DOCUMENT RESUME

ED 367 148 FL 021 870

AUTHOR Palmer, Adrian

TITLE The Role of Language Testing in Language Program Evaluation.

PUB DATE 91

NOTE 15p.; In: Anivan, Sarinee, Ed. Issues in Language Programme Evaluation in the 1990s. Anthology Series 27; see FL 021 869.

PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Comparative Analysis; Criterion Referenced Tests; *Evaluation Criteria; Evaluation Methods; Information Utilization; *Language Tests; Norm Referenced Tests; *Program Evaluation; Rating Scales; Research Design; Research Methodology; Scores; Second Language Instruction; *Second Language Programs; Test Content; Testing; *Test Results

ABSTRACT

A discussion of second language program evaluation focuses on the interpretability of test scores as a criterion in program evaluation. It looks at both test design and research design issues. First, eight method-comparison, program evaluation studies that compare acquisition-based and analysis/practice based methods are described. Acquisition methods are defined as those in which students are exposed to the language as a whole, anticipating that the student will acquire structure subconsciously; traditional methods also use analysis, practice, and explanation to build language competence. Then two test-design issues are examined: test content (achievement vs. proficiency) and the kind of scale (norm-referenced vs. criterion-referenced) used. The choices made in the eight studies are analyzed. Next, the effect of research design on interpretation of test scores is examined, with particular attention given to two issues: the purity of the instruction used and the background of the subjects. Two of the studies are selected as more easily interpreted than the others. The results of the studies are then summarized and compared, and implications for test development and valid use of test results in program evaluation are outlined. (MSE)

THE ROLE OF LANGUAGE TESTING IN LANGUAGE PROGRAM EVALUATION

Adrian Palmer

INTRODUCTION

First, let me say how happy I am to be here.* When I started to prepare for this talk, I recalled a paper Jack Upshur, my mentor in Language Testing presented at the SEAMEO Regional Language Center Seminar on Language Testing exactly twenty years ago. I remembered that the copy was made on a thermofax machine in pre-Xerox days, and I realized just how long I have been influenced by RELC. To be honest, in addition to my professional interest in language testing and program evaluation, I also value the opportunity to renew old friendships. This is why coming to a RELC seminar is a double treat for me.

What I plan to do today is focus on the interpretability of test scores in program evaluation studies. I will examine two main issues: test design issues and research design issues.

First, I will briefly describe eight method-comparison, program evaluation (MC-PE) studies comparing acquisition-based and analysis/practice based methods. (Since analysis/practice is both somewhat clumsy and also perhaps overly limiting, I will use the terms "traditional" or "eclectic" to refer to this method, even though these terms do not describe the basis of the method in the same way that "acquisition" does for the experimental method.)

Second, I will describe two test-design issues: the basis for the tests (syllabus vs. theory) and the choice of scales (norm referenced vs. criterion referenced). After introducing the issues, I will analyze the choices made in each of the studies to illustrate how these considerations were dealt with in actual research.

Third, I will describe two research-design issues: instructional purity and subject selection, and I will analyze the choices made in each of the studies.

Finally, I will summarize and analyze the results of the studies and suggest areas of test development that require our attention.

Fourth, I will present the results of a comparative analysis of the outcomes of the studies.

Finally, I will discuss some of the assumptions which must be made in order for the conclusions about overall results to be valid and make some suggestions for future directions in language test development and use.

THE METHODS

The eight program evaluation studies have a common theme: comparison between acquisition-based language teaching methods and traditional methods. I will define acquisition methods as those that expose the student to the language as a whole, anticipating that the student will pick up the structure, etc., subconsciously. Traditional methods are those that also use analysis, practice, and explanation in order to build overall competence.

* I would like to thank Paul Kramer for our many years of collaboration on issues related to testing the Input Hypothesis, as well as some of the specific analyses presented here. I would also like to thank Lyle Bachman for our discussions of interpretability of test scores and research results, as well as his specific comments on this paper.

In this paper, I will focus almost entirely upon a comparison of the effectiveness of acquisition based versus traditional instruction in promoting the development of general language proficiency. I would emphasize, however, that the individual studies also looked at program-specific outcomes (such as academic subject-matter learning).

The specific theory of language acquisition upon which the most of the studies were based is Stephen Krashen's Input Hypothesis (Krashen, 1985). In its strongest form, this theory states that two and only two factors are responsible for second language acquisition: comprehensible input and low affective filter strength.

