DOCUMENT RESUME

ED 243 943 TM 840 256

AUTHOR Yap, Kim Onn
TITLE Evaluating a Bilingual Test: Adding the Counsumer's Point of View.
PUB DATE Apr 84
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Secondary Education; English (Second Language); Evaluation Criteria; Language Proficiency; *Language Tests; *Limited English Speaking; Mainstreaming; Participant Satisfaction; Questionnaires; Teacher Attitudes; Test Reliability; *Test Reviews; *Test Use; Test Validity
IDENTIFIERS *Basic Inventory of Natural Language; Inventory of Test Use Satisfaction; Test Retest Reliablity

ABSTRACT

The purpose of this study was to review and evaluate the Basic Inventory of Natural Language (BINL) to help determine whether it should continue to be used in Honolulu's Students of Limited English Proficiency (SLEP) Program. The study looked at four critical aspects of the BINL: its content validity, its test-retest reliability, its effectiveness as a measure for exiting project students in terms of the students' post-SLEP performance in the regular classroom, and the satisfaction of the project staff with its use. All the BINL test items related to some of the SLEP instructional objectives, but not all of the SLEP objectives are measured by BINL items. The test-retest correlation in terms of BINL raw scores was .88. On the average, the exited students appeared to be doing more than satisfactorily in the regular classroom in all subject areas. The test use survey indicated that there was a high degree of support and enthusiasm on the part of the SLEP Program staff for the use of the BINL. (BW)

Evaluating a Bilingual Test: Adding the Consumer's Point of View

Kim Onn Yap

Northwest Regional Educational Laboratory

A paper presented at the annual meeting of the American Educational Research Association New Orleans, April 23-27, 1984

Evaluating a Bilingual Test: Adding the Consumer's Point of View

INTRODUCTION

Since the spring of 1979, the Honolulu district in Hawaii has been using the Basic Inventory of Natural Language (BINL) to measure English language proficiency of students participating in the Students of Limited English Proficiency (SLEP) Program. The test was selected by the district as an instrument for screening, diagnosis, placement and evaluation. The purpose of the present study was to review and evaluate the BINL to help determine whether the continued use of the test in the district is warranted.

The SLEP Program is designed to serve students whose dominant language is not English and whose limitation in the use of English prevents them from functioning effectively in the regular classroom. The overall objective of the program is to help these students to adjust to the American culture in the Hawaiian setting by acquiring basic communication skills to participate in the regular classroom instruction and school activities appropriate for the their age and grade level.

Students are selected to participate in the program on the basis of their language dominance ratings as determined by criteria specified in the Identification Assessment Programming System (Hawaii Department of Education, 1980). Only students who receive language dominance ratings of 1 and 2 are eligible to participate in the program. Participants are exited from the program when they reach a language dominance rating of 3 (or above) and score at the 25 percentile (or above) on the Metropolitan Achievement Test in reading, language arts, and mathematics.

Currently, the SLEP Program is offered in 55 schools in the district, serving some 5,000 students. Program staff include 7 permanent teachers, 51 temporary teachers, 8 school home aides, 15 educational assistants, and 47 part-time temporary teachers. Program funding is approximately $1.3 million per year.

The BINL purports to measure language proficiency in terms of the complexity of the language used by students in grades k-12. The test items consist of 40 large story starter pictures. The student selects 3 to 5 pictures and responds to the items by making up stories about the pictures or about some of the people and things in the pictures. The pictures may be used as a jumping off point for the student to speak about a personal experience. Student responses are tape recorded to facilitate scoring.

The BINL scores are based on 10 speech samples taken from the student during the test. The test provides a series of scores which may be used to build a language profile for each student. Fluency is indicated by the total number of words used by the student during language sampling. This score is said to be a general indicator of ability to use the vocabulary, structures and forms of a language. Level of complexity is a measure of the student's command of the structures of the language, including the use of modifiers, phrases and clauses. Average sentence length is a measure of the fluency count and the number of phrases or sentences used by the student.

The BINL complexity level scale ranges from 0 to 200. Average sentence length ranges from 0 to 15 words. Based on score ranges, language proficiency categories are established. These include non-English speaking, limited English speaking, fluent English speaking,

and proficient English speaking. Separate score ranges are used to establish the categories for different grade level groupings (e.g., k-2, 3-6, 7-8, 9-12). Students in the early grades are generally not expected to reach high school levels of complexity.

The BINL is administered individually. Test administration generally takes 10-15 minutes. The test is both hand- and machine-scorable.

PROCEDURES

The study looked at four critical aspects of the BINL. First, the validity of the test, primarily its content validity, was examined. Second, the test-retest reliability of the BINL was assessed. Third, the effects of the BINL as a measure for exiting project students was studied in terms of the students' post-SLEP performance in the regular classroom. Fourth, a test use survey was conducted to find out how satisfied the project staff were with the use of the BINL. Specific procedures used in the study are described as follows:

A. Test Review Committee

The formation of a test review committee was a critical step in the study. The committee was charged with assessing the content validity and other psychometric qualities of the BINL on the basis of pre-specified criteria. Committee members were selected on the basis of:

- background in bilingual education,
- knowledge of and experience with test and measurement in general,
- knowledge of and experience with assessment involving bilingual students,

- knowledge of and experience with bilingual instruments,
- knowledge and understanding of the approach to compensatory education in the Honolulu district,
- knowledge and understanding of SLEP instructional objects, and
- independence, self-assurance and capacity for critical thinking.

