DOCUMENT RESUME

ED 317 075 FL 018 429

AUTHOR	Stansfield, Charles W.; And Others
TITLE	The Guam Educators' Test of English Proficiency (GETEP). Final Project Report, Revised.
INSTITUTION	Center for Applied Linguistics, Washington, D.C.
PUB DATE	5 Apr 90
NOTE	129p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Databases; Elementary Secondary Education; *English (Second Language); Interviews; *Language Proficiency; *Language Tests; Multiple Choice Tests; Oral Language; Scoring; *Teacher Certification; *Test Construction; *Testing Programs; Test Items; Test Validity; Verbal Tests; Writing Exercises
IDENTIFIERS	*Guam Educators Test of English Proficiency

ABSTRACT

The development and field testing of a proficiency test in English as a Second Language for non-native speakers teaching on Guam is reported. The resulting instrument measures four language skills (listening, reading, writing, and speaking). The listening measure uses natural language that might be heard by a classroom teacher. The reading measure is based on authentic materials for Guam educators, including department of education publications and professional journals. The writing measure consists of a holistically scored essay on an educationally relevant topic and task, and a multiple-choice portion requiring the prospective teacher to identify errors in three simulated student essays. The speaking measure is an oral proficiency interview. The report describes the needs assessment, test construction, field testing and revision, and the setting of appropriate passing scores on each test section. Recommendations for the implementation of an operational testing program are made, taking into account the particular circumstances of the Guam Department of Education. Appended materials include the needs analysis report, item specifications and samples, instructions, scoring guides, and other information related to test development and administration. (MSE)

The Guam Educators' Test of English Proficiency (GETEP)

Final Project Report

Revised

Charles W. Stansfield

John Karl

Dorry Mann Kenyon

Center for Applied Linguistics
1118 22nd Street, NW
Washington, DC 20037

5 April 1990

The Guam Educators' Test of English Proficiency (GETEP)

Final Project Report

Revised

Charles W. Stansfield

John Karl

Dorry Mann Kenyon

Center for Applied Linguistics
1118 22nd Street, NW
Washington, DC 20037

5 April 1990

Table of Contents

Acknowledgements

1. Overview of the Project 1

2. Preparation for Test Development 2

3. Development of the GETEP Multiple Choice Sections 4

3.1 Development of Test and Item Specifications 4
3.1.1 Listening Comprehension Specifications 5
3.1.2 Reading Comprehension Specifications 6
3.1.3 Writing Proficiency Specifications (Error Detection Items) 7

3.2 Writing Items and Assembling Field Test Forms 8
3.2.1 Preparing Listening Comprehension Items 8
3.2.2 Preparing Reading Comprehension Items 9
3.2.3 Preparing Error Detection Items 10
3.2.4 Preparing the Field Test Forms of the GETEP . 11

3.3 Field Testing the Multiple Choice Sections 11

3.4 Statistical Item Analyses and Revisions to the Field Test Forms 15
3.4.1 Revision of Listening Comprehension Items . . 16
3.4.2 Revision of the Reading Comprehension Items . 18
3.4.3 Revision of the Error Detection Items 19

3.5 Preparation of the Final Forms 20

4. Development of the GETEP Writing Sample (GWS) 21

4.1 Specifications 21
4.2 Developing Writing Prompts 21
4.3 Field Testing 23
4.4 Developing the GWS Scoring Guide and Training Raters 23
4.5 Inter-rater Reliability 26
4.6 Standard Setting 26

5. Development of the GETEP Oral Proficiency Interview . . . 28

5.1 Background 28
5.2 Training Interviewers 28

6. Setting Passing Scores for the GETEP 32

6.1 Background 32
6.1.1 Judges 32
6.1.2 Defining Minimal Competency 33
6.1.3 Standard Setting Methods 34

6.2 Procedures Used in This Study 37
6.2.1 Selection of Judges 37
6.2.2 Approach to Standard Setting 39

6.2.2.1 Writing Sample 39
6.2.2.2 Oral Interview 40
6.2.2.3 Multiple-choice Sections 40

6.3 Findings 43
6.3.1 Description of the Minimally Competent Teacher 43
6.3.2 Writing Sample 45
6.3.3 Oral Interview 45
6.3.4 Multiple-choice Sections of the GETEP 46

6.4 Discussion 50

7. Recommendations for the Operational Testing Program . . . 53

7.1 Administration and Scoring of the GETEP on Guam . . 53
7.1.1 Appointment of a GETEP Program Director . . . 53
7.1.2 DOE to Issue Pass/Fail Scores 53
7.1.3 Examinee Handbook 53
7.1.4 Manual for Administering the GETEP 54
7.1.5 GETEP Answer Sheet 54
7.1.6 Printing of the GETEP 55
7.1.7 Test Security 56