The eight studies reviewed include three studies of content-based, second language instruction (Burger, 1989; Edwards, Wesche, Krashen, Clement, & Kruidenier, 1984; Hauptman, Wesche, and Ready,1988); two studies of content-based foreign language instruction (Lafayette & Buscaglia, 1985; Sternfeld, 1989); and three studies of non-content based, foreign language instruction (Asher, Kusudo, & de la Torre, 1983; Lightbown, 1989; Kramer, 1989).

TEST DESIGN ISSUES

The first issue I will investigate is the relationship between the test designs used in the studies and their interpretability. I will focus on four major test design options, involving two issues: the test content and the kinds of scales used. These options are outlined below in Figure 1.

Figure 1

Test Design Options

	Test Content	
Scales	Proficiency	Achievement
Norm referenced		
Criterion referenced		

TEST CONTENT: ACHIEVEMENT VS. PROFICIENCY

Basic Considerations

When developing or choosing tests for program evaluation, one of the first questions that arises is what to test. Berctta (1986a) describes three design patterns for testing: program specific achievement tests for each program, program-neutral proficiency tests, and a combination of achievement tests program-specific plus program-neutral measures.

The content of achievement tests is based upon a syllabus and samples what the students were taught. The strength of achievement tests is that one does not have to defend the course objectives. One has only to demonstrate that the tests cover a reasonable sample of the material taught. The weakness of achievement tests in MC-PE studies is that comparisons must be made at least partially in terms of the programs' effectiveness in covering material they were not designed to cover.

Proficiency measures are based upon a program-neutral theory of language and provide a way of directly comparing the relative effectiveness of different programs in reaching program neutral goals (but don't provide a means of evaluating the effectiveness of different programs in reaching their own specific goals). They allow us to ask the question "what is the relative effectiveness of these two programs in accomplishing thus and so?" Using proficiency tests requires us to address two major questions: What is the nature of the language competence, and what evidence do we have that the tests we are using actually measure that competence?

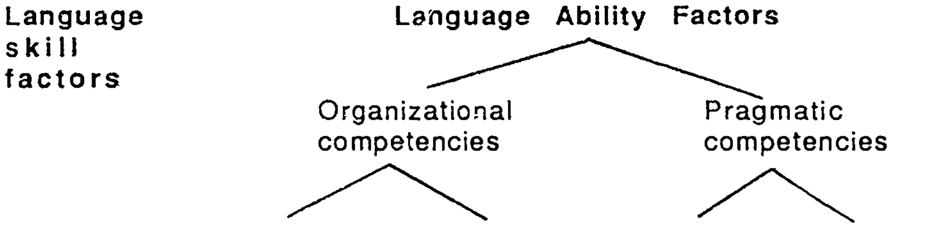
MODELS OF LANGUAGE ABILITY

The two models of language ability which seem to have attracted the most interest in the past decade are those inspired by Canale & Swain (C-S), and those inspired by Oller. These two models contrast in that the C-S model attempts to describe the various components of language ability, while the Oller model focuses primarily on the communality.

Lyle Bachman and I have worked extensively with the Canale-Swain model, and we have adopted a version which I will call the organizational-pragmatic model. In addition to these two major constructs, the model also includes the four language use skills of listening, reading, speaking, an writing.

Figure 2

Communicative Language Ability Constructs

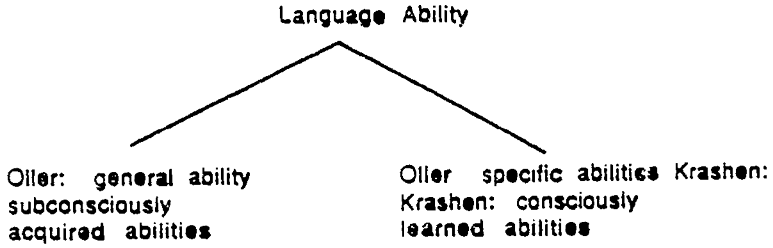


Language skill factors	Gram. comp.	Textual comp.	Illocut. comp.	Socioling. comp.
Listening				
Speaking				
Reading				
Writing				

The Oller inspired model(s) include two major constructs: a large general ability construct, and some smaller specific constructs. Krashen has sometimes interpreted these two constructs as "acquired" and "learned" competencies, an interpretation which I find reasonably compatible with Oller's.