Based on the above criteria, nine individuals were identified and selected as members of the test review committee. They included four district staff, a school-level project teacher, a state level evaluation staff, two university faculty members and the external evaluator.

B. Test Evaluation Criteria

Numerous sources were used to develop a set of criteria for test evaluation. These included various documents produced by the Center for the Study of Evaluation of UCLA (Hoepfner, et al., 1976), the Center for Bilingual Education (Silverman, et al., 1976; Silverman, et al., 1978) and the Assessment Projects at the Northwest Regional Educational laboratory (Nafziger, et al., 1975), the American Psychological Association, the American Educational Research Association, the National Council of Measurement in Education (Davis, et al., 1974), as well as individual researchers (e.g., Madaus, et al., 1982). The final set of criteria used in the present study thus represents a comprehensive compilation of generally accepted test standards which had been field tested and used in test evaluation.

More specifically, the criteria relate to four major areas of test characteristics: measurement validity, examinee appropriateness, technical excellence, and administrative usability. The criterial areas are further described as follows:

Measurement validity. This set of criteria looks at the nature of what a test measures, the range of behaviors sampled, the relationship of the test score to other measures, and the demonstrated usefulness of the test in theoretical or practical settings.

Examinee appropriateness. These criteria relate to the appropriateness of the test materials, including content of the stimuli (items) and mode of response, relative to the grade level of students taking the test.

Administrative usability. These criteria deal with practical concerns in administering and using a test. The ease with which the test can be given, scored, and interpreted, and the usefulness of the resulting score in making program or instructional decisions.

Technical excellence. These criteria are concerned with the test's reliability, replicability and refinement of measurement.

Each of the four criterial areas included several individual criteria. To facilitate test review, these individual criteria were transformed into questions to guide the test reviewers in test evaluation.

C. Test Review

The review session was preceded by a test administration demonstration provided by one of the district staff. The demonstration was provided to enhance committee members' understanding of the BINL. Strict protocols were observed during the review session. Committee members followed directions developed specifically for the test review. Criteria to be used to evaluate the BINL were explained to committee members by the external evaluator. All evaluative decisions were based on information presented in the Manual and related documents supplied by

the test developer. No attempt was made to verify the available information. On the other hand, when needed information was not available and was not readily inferrable from existing data, an unfavorable rating was to be given.

In assessing content validity of the BINL, the 40 starter pictures served as test items in the review process. The items were assessed in terms of the degree of match with a set of instructional objectives provided by the SLEP program staff. These objectives relate to students' ability to:

- express feelings, attitudes, and meaning through a variety of sentence structures;
- make statements;
- ask questions;
- interact with others to convey a message; and
- express ideas effectively and clearly.

For each BINL item, test reviewers were asked to determine whether the item provided a relevant measure of at least one of the SLEP objectives. Reviewers were also asked to determine the percentage of SLEP objectives measured by one or more of the BINL items.

D. Test-Retest Reliability

To assess test reliability of the BINL, a test-retest study was performed on the test as it is used in the SLEP program. A random sample of 192 students was drawn from different grade levels at different schools. Data elements included school name, grade level of student, student name, BINL raw score, NCE score and BINL level score.

Data coding was performed by the district staff. Completed data sheets were mailed to the author for key-punching and analysis.

E. Mainstreaming

A mainstreaming study was performed to evaluate the effects of the BINL as a measure for exiting students. Data were gathered for a random sample of over 200 students at different grades in different schools. Only students exited from the SLEP Program for at least six months were included in the sample. The data included school name, student name, grade level, BINL administration date, BINL raw score, MAT administration date, MAT percentile scores (language arts, reading and mathematics) and year-end school grades. For purposes of comparability, all school grades were converted to scores on a five-point scale (i.e., 1 = failing, 2 = barely passing, 3 = satisfactory, 4 = very good, 5 = excellent). Data coding was performed by the district staff. Completed data sheets were mailed to the author for key-punching and analysis.

F. Test Use Survey

Three separate surveys were conducted in the Honolulu, Leeward and Central districts in Hawaii to obtain a measure of test use satisfaction on the BINL and the Language Assessment Scales (LAS). The survey instrument consisted of 41 items developed essentially on the basis of the same set of criteria used in the test review study. Items specific to information contained in the BINL or LAS manual were excluded, however. See Appendix A.

All three surveys were conducted in April 1983. Data obtained from the surveys were coded and entered into the computer by district staff for preliminary tabulations. Analyses were performed separately for each district and an additional analysis was conducted by pooling data from the Leeward and Central districts.