7.2 Administration and Scoring of the GETEP on the Mainland 56
7.3 Administration and Scoring by Recruiting Teams . . . 57
7.4 Maintenance of a Database 58
7.5 Research 59
7.6 The Operational Speaking Test Program 60
7.7 The GETEP Writing Sample 61

References 63

Appendices

A....Complete Needs Analysis Report
B....Specifications for the Listening and Reading Comprehension Sections of the GETEP
C....Item-specification Table for the Final Listening Comprehension Short Dialog Section
D....Instructions for the GETEP Writing Sample
E....GETEP Writing Sample Scoring Guide and Benchmark Samples
F....Letter Describing Required Characteristics of Individuals Competent to make Judgements on the Passing Scores for the GETEP
G....Judges' Rating Sheet for the Passing Score Study
H....Complete Responses to the Judges' Questionnaire
I....Proposed NCS Machine-readable Answer Sheet

Acknowledgements

The successful development of the Guam Educators Test of English Proficiency (GETEP) could not have occurred without the cooperation and assistance of many individuals. First of all, we wish to thank the many individuals from Guam who contributed their time and energy to the construction of this test; they were invaluable in ensuring that the final product truly is the Guam Educators Test of English Proficiency. Among these people, we would like to mention Ms. Anita Sukola, the Director of the Guam Department of Education, for meeting with CAL staff during each of three visits to Guam. Ms. Evelyn Salas, Certification Officer, Ms. Janette Yamashita, Associate Superintendent of Elementary Education, and Ms. Beth Montague, Associate Superintendent of Secondary Education, were gracious hostesses and logistical guides during a week of visiting schools during the needs assessment. We would especially like to thank the many teachers and principals with whom we met. They gave us an abundance of helpful suggestions regarding the form and content of the instrument. Ms. Florence Northway, project monitor within the Guam DOE, recruited over seventy examinees to participate in the field testing of four forms of the test in October 1989. She also took charge of local arrangements for visiting CAL staff and provided much useful information. Finally, John Shaver, consultant to the Guam Department of Education, was exceedingly helpful during the three visits to Guam, and in the interim as well.

Our sincere thanks go to the three groups of teachers and administrators who participated in workshops on Guam. These include the four speech teachers (Art Wheeler, Gerri Diaz, Bonnie Sarempa, and Cathy Cardenas) who devoted at least five days of their time to training to become oral proficiency interviewers; the four language arts specialists (Margaret Camacho, Julie Sisson, Francis McDonald, and Marie Barretto) who spent two days in training to score the GETEP Writing Sample; and the eight judges (Julie Sisson, Ernestina Cruz, Carmen Rodriguez, Beth

McClure, Nerisa Shaffer, Evelyn Salas, Patrick Artero, and Allene Yamashita) who participated in the four-day-long standard setting process. We were very impressed by the professionalism and dedication of these fine educators.

We also appreciate the hard work of Dr. Dan Robertson, professor of English at the University of Guam and CAL consultant and liaison, who took charge of the printing of the field tests and their administration in Guam. He also administered the GETEP Writing Sample to students at the University of Guam so that a scoring guide could be developed and benchmark essays identified. He arranged for students to be interviewed during the training of oral proficiency testers. Professors Gene Bruce and Evelyn Salas of the University of Guam were also helpful in reviewing the prompts for the GETEP Writing Sample. Early in the project, professor Joyce McCauley provided a number of Freshmen student essays, which were useful in determining the type of errors that would be included on the error detection items.

Here in Washington, as we were developing items for tne test, we were kept in an appropriately "on-island" state of mind by Genevieve Arbitrario, a teacher and librarian for 11 years in five of Guam's schools, who reviewed the multiple-choice items for their verisimilitude to life in those schools. This was particularly important in the case of the listening comprehension items. Ms. Nancy Conklin of the Northwest Laboratory for Educational Research, who has spent many years doing educational and linguistic research on Guam and in Micronesia, provided very helpful reactions to initial items for the listening comprehension section of the test. Mr. Patrick Artero of the Governor of Guam Liaison Office provided useful information on the people of Guam and gave us access to the Liaison Office's collection of newspapers from Guam. Dr. Hector Nevarez, our COTR, provided important assistance at critical points.