Fig. 3

Oller (Krashen?) Model of Language Ability


Language Ability
Oller: general ability subconsciously acquired abilities
Oller specific abilities Krashen: Krashen: consciously learned abilities

EVIDENCE OF CONSTRUCT VALIDITY

Over the past decade and longer, researchers have been devoted considerable effort toward investigating the construct validity of these models. Here are four general conclusions which I believe the research supports. First, there is a distinction among the language use skills (listening, speaking, reading, and writing). Second there is a distinction between organizational and pragmatic competencies. Third, in addition to distinct abilities, a general ability factor affects all language test scores. And finally, language test scores are affected by test method (such as multiple-choice, cloze procedure, translation procedure, interactive interview, self-rating procedures, etc.) The evidence supporting these generalizations is found in a growing body of research, including Bachman 1982, Bachman & Palmer, 1981, 1982, 1989; Brütsch, 1979; Clifford, 1981; Fouly, Bachman, & Cziko, 1990; Oller 1979 & 1983; Palmer, 1972; Upshur & Homburg, 1983; and Upshur & Palmer, 1974.

The point of this is to emphasize that language testing researchers have been thinking about the nature of communicative language ability for some time and have been developing and evaluating proficiency tests based upon recent models of language ability.

ANALYSIS OF ACHIEVEMENT/PROFICIENCY CONSIDERATIONS

Given this relatively large body of language testing research on construct validity, it is interesting to examine the language tests used in MC-PE studies to investigate the extent to which they have been influenced by these developments in language testing research.

Table 1

Analysis of Tests Used in MC-PE Studies

Study	Specificity of Measures		Scoring Reference		Language Ability Model			
	Ach.	Prof.	NR	CR	Skls.	Meth.	Gen-spec./Acq.-Learn	C-S B-P
CB-SL								
Burger		•	•		•	•		
Edwards et al.		•	•		•	•		
Hauptman et al.		•	•		•	•		
CB-FL								
Lafayette et al		•	•		•			
Sternfeld		•	•		•	•		
NON-CB, FL								
Asher et al.	•	•	•		•			
Kramer		•	•		•	•	•	
Lightbown	•	•	•		•			

As can be seen from the first two columns in Table 1, two of the eight studies reviewed (Asher et al. and Lightbown) included syllabus-based achievement tests. All eight of the studies used proficiency measures reflecting distinctions among the language skills constructs (listening, speaking, reading, and writing), although not all studies used tests of all four skills.

In addition, as can be seen in the four columns under "Language Ability Model," about half of the studies also used proficiency tests classified by test method (translation, cloze, summary, and multiple choice. None of the tests used seems to have been directly influenced by the Canale-Swain/Bachman-Palmer model of communicative language ability. However, the use of cloze testing procedures in several of the studies does suggest the influence of Oller's work on general vs. specific factors, and possibly (by extension) Krashen's acquired/learned competence constructs.

In most of the studies, the tests used are named, and in some cases described, but not systematically classified by trait and method. Under Instruments, for example, we might find a list of tests such as "reading," "vocabulary" "grammar," "cloze," and "translation." Notice that such a list classifies tests sometimes by language use skill, sometimes by language ability, and sometimes by method. What we do not find are descriptions of the theory of language abilities upon which the tests were based. Nor do we find consistent distinctions made between language ability and test method.

In addition, we find almost no references to the specific language ability constructs which form the heart of Krashen's input hypothesis: acquired and learned competencies. And while many of the studies include tests commonly thought of as "integrative" (such as cloze and dictation) and "discrete-point," (such as multiple-choice grammar), reference is generally not made to the possible relationship between such tests and the primary language ability constructs in Krashen's theory.

One exception to this is Kramer, who provided a lengthy discussion of issues involved addressing the construct validity of the measures used. Analyzing the pa .ern of scores on his tests, he discussed the validity of the measures in terms of the basic constructs in Krashen's Input Hypothesis: "acquired:" and "learned" competence." While Kramer was not able to employ measures with prior demonstrated construct validity, because of the purity of his instruction (see below) he was able to assess whether the results of the research provided any evidence for the validity of the acquired/learned competence distinction.

In summary, with respect to the issue of the construct validity of the language tests used in methods comparison program evaluation studies, I believe what we see here is a general trend for such studies to employ tests that might be considered deficient in the following ways. They lag behind recent work in language testing research; they use tests which are based upon models different from those that the methods' developers had in mind when they developed their methods; and they use tests which tend to avoid the issue of the distinction between language trait and testing method.