RESULTS

This section presents the results of each of the evaluative activities performed in the study. First, the findings of the test review committee are discussed. This is followed by results obtained in the reliability study, the mainstreaming study, and the test use satisfaction surveys.

A. Test Review Findings

These findings are presented in terms of the general qualities of the BINL as viewed on the basis of the pre-specified criteria and then specifically in terms of its content validity.

General qualities. Responses suggest that the test reviewers generally perceived the BINL to be an instrument of high merit. With few exceptions, the test received favorable ratings from the committee members. Several items are worthy of particular mention.

First, the BINL item development process appeared to be of some concern to the committee members. While some of the rationale could be inferred from the test manual, the item development process as a whole appeared poorly documented.

Secondly, most reviewers appeared uncertained as to whether the BINL could be expected to correlate with student performance in school subjects. One reviewer suggested that the BINL measured only oral language proficiency and could not be expected to serve as a predictor of achievement in other academic areas.

Thirdly, some reviewers were not certain if the BINL scoring procedure can be described as objective and simple. The majority of the reviewers, however, thought it was.

Fourthly, while virtually all reviewers felt the norm groups used in the standardization of the BINL were of sufficient size, a few expressed reservations over the representativeness of the norm groups, particularly with respect to racial, ethnic, economic and sexual representation.

Fifthly, a couple of reviewers expressed concern over whether the BINL was capable of adequately differentiating among students at the upper and lower ends of the achievement spectrum and whether the test measured a diversity of skills of bilingual students. A majority of the reviewers, however, did not see the BINL lacking such capabilities.

As mentioned earlier, responses to most of the items were favorable. The overall positive perceptions of the test review committee are perhaps best reflected in their recommending the use of the BINL in the SLEP Program.

Content validity. Content validity was assess by posing two questions to the test reviewers: (a) What proportion of the BINL items appears to measure one or more of the SLEP instructional objectives as identified by the program staff? (b) What proportion of the SLEP instructional objectives is measured by one or more of the BINL items?

As mentioned earlier, for purposes of this study, the 40 starter pictures, being the primary stimuli for eliciting student responses, were regarded as test items. A set of five major instructional objectives was identified by the SLEP program staff. These objectives pertained to the student's ability to:

1. express feelings, attitudes, and meaning through a variety of sentence structures;
2. make statements;

3. ask questions;

4. interact with others to convey a message; and

5. express ideas effectively and clearly.

An examination of the BINL items relative to the SLEP instructional objectives produced a general consensus among the test reviewers with respect to the content validity of the BINL. Specifically, all of the BINL items were judged to be related to objectives 1, 2, and 5. None of the items was perceived to be a measure of one's ability to ask questions (objective 3) or to interact with others (objective 4). At best, the testing situation may yield an indirect indication of such abilities. Furthermore, the ability to express feelings and attitudes (as distinct from meaning) included in objective 1 is only partially or indirectly assessed by the BINL items.

Thus it appears that all the BINL test items (starter pictures) relate to some of the SLEP instructional objectives and are therefore capable of providing a measure of the student's ability in the relevant skill areas. On the other hand, not all of the SLEP objectives are measured by the BINL items. Two of the five objectives identified by the program staff are in fact only indirectly, if at all, assessed by the BINL items.

B. Test-Retest Reliability

The random sample chosen for the test-retest reliability study consisted of 192 students drawn from all grade levels (k-12) in 26 of the 55 schools. Numbers of elementary, intermediate and high school students included in the sample were proportionate to SLEP students in the district at the respective grade levels. Most of the students were

tested on January 3 through January 14, 1983 for the first test. The retest, in most cases, was administered between January 31 through February 17. Some of the delay in the first and second testings was due to logistical problems in recording students names and notifying the schools to retest the students. All testing was conducted by the SLEP program staff at the school sites. The completed score sheets were machine-scored by the test publisher in California.

The data indicate that the average BINL raw score (language complexity) for the first test was 90.53 with a corresponding Normal Curve Equivalent (NCE) score of 60.93. At retest, the students obtained an average BINL raw score of 99.58 corresponding to an NCE of 66.44. The gain was probably due to elapsed time between the two testings. The interval was, in some cases, longer than anticipated and a greater amount of learning than expected could have occurred.

The primary interest in the test-retest reliability study is of course the intercorrelations among the variables included in the study. Of particular importance is the correlation between the first testing and second testing. Data show that the test-retest correlation in terms of BINL raw scores was .88. The correlation in terms of NCEs was .87. These test-retest reliabilities should be viewed with some caveats.

First, different methods of obtaining test reliability generally yield different results. The parallel forms correlation is typically the lowest and the odd-even (e.g., split-half) reliability the highest (Gulliksen, 1950, p.215). The test-retest reliability coefficient of .88 obtained in the present study is probably quite comparable with the split-half coefficient of .92 reported in the BINL test manual (Herbert, 1979).