The GETEP benefited from the particular expertise of several other consultants to CAL, to whom we are most grateful: Ms. Elinore LeBaron, after being apprised of the findings of the needs assessment, contributed to the specifications for the test

and wrote many of the listening comprehension passages and items; Dan Kennedy wrote the error detection passages and instructions; Dr. Marsha Bensoussan, a Visiting Scholar at CAL from the University of Haifa, Israel, used her expertise in the construction of cloze tests to evaluate and revise our cloze passages and items; and Dr. David Carroll, also a Visiting Scholar at CAL and a language testing specialist with the British Council, helped write the reading comprehension items and organized these items into four parallel forms.

From the CAL staff itself, we thank Dr. JoAnn Crandall who helped lay the groundwork for the test by spearheading the needs assessment. She also aided in the transition from the needs assessment to test development by reviewing the draft specifications and the initial items. Ann Kennedy selected reading comprehension and cloze passages, and developed items based on them. Laurel Winston assisted in a variety of ways as the tests were being developed.

To all these individuals we owe a hearty "thank you."

Charles W. Stansfield
John Karl
Dorry Mann Kenyon

1. Overview of the Project

Although English is the language of instruction on the U.S. Territory of Guam, the local language is Chamorro, and other language communities are present on the island as well. As a result, many teachers on the island are not native speakers of English, and there is a need to ensure the English language competence of public school teachers, who teach children from Guam and other islands of Micronesia and the Pacific rim, as well as the children of U.S. military personnel and civilians from the mainland. As a result of this situation, the Center for Applied Linguistics received contracts from the Department of Defense Dependent Support Policy Directorate and the Guam Department of Education to develop an English language proficiency test for use in the certification of teachers.

The Guam Educators' Test of English Proficiency (GETEP) is a four communicative skills, job-relevant test of English for educators at the K-12 level. The listening measure employs natural language as might be heard by a teacher in a classroom or a school. The reading measure is based on authentic materials for educators in Guam, including publications of the Guam Department of Education, professional journals, etc. It includes a multiple-choice cloze format and traditional reading comprehension items. The writing measure consists of a holistically scored essay, on an educational relevant topic and task, and a multiple-choice portion that requires the prospective teacher to identify errors in three simulated student essays. The speaking measure is an oral proficiency interview.

The report describes the development of the GETEP, which began with a needs assessment, its field testing and revision, and the setting of appropriate passing scores on each section of the test. It makes recommendations for the implementation of an operational GETEP program, taking into account the particular circumstances of the Guam Department of Education.

2. Preparation for Test Development

The preparation for the development of the GETEP began in November, 1988, with a trip to Guam by Dr. JoAnn Crandall and Mr. John Karl of CAL. The purpose of the trip was to conduct a needs assessment that would ensure that the test developed would be responsive to the specific needs of teachers and administrators on Guam, and officials of the Guam Department of Education (DOE). To this end, Crandall and Karl met with DOE officials, University of Guam (UOG) faculty, principals and teachers from a sample of elementary and secondary schools, and with Department of Defense (DOD) administrative and education personnel. At each of these meetings, they solicited input regarding the kinds of test tasks that would correspond to the communicative tasks that teachers actually perform in their day-to-day work. Teachers and administrators suggested material that could be used for reading comprehension passages, and they suggested scenarios around which listening comprehension passages and items could be constructed. They also suggested topics that could be used in the speaking and writing sections of the test.

Possible item types were discussed, and there was widespread agreement that the following item formats, put forth by CAL in its proposal, would be appropriate for testing the various skills:

Test Section	Item Types
Listening comprehension:	* Multiple-choice (MC) short dialog items * MC extended dialog and monolog items
Reading comprehension:	* MC reading comprehension * MC cloze items
Writing proficiency:	* MC error detection items * holistically scored writing sample on relevant topics
Speaking proficiency:	* Oral interview on suggested topics

Sample items were presented to teachers and administrators to ensure that all parties understood what the final test would look like. Furthermore, during discussions with the DOE, it was agreed that the two forms originally envisioned should be expanded to four forms. The availability of additional forms would help maintain the security of the test.

To ensure that the test reflected communication as it actually occurs in classrooms on Guam, Crandall and Karl observed classes at four elementary schools, one middle school, and one high school. During these observations they took notes on the type of language teachers use in the classroom, and on the nature of oral exchanges between teachers and students. The complete report on the needs assessment trip is available in Appendix A of this report.

3. Development of the GETEP Multiple-Choice Sections

This chapter describes the process of developing the multiple choice parts of the GETEP. The preparation of test and item specifications; the process of writing, reviewing, and revising items; the results of the field testing of the pilot forms of the GETEP, and its subsequent revisions are discussed in this chapter.