SCALING ISSUES

Overview

I now turn to the choice of frame of reference for interpreting test scores. Norm-referenced (NR) scores are interpreted only in reference to the performance of a particular group of individuals. In contrast, criterion-referenced (CR) tests scores are "...interpreted as an indication of an individual's attainment with respect to a given domain of proficiency" (Bachman & Clark 1987:28). Each frame of reference has its own strengths and weaknesses (see Brown, 1989).

One of the main reason that norm referenced are so widely used is that they are available. And probably one reason they are so available :s that they are easy to construct. We can get away without defining what it is that we are measuring at all! People are compared to people, not to levels of ability, which means that the nature of language ability can go unspecified. This factor which contributes to their ease of construction is also, of course, one of their main weaknesses. Another weakness of NR scales is that they do not provide us with a measure of how much of a body of knowledge one controls. Thus, they do not provide us with the kinds of information we might want in assessing the relative importance of attained levels of ability.

The main strength of criterion referenced scores of language ability is that ratings are comparable across a wide range of contexts and content areas (Bachman & Savignon, 1985), which is precisely what Bachman (1989) suggests would be useful in MC-PE studies. Criterion referenced scores would allow us to address very interesting issues, such as the amount of given ability that students have mastered. This, in turn, would allow us to assess the importance of that level of mastery.

The main weakness of these scales is the practical difficulty in constructing them for tests of general language ability. Specifically, one needs to define language ability precisely, to keep as distinct as possible the roles of language ability and test method, to keep distinct language ability and context, and to specify zero and complete mastery levels.

Bachman and I tried to define such scales in our work, and Bachman and Clark (1987) have suggested a general program to further develop, refine, and operationalize such scales. Much work remains to be done, both on the conceptual and operational level, before we have available a battery of CR language ability measures, but, as I hope to show, given the kinds of outcomes we are getting in our much of our MC-PE research, and given the difficulties presented in interpreting these outcomes, I think this kind of research and development work is warranted.

ANALYSIS OF THE SCALE OPTIONS USED IN THE STUDIES

All of the studies reviewed used NR scales (see "Scoring Reference" in Table 1). Thus, while we can say that one group of students performed better than another group, we cannot say how much of the language either group controlled. This means that we cannot reach conclusions about the importance of the levels of competence reached, nor about the relative effectiveness of each program in reaching its own unique objectives.

Some of the MC-PE researchers have noted this problem and provided additional information to make the results more interpretable. Both Kramer and Sternfeld provided examples of what the students were able to do after completing the program of study. Kramer also provided descriptions of student performance.

The need to go beyond the typical NR comparative statistics became particularly obvious to me when I attended a meeting between a dean at the University of Utah and a group of researchers. The Dean took one look at the summary statistics and immediately said, "Setting the differences between groups aside, just how much of the language have these people learned in one year of instruction?" The Dean was more concerned with the amount the students had learned than with what appeared to be minor (though possibly statistically significant) differences between groups.

RESEARCH DESIGN ISSUES

So far, I have examined the effect of test design on interpretation of the results of the research. Now, I turn to the influence of the research designs and their effect on interpretation of test scores. Specifically, I will examine two design issues: the purity of the instruction used and the backgrounds of the subjects involved in the studies. While these are by no means all of the relevant internal validity considerations, they strike me as being particularly important in MC-PE research.

INSTRUCTION PURITY

Overview

Instructional purity is the extent to which the treatments of the two groups of subjects are faithful to the theories on which they are based. Instructional purity affects the internal validity (Beretta 1986b, Brown 1988) of the study, which is the extent to which we can attribute results (as measured by tests) to differences in treatment. Studies in which such attributions can be logically made are said to exhibit a high degree of internal validity.

If we wish to compare the effectiveness methods based upon different principles, such as acquisition versus analysis/practice we need to start out with definitions of the methods based upon the principles, "descriptive data" (Richards & Rodgers, 1986: 181-3). According to Krashen's definitions, language acquisition is the result of comprehensible input and low filter strength, and nothing else. This indicates what must be provided in "acquisition" classes. In contrast, "traditional" or "eclectic" instruction might be operationalized as instruction which also provided conscious learning, drills, production oriented activities, practise, explanation, analysis, etc. (along with comprehensible input).