Second, reliabilities of oral language tests are generally lower than most standardized achievement tests, oral language proficiency being a relatively more difficult trait to measure (Silverman et al., 1976; Perlman and Rice, 1979).

Third, as indicated earlier, the time interval between the first test and the retest was longer than anticipated at least in some cases. While the elapsed time was, by all indications, not long enough to seriously confound the results, it most likely served to attenuate the test-retest reliability.

C. Mainstreaming

A random sample of 236 students was used in the mainstreaming study. These students were mainstreamed between September 1980 and June 1981. The sample covered all grade levels (k-12) and 40 of the 55 schools in the district. Numbers of elementary, intermediate and high school students included in the sample were proportionate to SLEP students in the district at the respective grade levels. Farrington High School which enrolled 14 percent of the SLEP students in the district was slightly overrepresented in the sample. Twenty-two percent of the sample was obtained from that school.

In selecting the sample, student folders were randomly picked from file boxes containing all exited students during the specified period. No attempt was made to randomly select students from the various language groups. An examination of the final sample by the SLEP program staff indicated that the sample did appear representative of the language groups in the district.

After the sample was selected, year-end grades for the 1981-82 school year were obtained. These grades represented their post-SLEP achievement in the regular school setting after the students had been mainstreamed for at least a year. At the secondary level, grade point averages for mathematics and language arts were obtained. At the elementary level, a single grade point average for mathematics was provided; language arts was divided into reading and speaking/listening--except for two schools which provided a single language arts grade point average.

The somewhat diverse grading schemes used at different schools were converted to a common five-point scale as follows:

1 = Failing

2 = Barely passing

3 = Satisfactory

4 = Very good

5 = Excellent

The exited students might have been tested with the BINL on several occasions. In such cases, BINL scores obtained immediately prior to mainstreaming were used. Both BINL raw scores and BINL levels were provided.

The primary interest of the mainstreaming study was to assess the effects of the BINL (in conjunction with the MAT) as an instrument for exiting SLEP students. More specifically, the question of primary interest was whether students mainstreamed on the basis of the BINL (and MAT) were performing satisfactorily in the regular school setting.

The data suggest that on the average the exited students appeared to be doing more than satisfactorily in the regular classroom in all subject

areas, particularly in mathematics and language arts. On a five-point scale, the average mainstreamed student earned grades ranging from 3.19 to 3.62. When the data were further analyzed in terms of percent of students achieving various school grades, a similarly positive achievement pattern emerged. Data indicate that less than 4 percent of the exited students were actually failing in some subjects. The predominant majority demonstrated satisfactory or better than satisfactory achievement (i.e., 85.7 percent) in mathematics, 82.5 percent in language arts, 93.1 percent in reading, 91.1 percent in speaking/listening). Approximately one-half of the students showed "very good" or "excellent" performance in mathematics (46.5 percent) and language arts (57.5 percent). Over one-fifth of these students had similarly high achievement in reading (28.4 percent) and speaking/ listening (20.5 percent).

D. Test Use Satisfaction

The survey on test use satisfaction was conducted to assess how satisfied the SLEP staff were with the use of the BINL in the program. The Inventory of Test Use Satisfaction (IOTUS) was administered to all SLEP program staff in the district in April 1983. Similar surveys were also conducted at the same time in the Leeward and Central districts to provide data for comparison purposes. As indicated earlier, items in the IOTUS were developed on the basis of pre-specified criteria for test evaluation. The instrument consists of two parts. Part I is made up of 6 items relating to the respondent's general knowledge of and experience with the test in question. part II consists of 35 items mostly relating to the specific test evaluation criteria. The following is a presentation of major findings.

Sixty-one SLEP program staff in the Honolulu district responded to the survey. These included four educational assistants, 12 part-time temporary teachers and 34 teachers. The others did not specify their job positions. A predominant majority of the respondents (83.7 percent) rated their knowledge of the BINL as good or excellent. Over 90 percent had administered the BINL 8 or more times. A majority (83.3 percent) reported that it took 20 minutes or less to administer the BINL.

Over one-half (57.8 percent) of the respondents indicated that at least 50 percent of the skills taught in the project were measured by the BINL, with a sizeable number (29.8 percent) indicating that 71 percent or more of the skills were covered by the test. With respect to test use, equal emphasis appeared to have been placed on evaluation (80 percent), student selection (70 percent), diagnosis (68 percent), instructional planning (68 percent), and student placement (90 percent).

With respect to measurement validity, the responses were highly favorable. In all cases, a majority of the respondents felt that (a) they knew what the test was supposed to measure (93.4 percent); (b) the items in the test seemed conceptually sound (54.1 percent); (c) the test measured what it was supposed to measure (78.6 percent); and (d) the test measured something distinct from what was measured by other similar tests (51.7 percent). The respondents seemed less certain about the ability of the test in predicting how well a particular student would do in other school subjects. Less than a quarter (24.5 percent) of the respondents felt the test possessed such predictive validity.