3.1 Development of Test and Item Specifications

In light of the findings of the needs assessment study, CAL staff drew up specifications for the test. The specifications were used as a guide by all item writers who worked on the project. This procedure ensured that each item was designed to test understanding of specific linguistic features of English, and that the test was developed in a way that was responsive to local needs as determined in the needs assessment. A complete copy of the specifications are included in Appendix B.

Below are the number of items that were chosen to be included in each part of the multiple-choice sections of the GETEP.

LISTENING COMPREHENSION	
* Short Dialogs	20 items
* Extended Dialogs and Monologs	30 items
Total	50 items
READING COMPREHENSION	
* Cloze	20 items
* Reading Passages	30 items
Total	50 items
WRITING PROFICIENCY	
* Error Detection	30 items
TOTAL FOR MULTIPLE-CHOICE SECTIONS	120 items

3.1.1 Listening Comprehension Specifications

As shown above, the original specifications for the listening comprehension section called for two parts containing a total of 50 items. (Note: After field testing, the short dialog part was reduced from 20 to 15 items, and the extended dialog and monolog part was reduced from 30 to 25 items. See section 3.5.1 for details on these revisions.)

For the short dialog items, twenty categories of grammatical, syntactical, rhetorical, and phonological features of spoken English were specified, all of which were to be represented on each form of the test. Each item, then, tests a specific aspect of listening comprehension. To simulate classroom conditions, all the short dialogs would be based on one of three types of interaction: teacher/student, student/student, and teacher/other adult (e.g., school principal, nurse, counselor, custodian or parent). Thus, each stimuli would involve two of four possible speakers: adult male, adult female, student male, and student female.

It was decided to have the test focus on the language the teacher would have to understand in a real life situation. Thus, in a short dialog between a teacher and a student, the question should require understanding of the student's statement in the dialog, and in addition, should require comprehension of both utterances (that of speaker 1 and speaker 2) to answer.

Since in the classroom the teacher often has to comprehend a student initiated question, this type of listening activity would be tested using the following basic format:

SM: Sentence. Sentence. Question.
AF: Response.

Question: What is the student's problem/concern, etc.?

SM = Student/Male
AF = Adult/Female

Of course, sometimes the teacher makes statements in front

of the class and then follows them with a question to a student, in which case the student would give a response. The test question would then focus on the details of the student's response. Thus, the test questions would focus on comprehension of the student's language in a student-teacher exchange or of the other party's language in an exchange between a teacher and another adult.

In contrast to the short dialog items, the extended dialog and monolog items were intended to test less specific, more global aspects of listening comprehension. They would focus on testing: 1) understanding of the main topic of the dialog or monolog, 2) comprehension of the use of supporting ideas presented in the conversation or talk, and 3) the ability to make inferences based on information presented in the dialog or monolog. The extended dialogs, like the short dialogs, would be between teacher and student, student and student, or teacher and other adult. Likewise, the monologs would all be related to an educational setting.

3.1.2 Reading Comprehension Specifications

As shown above, the specifications for the reading comprehension section called for two parts, both containing multiple-choice items.

It was decided to use the multiple-choice, rational deletion, cloze format as the first part of the reading comprehension section. Previous research has shown that this item type has the ability to test language skills in a way that is both integrative and communicative. The specifications called for two passages for this part, with 10 cloze items in each passage. The majority of the items would focus on testing reading comprehension, as opposed to testing the vocabulary or syntax that was called for by local constraints within a single sentence. However, the individual cloze items did make use of either vocabulary or syntax to test passage comprehension. In developing specifications and writing cloze items for this part,

staff were influenced by the approach used by Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988).

Specifications for the reading passages part called for more traditional reading comprehension items: passages of up to several paragraphs in length followed by five to six questions testing (as in the extended dialog and monolog part of the listening comprehension section described above): 1) understanding of the main idea or topic of the passage, 2) comprehension of supporting ideas and 3) the ability to make inferences based on information presented in the passage. Questions involving an awareness of rhetorical organization or requiring an analogy or interpretation of the passage would be de-emphasized, since they invoke reasoning ability as well as language proficiency. It was decided that each form would contain 30 items in this part (5-6 passages with 5-6 items each).

3.1.3 Writing Proficiency Specifications (Error Detection Items)

Since teachers are frequently required to read and correct students' writing and to write original material themselves, it was decided that the writing test should include two parts, corresponding to these two activities. Thus, it consists of a multiple-choice error detection exercise and a holistically scored written essay.

