sp   fishing accounts..." (The correct response is "sport.")

As an overall observation, the teachers-in-training noted that what this instrument lacked was a consistent discrimination of the lower proficiency students. Take an item in section III, which was a cloze item: Commercial f   produces..." (The correct response is "fishing.") The students with the scores 84, 81, 79, 79, 79, 77, and 75 missed it. The high scorer got it correct; however, students with scores 74 and 72 also got it right, so the item is not a good discriminator. Maybe 72 and 74 were sitting on either side of 91!

The teachers-in-training made other useful observations. For example, it was also observed that, although all test-takers answered correctly the items in Part I of both A and B versions of the test, the pictures were poor in quality. The "dirt" on the dress, for example, looked like a pattern on the material.

Also, in Part IV, three of the five items tested for the use of prepositions, which differs from the item testing for adjective use. There were questions raised as to the usefulness of testing for "shall/will," for articles, and for "go/leave." In sum, the teachers-in-training decided that the test was weak and, were it to be used, would need much rethinking and revision.

Student-Teacher Responses

This exercise was scheduled towards the end of the course and served as a tool for the integration of all that had been taught. As one student explained,

> The writing of the test was to allow us some hands-on experience with coming up

> with the testing instrument and then going through, at least to the limited degree that we were able, the process to establish the reliability and the validity of that test.

Beyond being an opportunity for the integration of course material, writing, administering, and analyzing the results of their own test was important to the teachers-in-training because of the meaningfulness of the activity. Writing and analyzing their own test brought the reality of testing home to them in ways that merely examining standardized and classroom tests had not. The meaningfulness of this exercise became clear in a series of interviews the researchers conducted with the students. One said,

> If she (the instructor) had just given us a test and said, "This is a test about that, and this is what happens with a test," then for me I would probably hear it and maybe forget it tomorrow. But having done it myself and torn it apart (analyzed it) and changed it, to me these things (testing issues) have stayed.

Another believed that having teams working on their own tests was important:

> The difficulty of test writing as illustrated in the class is in attempting to get the test to accurately fit the instruction of those being tested so you can have some type of confidence that you are doing your students justice by evaluating them on what they've been instructed.

In the interviews, all class members expressed surprise at their own inability to predict which items would be good discriminators. One student said that having the class write its own test taught her

> ...to look a lot more carefully at the way questions are asked and formulated. I think I learned a lot from doing that. I think it worked pretty well. In fact, I did not give much thought to the writing of the test items, so I learned a lot because there were surprises when the results came in.

Another student said,

> I was surprised that things (items) I thought were lousy were actually good discriminators. Like one section I didn't think was very good, but it was excellent for discriminating among the (ESL) students. It was surprising, because what I thought was bad wasn't so bad and what I thought was good wasn't so good. So I thought it was useful. And then, was the purpose of the assignment to write a good test? Would we have learned as much had we written an excellent test and then analyzed it?

The teachers-in-training were unable to predict which items were good discriminators; therefore, they found the test analysis to be both interesting and informative. For example, why would a particular ESL student miss item III.-1 and item VII.-1, both in Test A? What do the items have in common? One of the teachers-in-training said,

> There were some multiple choice questions where people had chosen a certain

> distractor that they weren't supposed to have chosen, but several people chose that answer, so (understanding) that was kind of insightful. And you really had to think, well, how was I writing this and was that what I intended to do.

This analysis resulted in the teachers-in-training looking beyond the scores to understand what it was about a test that discriminates among the test-takers. This in turn led them to a new interest in test results. One stated,

> If I was in a situation where I was doing a lot of testing and working with students, then I would definitely sit down and take a closer look at the results instead of just grading the test and handing it back.

Thus, having teacher-trainers write their own test resulted in a meaningful integration of theory and practice, well-suited to the final component of a language testing course. Surprises among the test items produced sensitivity within the test-writers, making them more careful in constructing tests. The question then became how much would the teachers-in-training transfer these test-writing strategies and test-writing wiseness issues to their own teaching and test construction? When asked how many of the issues raised in this exercise she would use in writing her own "real" tests, one teacher-in-training replied,

> I think writing the test was good, was really good in the sense that for me I was able to focus a lot. I have had to write tests, and I know I am going to have to write tests when I get back (to her home country) and so it made me more aware, more conscious of the things I need to be aware of when I'm writing a test.