A predominant majority of the respondents (85 percent) indicated that the BINL provided reliable information for its intended use. Such uses

included evaluation, student selection, placement, diagnosis and instructional planning. Virtually all (95 percent) of the respondents reported that the test results generally turned out to be what they would expect. Over 88 percent indicated that they generally made use of the test information in some way.

With respect to examinee appropriateness, a majority of the respondents (73.8 percent) felt that the layout of the test (including print size, illustrations, use of white space and color) was attractive and helpful. The respondents, however, appeared having difficulty in treating the starter pictures as "items" and most (78 percent) did not respond to the question regarding how well the items were written. Some did not respond to the question regarding item relevance or item bias. Over one-half (50.8 percent) did indicate that the items appeared free of cultural, sexual and ethnic bias.

Virtually all (97.3 percent) respondents indicated they had no difficulty in administering the test to students. Most (83.6 percent) reported that they were able to administer the test in the same way each time they tested the students, with 59 percent indicating that administering the test was an enjoyable and rewarding experience. A predominant majority (88.5 percent) believed that the way in which students were required to respond to the test items was simple and direct. Less than 10 percent felt that it took too long to adminster the test.

The BINL also received very favorable ratings in terms of administrative usability. For instance, the test manual was rated as clear, well-organized, consistent, thorough and helpful by 82.8 percent

of the respondents. Similarly, the instructions for administering the test were deemed to be clear and easy to follow by 95 percent of the respondents. The answer or scoring sheet was easy to use, according to another 78.6 percent of those responding to the survey.

While a majority (60.6 percent) of the respondents perceived the scoring procedure as straightforward and objective, a sizeable number (41 percent) did not indicate whether it would be difficult to hand-score the test--presumably they had never had to do so. Most (70 percent) indicated there would be no difference between hand- and machine-scoring. Over one-half (51.7 percent) of the respondents felt the test provided an important source of information for program improvement. A majority (68.8 percent) reported that they often used the test results to make instructional decisions.

A sizeable number (41.7 percent) of the respondents apparently had never had to convert raw scores to normed or interpreted scores for the BINL and did not respond to the question regarding score conversion. Forty percent, however, did respond, indicating that the score conversion process was easy. Most respondents (70 percent) felt that it was easy to understand the various scores provided by the test. A majority (76.7 percent) indicated that they saw no problem in using the various test scores for the intended purposes such as evaluation, student selection, placement, diagnosis, and instructional planning.

With regard to technical excellence, the respondents felt that the BINL had enough items to include a sufficient range of difficulty (60 percent) and that both the raw scores and converted scores had a sufficient range to differentiate adequately among students (78.4 and 73.4 percent, respectively). However, a substantially lower percentage (40 percent) of the respondents believed that the BINL measured a wide

range or diversity of skills. About 10 percent indicated that the costs of the test were too high for the kinds of information it provided. Approximately 30 percent felt otherwise. One-third of the respondents did not respond to the item.

The overall perceptions on the BINL were perhaps best reflected in the respondents' expression of satisfaction with the use of the test in the SLEP Program. A predominant majority (78.4 percent) indicated they were satisfied. A few (3.3 percent) felt otherwise; the others (18.3 percent) apparently did not have strong feelings one way or the other. Over two-thirds (70 percent) would recommend the test for use in programs similar to the SLEP Program.

As indicated in an earlier section, test use surveys were also conducted in the Leeward and Central districts in which the LAS was used for student selection and other purposes. The combined sample of 81 consisted of 35 part-time teachers and 27 permanent teachers. The others did not specify their job positions.

Based on the survey results, several items appear worthy of mention for purposes of comparison. These points of interest are listed as follows:

1. The SLEP program staff in the respective districts appeared quite comparable in terms of both their knowledge of and experience with the respective tests--most having administered the BINL or the LAS 8 or more times.
2. A comparison of responses on content validity suggests that

there was a better match between the BINL and the Honolulu SLEP Program. While over one-half (57.8 percent) of the Honolulu respondents indicated a match of 50 percent or better, only slightly more than one-third (35.1 percent) of the Leeward/Central respondents felt the same way about the LAS. Furthermore, proportionately more Honolulu staff used the BINL for diagnosis (68.3 percent) and instructional planning (68.3 percent) as compared with Leeward/Central staff using the LAS for similar purposes (53.0 percent and 42.5 percent, respectively).

3. With respect to measurement validity, over one-half (51.7 percent) of the Honolulu respondents believed that the BINL measured something distinct from what was measured by other similar tests. Approximately to one-third (32.4) of the Leeward/Central respondents felt the same way about the LAS. In both cases, a much lower percentage of respondents (24.5 percent for Honolulu, 27.2 percent for Leeward/Central) believed that the respective tests provided results capable of predicting how well students may perform in other school subjects.

4. A substantially higher percentage of the Honolulu respondents (85 percent) indicated that their test provided reliable information for its intended use. Only 55 percent of the Leeward/Central respondents felt the same way about the LAS. Also, proportionately more Honolulu respondents (88.1 percent) reported using test information they received from the BINL.