Once we have defined the theoretical bases for and differences between the methods being compared, we need to look for some sort of evidence (observational data) that the activities that took place in the classroom were faithful to these definitions and distinctions.

ANALYSIS OF THE STUDIES WITH RESPECT TO PURITY OF INSTRUCTION

Paul Kramer and I analyzed the reports of the eight MC studies to determine the extent to which conscious learning and production activities were included in the experimental (acquisition based) classes, these being the primary activities which are said to contribute to learning, but not to acquisition (Kramer and Palmer, 1990). The results of our analysis are given in the first two columns of Table 2.

Table 2

Analysis of Factors
Affecting the Internal Validity of the Studies

	Learning Activities		Stud. Background
Study	Consc. Learn.	Production	(Prior Instructio ,
CB-SL			
Burger	yes	yes	advanced
Edwards et al.	no gram. voc (?)	yes	high intermediate
Hauptman et al.	no gram. voc (?)	yes	high intermediate
CB-FL			
Lafayette et al.	no	yes	fourth course
Sternfeld	no	some	some beginning
NON-CB, FL			
Asher et al.	no	yes	beginning (?)
Kramer	no	no	beginning
Lightbown	no	no	beginning

As can be seen in columns 1 and 2 under "Learning Activities" in Table 2, most of the studies seem to have avoided the use of conscious learning activities in the "acquisition" treatment, perhaps with the exception of the conscious teaching of vocabulary. On the other hand, most of the studies seem to have provided situations, such as discussion groups, in which the students were expected to produce the language, as opposed to just processing input. While Krashen's Input Hypothesis certainly does not include a rule prohibiting speech, it does specifically state that speaking is not a cause of language acquisition.

To test Krashen's theory, it seems to me that we must try our best to keep it as distinct as possible from other theories. The two main differences between Krashen's theory and others seems to me to lie in the two negatives: neither conscious learning nor production are required for acquisition. Thus, If we include either of these in the method which is supposed to be an operationalization of Krashen's theory and not other theories, it is difficult to claim that the two methods are distinct.

On the whole, it appears only two of the eight studied (Kramer & Lightbown) provided the students with relatively pure operationalizations of acquisition-based instruction, as defined in Krashen's theory. The research reports generally did not include descriptions of the traditional instruction used for the control groups, but I think it is reasonable to assume that this instruction was eclectic enough to be distinct from the narrower range of activities found in the studies using relatively pure acquisition-based instruction. Nevertheless, the fact that the traditional instruction is not carefully described is a weakness of the reports of these studies.

STUDENTS BACKGROUND

Overview

Another research design factor affecting the interpretation of test scores in MC studies is the nature of the abilities that the students bring with them to the study. If we are comparing the relative effectiveness of two methods of language teaching, and if amount of treatment to which we expose them is small relative to the total amount of prior instruction they have received, and if the testing indicates some sort of positive outcome, it would still be risky to advocate the experimental treatment as the basis for all of a student's language learning activity. It would also be risky to infer that the experimental treatment alone (and the treatment upon which it was based) explained the outcomes obtained.

Students at intermediate or advanced levels of language proficiency might be presumed to have been exposed to a fairly wide range of language learning activities. Adding even a fairly narrowly focussed type of instruction (such as comprehensible input) might have fairly little effect on the overall range and quantity of language learning activities to which these students were exposed; and a method narrowly defined as containing only what was added (such as comprehensible input) would not resemble the total range of instructional activities affecting the results of the research.

Moreover, particularly if you do not employ random assignment to groups, you are likely to run into problems caused by differential backgrounds between the two groups. (This and other design problems are dealt with in some detail in Kramer & Palmer, 1990).

ANALYSIS OF THE STUDIES IN TERMS OF STUDENTS' BACKGROUND

The results of Kramer's and my analysis of the eight studies is given in the third column of Table 2. This indicates that the students were about equally divided between those at the intermediate-advanced level and those who were relative beginners.

So far, I have discussed test and research design and their effects on the interpretation of results, and I have noted that two of the studies (Lightbown's and Kramer's) seem to be more interpretable than others. I now turn to the outcomes of the studies.

QUANTITATIVE RESULTS

When one employs a treatment as radical as a "pure" implementation of acquisition based instruction (no conscious learning, no focus on form, no production activities) with groups of students differing markedly in initial language proficiency, as well as age, it is reasonable to hypothesize that there would be significant interaction between treatment and level, or between treatment and age. Such interaction might render invalid any global interpretation of the treatment as "effective" or "ineffective." I will now present a comparative analysis of the results of the studies, which, I believe, illustrate just this sort of interaction.