There was consensus among the teachers-in-training that the class on testing promoted a sensitivity to the way tests are written, and that this sensitivity extends not only to tests that are written and administered by teachers, but also to tests that teachers take. One commented,

> It was interesting that I had just taken a test some days ago (in one of my graduate level courses). (Issues raised in this class) helped me look really closely at that test. It really did that for me.

Another student said,

> If I were taking a test there might be things about it that I would think of and consider about it that I wouldn't have before.

In a follow-up survey two years later, the teachers-in-training had the opportunity to answer whether or not they had retained the information about testing and whether or not they had used their knowledge of testing in their teaching. Of the seven teachers-in-training, three are at present in non-teaching jobs. Of the remaining four, three sent a reply to our survey.

On the retention of information, one teacher commented:

> I am embarrassed to say that I do not remember a lot of the information we learned,

> for example, terminology. What I do remember, however, was the assignment we were given to construct and evaluate a test. I learned that this was not an easy task, that it was to be done carefully, with many issues in mind. I learned to question a newly-constructed test of mine as quickly as I judge students based on the results of that test.

Another student, now an ESL teacher in the field, wrote:

> I learned that it is difficult to make a "perfect" test. One must always expect to find flaws. The directions may be unclear; answers to a multiple choice item may be ambiguous or the distractor may not serve its purpose; there will be misspelled words and "typos" that can mislead the test takers; item stems can be poorly constructed.

Finally, on the issue of the uses of norm-referenced versus criterion-referenced tests a teacher said:

> I recall that norm-referenced tests indicate how the scores of one student compare with those of other students who have taken the same test. Criterion-referenced tests indicate if the student has learned what s/he was supposed to have learned as a result of instruction. I favor using a norm-referenced test at my place of work when we have to make decisions about who to admit and there are a limited number of admissions. On the other hand to make decisions regarding the advancement of students through our program, I favor using criterion-referenced tests, although we are using norm-referenced ones at present.

Conclusions and Recommendations

Clearly, the single most important benefit of this exercise for teachers-in-training was its direct relevance for their stage of development in the M.A. TESOL program: the students had the opportunity to see in "real life" what they had been taught about testing issues and testing concerns. The exercise was meaningful and sufficiently engaging as to be memorable after a considerable period of time, while more rote items had been forgotten.

Several of the students commented that in the research on testing and in courses on testing, the importance of tests to teachers, of testing, and of test writing by teachers is undervalued. Instead, emphasis is placed on teaching test-takers strategies of test-taking and awareness of test-wiseness. As test-takers are becoming more and better informed about tests, so too do test-writers need to be more and better informed. Furthermore, at the M.A. TESOL program level, the focus must be on classroom teachers—to teach them how to write good tests—since when they enter the field and take teaching positions in various programs or situations there is no guarantee that tests will be immediately available for their materials or for the continuing situations of changing texts and the need to adapt to shifting points of focus within the curriculum.

Within the course of testing, the opportunity to write their own tests, to experience the practical side of testing, is what is meaningful to the student teachers. In addition, these students need to actually take the tests that their students will take, both standardized tests and tests that the students themselves write. They need to analyze tests. All of this reveals the strengths and weaknesses inherent in each type of test, beyond what the students read in the literature. Actually handling tests in all their stages brings to life for the students the statistical issues involved in testing, resulting in their having a better understanding of standardized tests and their uses and shortcomings.

Current trends in pedagogy are imbalanced—weighted towards the test-taker. Exercises such as the one outlined in this article address this imbalance by offering a way for student teachers to experience real test-writing, test-administering, and test-analyzing.

The course on second-language testing fits nicely into an M.A. TESOL program, especially when the design of the course focuses on relevant issues for the teachers-in-training. It offers to the graduate students a more practical experience in testing which parallels their practicum courses in teaching oral and written English and complements their more theoretical studies in such courses as Pedagogical Grammar and Theory and Methods of TESOL. Students go beyond learning about testing, and actually learn how to test; for example, learning test construction rather than just test scoring, as Davies (1990) recommends. This type of course can serve to round out their experience and provide applications to a full range of issues in the teaching (and testing) of English as a second language. With this practical emphasis on language testing, the students sent into the TESOL field are more prepared for the vagaries that their real-life teaching experiences will provide them, and the field itself will benefit by having a more aware constituency of practitioners who are more literate in the area of second-language-testing issues.