Only 63.7 percent of their Leeward/Central counterparts indicated using test information provided by the LAS.

5. There were differences in perception with respect to ease and appropriateness of test administration. Again, these differences were generally in favor of the BINL. For example, virtually all (98.3 percent) the Honolulu respondents indicated that they had no difficulty in administering the BINL while 81.5 percent of their Leeward/Central counterparts felt the same way about the LAS. Furthermore, 83.6 percent of the Honolulu respondents reported that they were able to administer the BINL in the same way each time they tested their students. The corresponding figure for the Leeward/Centeral respondents was 72.8 percent. Approximately 88 percent of the Honolulu respondents believed that the way in which students were required to respond to the BINL test items was simple and direct. About 76 percent of their counterparts in Leeward/Central felt the same way about the LAS.

6. Comparisons with respect to the test manual, instructions for test administration, use of answer sheets and scoring procedures were also generally in favor of the BINL. For instance, 82.8 percent of the Honolulu respondents agreed that the BINL test manual was clear, well-organized, consistent, thorough and helpful. About 68 percent of the Leeward/Central respondents felt the same way about the LAS. Practically all (95 percent) of the Honolulu respondents indicated that the instructions for administering the BINL were clear and easy to follow. About 88

percent of the Leeward/Central respondents felt the same way about the LAS. Similarly, a higher percentage of the Honolulu respondents (78.6 percent versus 65.4 percent) indicated that the BINL answer sheet was easy to use. Although a sizeable number of the respondents (both in Honolulu and in Leeward/Central) did not respond to the item, a higher percentage of the Honolulu respondents (60.6 percent versus 35.0 percent) agreed that the scoring procedure for the BINL was straightforward and objective. Few (21.3 percent), however, indicated that they would have no difficulty in hand-scoring the test.

In other aspects of administrative usability, the responses were also generally more in favor of the BINL than the LAS. As a case in point, about 40 percent of the Honolulu respondents reported that it was easy to convert raw scores to normed or interpreted scores for the BINL. Approximately 23 percent of the Leeward/Central respondents indicated the same for the LAS. Proportionately, a far greater number of the Honolulu respondents indicated that it was easy to understand the meaning of the various scores provided by the BINL (70.0 percent versus 51.2 percent) and that they saw no problems in using the various test scores for the intended purposes (76.7 percent versus 49.4 percent). Moreover, a greater proportion of the Honolulu respondents (68.8 percent) used the test results to make instructional decisions than did their Leeward/Central counterparts (42.5 percent).

8. Even though a sizeable number (20 percent) of the Honolulu respondents did not respond to the item, 45 percent of them did indicate that the items were relevant to their students. Approximately 30 percent of the Leeward/Central respondents felt the same way about the LAS items. In both cases, approximately one-half of the respondents (50.8 percent for Honolulu and 46.9 percent for Leeward/Central) indicated that the test items were free of cultural, sexual and ethnic bias. Proportionately fewer Honolulu respondents (14.8 percent versus 25.9 percent) felt that the items were not free of bias.

9. There appeared to be some evidence that the BINL provided a wider range of coverage than the LAS. A predominant majority of the Honolulu respondents indicated that the BINL raw scores (78.4 percent) and converted scores (73.4 percent) had a sufficient range to differentiate adequately among students. Less than one-half (31.2 percent and 34.7 percent, respectively) of the Leeward/Central respondents felt the same way about the raw and converted scores for the LAS. About 60 percent of the Honolulu respondents agreed that the BINL had enough items to represent a sufficient range of difficulty. Approximately 40 percent indicated that the test measured a wide range or diversity of skills. The corresponding figures for the Leeward/Central respondents were 43.7 percent and 37.7 percent, respectively. It should also be noted that a sizeable number of the respondents (28.4 percent in Honolulu and 41.6 percent in Leeward/Central) did not think the respective tests measured a wide range or diversity of skills.

10. In the Honolulu district, a predominant majority (78.4 percent) of the respondents indicated that they were satisfied with the use of the BINL in the SLEP Program. Only 3.3 percent expressed dissatisfaction. In the Leeward and Central districts, 36.2 percent of the respondents reported that they were satisfied with the use of the LAS in their program while another 36.2 percent expressed dissatisfaction. About 70 percent of the Honolulu respondents would recommend the use of the BINL in programs similar to the SLEP Program. Approximately 34 percent of the Leeward/Central respondents would recommend the use of the LAS.

While it is possible to overinterpret perceptual data, the responses obtained from the three surveys did present a clearly discernible trend supporting the continued use of the BINL in the Honolulu district. Not only did the respondents think highly of the test, comparative data suggest that their support and enthusiasm for the BINL appeared to be greater than that expressed by users of the LAS for that test in the Leeward and Central districts. In most cases, survey responses were more favorable to the BINL than they were to the LAS.