A between-group comparison of end-of-treatment scores on proficiency tests is given in Table 3 (Kramer & Palmer, 1990).

Table 3

Between-Group
Post-Test Proficiency Comparisons

	TRADITIONAL GROUP SIG. BETTER	NO SIGNIFICANT DIFFERENCES	ACQUISITION GROUP SIG. BETTER
"Pure" Studies	(Adult L2) K-oral int. K-reading trans. K-writing summ. K-vocabulary (4 outcomes)	K x 3 (3 outcomes)	(Child L2) Li-vocabulary Li-pictures Li-speaking* (3 outcomes)
"Impure" Studies	H-cloze S-writing La-reading La-writing (4 outcomes)	H x 11 S x 4 E x 7 La x 1 B x 5 (28 outcomes)	H-translation H-total prof. E-cloze L-speaking (4 outcomes)

NOTES: *Between-group differences on Lightbown's speaking tests were large but were not tested for significance (small N).
No post-test *proficiency* comparisons provided in Asher et al. so no outcomes for this study are included in this table.

The table is constructed to call attention to the interaction between students' age, purity of treatment, and effectiveness of the methods.

Within cells are comparative post-instruction proficiency test outcomes (or gains in those studies employing ANCOVA's with pre-test scores as covariates), designated by the initials of the first author (E = Edwards et al, La = Lafayette, etc.) and test content (speaking, vocabulary, etc.). In the top row are the outcomes for the two relatively "pure" studies. In the bottom row are the data points for the six relatively "impure" studies. In the left column are outcomes significantly favoring traditional treatments. In the center column are outcomes for which no significant post-instructional proficiency differences were found. In the right column are outcomes significantly favoring the experimental (acquisition-based) treatment.

Notations such as "K x 3" (as in the top center cell) indicate that no significant differences were found on three of the post-test proficiency measures in the Kramer's study.

In a few studies, we had to make arbitrary decisions as to whether to include both part and whole test scores as data points, so others who might analyse these studies on their own might arrive at slightly different totals from those than presented here. I believe, however, that the overall trends would likely be the same.

Our first general observation is that in the two "pure" studies, overall effectiveness of instruction was related to the students' age. The students in Kramer's study were adults, while those in Lightbown's study were children. Kramer's MANOVA indicated that the traditional students performed significantly better than the acquisition students both overall and on four of the seven tests. The experimental students in Lightbown's study performed better on all three proficiency measures, significantly better on two of them. Due to the small number of students taking the speaking test, no tests of significance were performed, although the differences appear to be large.

In the "impure" studies, there appears to have been no significant main effect (type of instruction). In addition, most of the comparisons between groups on individual tests indicate no significant differences: 28 non-significant differences versus 8 significant differences (four favoring the traditional group, four favoring the acquisition group). In addition, on those individual tests for which significant differences were found, I do not see any obvious interaction between treatment and specific language ability. For example, on the cloze test in Hauptman et al., the control group outperformed the experimental group; whereas in Edwards et al., the experimental group outperformed the control group.

Normally, when one encounters a large number of non-significant differences (as was the case for the "impure" studies), one would be concerned about the reliability of the measures. With unreliable tests, one would find non-significant between-group differences over and over, and conclusions that one group performed "at least as well" as another group would be meaningless. In addition, statistical logic requires that we first reject the null hypothesis that no learning took place before drawing a conclusion that two groups performed comparably.

Test reliability does not appear to have been a problem in these studies. Most of the studies included evidence of test reliability. And the null hypothesis of no learning has also been addressed, although the fact that some of the studies were conducted in a second-language environment tends to make rejecting the null hypothesis somewhat more problematic.

DISCUSSION

The analysis of the language testing outcomes presented above appears to point to fairly straight forward and strong conclusions both about language acquisition theory and about method. It suggests that theories of child and adult language acquisition are different. It suggests that different methods work for children and adults. It suggests that balanced methods are more effective for adults than methods with a narrow focus. And it suggests that more than comprehensible input (with low filter) is needed for efficient adult L2 acquisition. Yet I would immediately like to caution against taking overly strong positions about the validity of these inferences. Specifically, I would like to point out some of the limitations in our testing procedures which ought to raise caution flags at a number of points.