Note

1. The ESL students who participated in this exercise were given informed consent letters stating that their participation was voluntary and that their test results had no bearing on their assessment in their class in the IEP Program. All of them readily agreed to take the test and all seemed to take the test seriously.

References

Aogain, M.E. (1987). Training teachers to test communicative competence: The issues. In P.S. Green (Ed.), Communicative language testing: A resource handbook for teacher trainers (pp. 78-84). Strasbourg, Germany: Council for Cultural Co-operation.

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

Bachman, L.F. (1991). What does language testing have to offer? TESOL Quarterly, 25(4), 671-704.

Bailey, K.M., & Brown, J.D. (1996). Language testing courses: what are they? In A.

Cumming & R. Berwick (Eds.), Validation in language testing. Philadelphia, PA: Multilingual Matters.

Brindley, G. (1989). Assessing achievement in the learner-centered curriculum. Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Cohen, A.D. (1994). Assessing language ability in the classroom (2nd ed.). Boston, MA: Heinle & Heinle.

Connor-Linton, J. (in press). An indirect model of service-learning: Integrating research, teaching, and community service. Michigan Journal of Community Service Learning.

Davies, A. (1990). Principles of language testing. Cambridge, MA: Basil Blackwell.

Hamp-Lyons, L. (1991). Assessing second language writing in academic contexts. Norwood, NJ: Ablex.

Henning, G. (1987). A guide to language testing. Cambridge, MA: Newbury House.

Hughes, A. (1989). Testing for language teachers. New York: Cambridge University Press.

Lynch, B.K., & Davidson, F. (1994). Criterion-referenced language test development: linking curricula, teachers, and tests. TESOL Quarterly, 28(4), 727-743.

Madsen, H.S. (1983). Techniques in testing. New York: Oxford University Press.

Oller, J.W. (1979). Language tests at school. London, UK: Longman.

Raatz, U. (1981). Are oral tests tests? In C. Klein-Braley & D.K. Stevenson (Eds.), Practice and problems in language testing I: proceedings of the First International Language Testing symposium of the Interuniversitare Sprachtestgruppe (pp. 197-212). Frankfurt, Germany: Verlag Peter D. Lang.

Rohrer, J. (1981). Problems and practice in language testing: a view from the Bundessprachenamt. In C. Klein-Braley & D.K. Stevenson (Eds.), Practice and problems in language testing I: proceedings of the First International Language Testing symposium of the Interuniversitare Sprachtestgruppe (pp. 31-34). Frankfurt, Germany: Verlag Peter D. Lang.

Ross, J.R. (1979). “Where’s English?” In C.J. Fillmore, D. Kemper, & W.S.-Y. Wang (Eds.), Individual differences in language ability and language behavior (pp. 127-163). New York: Academic Press.

Shohamy, E. (1985). A practical handbook in language testing for the second language teacher. Ramat Aviv, Israel: Tel Aviv University.

Shohamy, E. (1988). A proposed framework for testing the oral language of second/foreign language learners. Studies in Second Language Acquisition, 10(2), 165-179.

Stevenson, D.K. (1981). Problems and practice in language testing: the view from the university. In C. Klein-Braley & D.K. Stevenson (Eds.), Practice and problems in language testing I: proceedings of the First International Language Testing symposium of the Interuniversitare Sprachtestgruppe (pp. 35-53). Frankfurt, Germany: Verlag Peter D. Lang.

Thissen, D., Baker, L., & Wainer, H. (1981). Influence-enhanced scatterplots. Psychological Bulletin, 90(1), 179-184.

Vollmer, H.J. (1981). Why are we interested in “general language proficiency”? In C. Klein-Braley & D.K. Stevenson (Eds.), Practice and problems in language testing I: proceedings of the First International Language Testing symposium of the Interuniversitare Sprachtestgruppe (pp. 96-123). Frankfurt, Germany: Verlag Peter D. Lang.

Weir, C.J. (1990). Communicative language testing. New York: Prentice Hall International.