CONCLUSIONS

The primary purpose of the present study was to obtain and interpret data pertaining to the BINL to help determine whether the continued use of the test in the Honolulu district is warranted. To that end several studies were conducted to obtain information on the psychometric qualities (particularly with respect to the content validity and

test-retest reliability) of the BINL, the effects of the BINL as an instrument for mainstreaming SLEP students, and perceptions of test users with respect to their overall test use satisfaction.

Several approaches were taken to obtain the relevant data, including the formation of a test review committee, the testing and retesting of a random sample of SLEP students, an in-depth examination of school grades of a random sample of exited students, and the conduct of surveys on test use satisfaction in three school districts. Results of the study appear to support the following conclusions:

1. Findings obtained from the test review session suggest that the BINL possesses favorable psychometric qualities as a measure of oral language proficiency. The match between processes employed in the test and the psycholinguistic principles which form the philosophical bases of the SLEP Program is considered excellent and perhaps unique, providing high content validity for the BINL. It is also obvious, however, that the BINL items do not measure all the skills which are emphasized in the SLEP Program. The BINL's content coverage is somewhat limited to oral proficiency in English and does not include such important skills as reading comprehension.
2. Results of the test-retest study suggest that the BINL as a language proficiency test possesses an adequately high degree of reliability. In spite of the attenuating factors which inadvertently occurred in the study, a test-retest reliability coefficient of around .88 was obtained for the test. It is also

noted, however, that responses obtained from the test use survey suggest that the scoring procedures were perhaps not as straightforward and objective as they could have been.

3. The mainstreaming study showed that in most cases exited students were performing satisfactorily, if not better than satisfactorily, in the regular school setting. When their school grades were converted to a common five-point scale, above average performance was indicated in all subject areas included in the study. Only in very few cases (less than 4 percent) were exited students shown to be failing in some subjects. A sizeable number of the exited students (20-57 percent) were doing "very good" or "excellent" work in the regular classroom following their exit from the SLEP Program.

4. The test use surveys conducted in the Honolulu, Leeward and Central districts indicated that there was a high degree of support and enthusiasm on the part of the program staff for the use of the respective tests (the BINL and the LAS) in these districts. It also appeared that the degree of support and enthusiasm was greater in Honolulu than in the other two districts. There was clear evidence that the Honolulu program staff were highly satisfied with the use of the BINL in the SLEP Program and believed the test served all the functions it was intended to serve.

REFERENCES

Davis, F.B. (Chair) Standards for Educational and Psychological Tests. Washington, D.C.: American Psychological Association, 1974.

De Avila, E.A., and Duncan, S.E. Language Assessment Scales. Corte Madera, CA: Linquametrics Group, 1977.

Gulliksen, H. Theory of Mental Tests. New York: John Wiley and Sons, 1950.

Hawaii State Department of Education. Identification, Assessment and Programming System for Students of Limited English Proficiency: A Systems Manual 1980-81. Honolulu: Hawaii State Department of Education, 1980.

Herbert, C.H. Basic Inventory of Natural Language. San Bernadino, CA: CHECpoint Systems, 1979.

Hoepfner, R., Bastone, M., Ogilvie, V., Hunter, R., Sparta, S., Grothe, C.R., Shani, E., Hufano, L., Goldstein, E., Williams, R., and Smith, K.O. CSE Elementary School Test Evaluations. Los Angeles: Center for the Study of Evaluation, UCLA, 1976.

APPENDIX A

Madaus, G.F., Airasian, P.W., Hambleton, R.K., Consalvo, R.W., and Orlandi, L.R. Development and application of criteria for screening commercial, standardized tests. Educational Evaluation and Policy Analysis, 1982, 4(3), 401-415.

Nafziger, D.A., Thompson, R.B., Hiscox, M.D., and Owen, T.R. Tests of Functional Adult Literacy: An Evaluation of Currently Available Instruments. Portland, OR: Northwest Regional Educational Laboratory, June 1975.

Perlman, C.L., and Rice, W.K. A normative study of a test of English language proficiency. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Silverman, R., Noa, J.K., and Russell, R.H. Oral Language Tests for Bilingual Students: An Evaluation of Language Dominance and Proficiency Instruments. Portland, OR: Northwest Regional Education Laboratory, July 1976.

Silverman, R., and Tupper, N. Assessment Instruments in Bilingual Education. Portland, OR: Northwest Regional Educational Laboratory, 1978.

DIRECTIONS FOR COMPLETING THE INVENTORY OF TEST USE SATISFACTION

The purpose of the Inventory is to assess how satisfied you are with the use of tests in your program. A separate Inventory should be completed for each test. For this particular study, Honolulu District respondents should complete the Inventory for BINL. Leeward and Central Distict respondents should complete the Inventory for LAS. Keep the following directions in mind when you respond to the Inventory.

1. Fill in your name (optional), position, school district and date on page 1 of the Inventory.

2. Indicate the name of the test for which you are completing the Inventory. Honolulu District respndents should complete the Inventory for BINL. Leeward and Central District respondents should complete the Inventory for LAS.

3. Be as thorough and candid as you can in responding to the items. Responses to the Inventory will not be identified with names of individual respondents.