First, the conclusion that certain programs seem more or less efficient than others in teaching language depends upon our confidence that the tests adequately sample what we believe to be the important components of language ability. If the tests are biased toward one program or another, the results will also be biased. It strikes me that one of the best ways to avoid bias in the selection of program neutral tests of general language ability is to use tests which are clearly related to a theory of language ability, preferably one which has been validated in independent research. If the selection of tests is at all haphazard, to that extent the results of the tests might be expected to be biased.

Second, if we observe significant differences between groups on measures of language ability and make anything of these differences, we are assuming that significant differences are also meaningful differences. As J. D. Brown points out, significance and meaningfulness are two different issues and must be addressed separately (Brown 1988: 141). As long as we continue to use norm-referenced scoring procedures, I am afraid that we will find it difficult to obtain the kind of information we need to distinguish the significant from important.

Third, because the analysis of the eight studies presented in this report addresses the issue of the program's effectiveness in promoting language acquisition, one might interpret this as a comment on the general value of the programs. This is clearly an unwarranted interpretation. Immediate gains in language proficiency are only one possible measure of a program's value. The researchers, however, were interested in other outcomes as well. For example, in the sheltered subject matter programs, mastery of the subject matter was an important consideration. Also, many universities are interested in promoting area studies programs, and based upon evidence from students' journal entries, students in the University of Utah's acquisition-based programs seem to have become increasingly aware of the L2 culture. What is the relative importance of this outcome compared with the development of language proficiency? And what is the relative importance of affect and attitude, variables measured in many of the studies reviewed?

Another measure of a program's success might be the extent to which its students continued on with additional language study. For example, follow up observation of subjects in Sternfeld's study indicates that a much larger percent of students in the immersion program continued on to more advanced courses. In this case small, (but possibly significant) between-group differences in language ability at the end of one year might prove inconsequential in the long run. If students are a little better than their peers at the end of one year of language study but then quit, within a couple of years this initial difference would be meaningless. After all, length of exposure is an important variable in language acquisition.

Additionally, the apparent clarity of the findings hinge to a considerable degree on two studies: Lightbown's and Kramer's. Just how confident should we be that these findings would be replicated? They might be, but again they might not. Should we be more confident of the patterns observed with a relatively large number of replicated."impure" studies than with a few unreplicated "pure" studies?

CONCLUSIONS

In the 12th annual Language Testing Research Colloquium held in San Francisco, in March of 1990, both Lyle Bachman and I expressed a fear that we as language testing researchers were becoming isolated from other users of language tests, and as a result what we are learning about language and language tests is not having an effect outside of our interest group.

I see the testing components of the eight program-evaluation studies reviewed in this paper as evidence that this fear is justified. I have experienced first hand the practical problems we face in trying to put to use what we have learned in our research. I consulted in both the Kramer and Sternfeld studies and had ample opportunity to influence the testing efforts, yet both of these studies were conducted by under conditions which would have made it difficult to develop or use the kinds of testing designs and procedures being advocated in the field of language testing research. Moreover, individual researchers naturally have their own testing interests and agendas, and who is to say that ours are more important than theirs?

If we as language testers are to have an impact on the use of language tests in program evaluation, or anywhere for that matter, we have to make it practical for others to use what we have discovered. We cannot expect researchers whose interests may be primarily in methodology or theory to find the time and develop the expertise necessary to create new, practical, construct valid criterion-referenced tests of communicative language ability. We need to do this development work on our own and then make tests of this sort available to others.

REFERENCES

Asher, J., Kusudo, J., & de la Torre, R. 1983. Learning a second language through commands: the second field test. In Oller, J. & Richard-Amato, J. Methods That Work: A Smorgasbord of Ideas for Language Learners. Rowley, Mass.: Newbury House. 58-72.

Bachman, L. 1982. The trait structure of cloze test scores. TESOL Quarterly, 16, 61-70.

Bachman, L. 1989. The development and use of criterion-referenced tests of language ability in language program evaluation. In Johnson, R. (ed.). The Second Language Curriculum. Cambridge: Cambridge University Press. 1989.

Bachman, L., and Clark, J. 1987. The measurement of foreign/second language proficiency. ANNALS, AAPSS, 490. March 1987.

Bachman, L., and Palmer, A. 1989. The construct validation of self ratings of communicative language ability. Language Testing, 6.1.