4. Read the items carefully before you respond. Throughout the Inventory the term "program" means the SLEP Program. The term "students" means students participating in the SLEP Program or students being tested for participation in the Program. Unless otherwise indicated or implied, the test means the test for which you are completing the Inventory.

5. Use your general impression of the test as a guide in responding to the items. We want your best professional judgments on the test--not scientific facts.

6. Return the completed Inventory to Dr. Don Enoli of the Honolulu District office.

INVENTORY OF TEST USE SATISFACTION

Name (optional): _____

Position: _____

School District: _____

Date: _____

Name of test (one only) for which you are completing this inventory:

_____

PART I:

Check one of the choices for each of the following items.

1. How would you rate your knowledge of the test?

   _____ Little

   _____ Moderate

   _____ Good

   _____ Excellent

2. How many times have you administered the test to students in the program?

   _____ Never

   _____ 1-3 times

   _____ 4-7 times

   _____ 8 or more times

3. How long does it take to administer the test to the average student?

_____ 1-10 minutes

_____ 11-20 minutes

_____ 21-30 minutes

_____ 31-40 minutes

_____ more than 40 minutes

4. What percentage of the skills that you teach students in the program is covered by the test?

_____ 20% or less

_____ 21-30%

_____ 31-40%

_____ 41-50%

_____ 51-60%

_____ 61-70%

_____ 71% or more

5. The test is primarily used for: (Check all that apply)

_____ Evaluation

_____ Student Selection

_____ Diagnosis

_____ Instructional Planning

_____ Student Placement

_____ Other

6. How long has the test been in use in the program for the purpose(s) indicated above?

_____ Less than 1 year

_____ 1-2 years

_____ 3-4 years

_____ More than 4 years

_____ Not sure

PART II:

Indicate whether you agree or disagree with each of the following statements by circling SA for Strongly Agree; A for Agree; N for Neutral; D for Disagree; and SD for Strongly Disagree. Circle NA only if the item is not applicable or inappropriate.

1. I know what the test is supposed to measure.

SA A N D SD NA

2. The items in the test seem conceptually sound in that the items are based on theory of linguistics, education, psychology and learning.

SA A N D SD NA

3. The test measures what it is supposed to measure.

SA A N D SD NA

4. The test measures something distinct from what is measured by other similar tests.

SA A N D SD NA

5. The test provides results which generally tell me how well a particular student is doing in other school subjects too.

SA A N D SD NA

6. The test provides reliable information for its intended use (e.g., evaluation, student selection, placement, diagnosis and instructional planning).

SA A N D SD NA

7. The test results generally turn out to be what I would expect.

SA A N D SD NA

8. I generally make use of information I get from the test.

SA A N D SD NA

9. I have no difficulty in administering the test to students.

SA A N D SD NA

10. I am able to administer the test in the same way each time I test the students.

SA A N D SD NA

11. The way in which students are required to respond to the test items is simple and direct.

SA A N D SD NA

12. Administering the test is an enjoyable and rewarding experience.

SA A N D SD NA

13. It takes too long to administer the test.

SA A N D SD NA

14. The test manual is clear, well-organized, consistent, thorough and helpful.

SA A N D SD NA

15. Instructions for administering the test are clear and easy to follow.

SA A N D SD NA

16. The answer or scoring sheet is easy to use.

SA A N D SD NA

17. The scoring procedure for the test is straightforward and objective.

SA A N D SD NA

18. I would have no difficulty in hand-scoring the test.

SA A N D SD NA

19. When the test is machine-scored, the results are often somewhat different from what I would expect.

SA A N D SD NA

20. The test provides an importaht source of information for program improvement.

SA A N D SD NA

21. I often use the test results to make instructional decisions.

SA A N D SD NA

22. The layout of the test (including print size, illustrations, use of white space and color) is attractive and helpful.

SA A N D SD NA

23. The test items are generally well written.

SA A N D SD NA

24. The test items are relevant to my students.

SA A N D SD NA

25. The test items are free of cultural, sexual and ethnic bias.

SA A N D SD NA

26. It is easy to convert raw scores to normed or interpreted scores for the test.

SA A N D SD NA

27. It is easy to understand the meaning of the various scores provided by the test.

SA A N D SD NA

28. I see no problems in using the various test scores for the intended purpose (e.g., evaluation, student selection, placement, diagnosis, and instructional planning).

SA A N D SD NA

29. The test has enough items to include a sufficient range of difficulty.

SA A N D SD NA

30. The test measures are wide range or diversity of skills.

SA A N D SD NA

---

31. The raw scores have a sufficient range to differentiate adequately among students.

SA A N D SD NA

32. The converted scores have a sufficient range to differentiate adequately among students.

SA A N D SD NA

33. The costs of the test (including test materials, administration, scoring and interpretation) are too high for the kinds of information it provides.

SA A N D SD NA

34. I am very satisfied with the use of the test in my program.

SA A N D SD NA

35. I would recommend the test for use in programs similar to mine.

SA A N D SD NA