Bachman, L., and Palmer, A. 1982. The construct validation of some components of communicative proficiency. TESOL Quarterly, 16.4. 449-463.

Bachman, L., and Palmer, A. 1981. The construct validation of the FSI Oral Interview. Language Learning, 31.1. 67-86.

Bachman, L., and Savignon, S. 1985. The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview and suggestions for its revision and development. Paper presented at the "Perspectives on Proficiency" Forum, 1985 Modern Language Association of America Convention, 29 December 1985, Chicago, Ill.

Beretta, Alan. 1986a. Program-fair language teaching evaluation. TESOL Quarterly, 20.3. 431-444.

Beretta, Alan. 1986b. A case for field experimentation in program evaluation. Language Learning, 36.3. 295-309.

Brown, J. D. 1988, Understanding Research in Second Language Learning. Cambridge: Cambridge University Press. 36-40.

Brown, J. D. 1989. Language program evaluation: a synthesis of existing possibilities. In Johnson, R. K. (ed.) The Second Language Curriculum. Cambridge: Cambridge University Press. 222-241.

Brütsch, S. M. 1979. Convergent-discriminant validation of prospective teacher proficiency in oral and written French by means of the MLA cooperative language proficiency test, French direct proficiency tests for teachers (TOP and TWO), and self-ratings. Unpublished Ph.D. dissertation, University of Minnesota.

Burger, S. 1989. Content-based ESL in a sheltered psychology course: input, output and outcomes. TESL Canada Journal/Revue TESL du Canada, 6:2, 45-59.

Clifford, Ray T. 1981. Convergent and discriminant validation of integrated and unitary language skills: the need for a research model. In Adrian S. Palmer, Peter J. M. Groot and George A. Trosper, eds. The Construct Validation of Tests of Communicative Competence. Washington, D. C.: Teachers of English to Speakers of Other Languages. Pp. 62-70.

Edwards, H., Wesche, M., Krashen, S., Clement, R., & Kruidenier, B., 1984. Second-language acquisition through subject-matter learning: a study of sheltered psychology classes at the University of Ottawa. The Canadian Modern Language Review, 41.2. 268-282.

Fouly, K., Bachman, L., and Cziko, G. 1990. The divisibility of language competence: a confirmatory approach. Language Learning, 40:1. 1-21.

Hauptman, P., Wesche, M., and Ready, D. 1988. Second-language acquisition through subject-matter learning: a follow-up study at the University of Ottawa. Language Learning, 38.3. 433-475.

Kramer, P. 1989. The Classroom Acquisition of German and the Input Hypothesis. Salt Lake City: University of Utah Ph.D Dissertation.

Kramer, P. and Palmer, A. 1990. Comparative program evaluation studies as tests of the Input Hypothesis. Paper presented at the TESOL '90 Convention, March 6-10. San Francisco.

Krashen, Stephen. 1985. The Input Hypothesis: Issues and Implications. London: Longman, Inc.

Lafayette, R. and Buscaglia, M. 1985. Students learn language via a civilization course--a comparison of second language classroom environments. Studies in Second Language Acquisition, 7.3.

Lightbown, P. 1989. Can they do it themselves? a comprehension-based ESL course for young children. To appear in the proceedings of the Conference on Comprehension Based Second Language Teaching: Current Trends. University of Ottawa, May, 1989.

Oller, J. 1979. Language Tests at School. London: Longman, Inc.

Oller, J. (ed.). 1983. A consensus for the eighties? In Issues in Language Testing research. Rowley, Mcss.: Newbury House. 351-386.

Palmer, A. 1972. Testing communication. IRAL 10.1. 35-45.

Richards, J., & Rodgers, T. 1986. Approaches and Methods in Language Teaching: A Description and Analysis. London: Cambridge University Press.

Sternfeld, S. 1989. The University of Utah's immersion/multiliteracy program: an example of an area studies approach to the design of first-year college foreign language instruction. Foreign Language Annals, 22.4. 341-354.

Upshur, J., and Homburg, T. 1983. Some relations among language tests at successive ability levels. In Oller, J. (ed.). Issues in Language Testing Research. Rowley, Mass.: Newbury House. 188-202.

Upshur, J. and Palmer, A. 1974. Measures of accuracy, communicativity, and social judgements for two classes of foreign language speakers. Selected Papers from the Third International Congress of Applied Linguistics, vol. 2. Heidleberg: Julius Groos Verlag.