The specifications for the error detection part call for three paragraph-long passages simulating student-written compositions. In each line of these passages three words are underlined and the letter A, B, or C is written below each. In the right hand margin of the passage, there is a fourth option, labeled "no error," and the letter D is written below it. In each line, one of the underlined words may, in the context of the sentence, contain a lexical, grammatical or syntactic error. The examinee must decide which of the underlined words needs to be changed, or whether there is no error in the line.

The development of specifications for this section was based

on an analysis of writing samples from Freshman students at the University of Guam. Common errors were identified and these were incorporated into the error detection item stimuli. The errors made fall into two general categories: the kinds of errors typically made by nonnative English speakers, and the kinds of errors made by native English speakers. Thus, the errors to be detected on this part of the test are similar to those found on the Written Expression portion of the TOEFL and on the Errors in Usage portion of the Test of Standard Written English. A number of other standardized tests also contain this type of item. The errors made in these essays were considered to be representative of both the kinds of errors that some teachers on Guam might make, and the kinds of student errors teachers on Guam should be able to correct.

3.2 Writing Items and Assembling Field Test Forms

The writing, reviewing, and revising of the GETEP multiple choice items represent the collaborative effort of a team of language testing specialists from CAL staff and consultants, together with consultants with experience in the Guam public school system.

3.2.1 Preparing Listening Comprehension Items

After the first draft of all the listening comprehension dialog and monolog stimuli and the test questions was completed, the items were reviewed both by CAL staff and by consultants familiar with the educational situation on Guam. During the review, the items were inspected for any possible problems, such as unclear questions, more than one key (correct answer), no key, or dialogs or language unlikely to occur on Guam. The dialogs were checked to ensure that they were as realistic as possible, and that each might conceivably occur in a classroom or school on Guam.

After the first review, the items were subsequently revised and then reviewed again. This process continued as necessary

regulations. Textbooks and scholarly papers on education on Guam and in Micronesia available on microfiche in CAL's ERIC collection also served as sources. These passages were reviewed for level of difficulty, rhetorical organization, clarity of expression, accessibility to a general educational audience, relevance to education on Guam, and suitability as texts for multiple-choice reading comprehension and cloze items.

Once passages were selected, a first draft of the cloze items was prepared. These were reviewed and revised by CAL staff and consultants. Although the test specifications called for 10 items per passage, 11 or 12 were prepared for each selection. This was done so that poorly performing items could be deleted after the field testing. Thus, it would not be necessary to pretest new items for the final form.

After reading comprehension questions were drafted for the longer passages, they were reviewed by CAL staff and consultants and then revised by the original authors. All items were again reviewed by the entire project personnel, and revised as necessary. In this manner, GETEP items received extensive and multiple reviews by language testing specialists.

Once the two parts of the reading comprehension section were completed, they were trialed on CAL staff who were not involved in the construction of the test. Two or three people took each form, for a total of 10 examinees. Five were native and five were non-native English speakers. Their performance on these items was examined and items that appeared to be too easy, too difficult, or confusing were revised. The examinees were also asked to point out any problems they saw with the test directions or test items. These comments were considered in making the final revisions for the field test version of the test.

3.2.3 Preparing Error Detection Items

In preparing these items, suggestions for topics for passages to be used in this part were put on paper and reviewed by the test development staff. Once approved, draft passages of

until all parties were satisfied with the items. Once all revisions were completed, a script of the listening comprehension stimuli were prepared for recording in a professional sound studio in Washington, DC. The use of a commercial, advanced-technology facility ensured the fidelity of the voices in the recordings. Then, universities and high schools in the area were asked to identify drama students who could serve as voices for the tape. Auditions were held at the recording studio and the suitability of each potential speaker was discussed by CAL staff and the staff of the recording studio. Eventually, a cadre of high quality speakers from appropriate age groups was identified. A professional radio announcer was contracted to read the directions to the test and to announce the number of each new item. To minimize the possibility that an examinee may confuse the speakers in a short dialog or extended conversation, it was decided that a male and a female voice should alternate in speaking. Further contrast between speakers is provided by the use of a student and teacher's voice in many dialogs. In this case, the teacher's voice is always more mature sounding than the student's voice.

After the tapes were recorded, they were listened to by CAL staff and any infelicities that were not detected and corrected at the original recording session were identified. Subsequently, the studio carried out minor editing on the tapes for the four versions of the test.

3.2.2 Preparing Reading Comprehension Items

Passages were selected by CAL staff to use in both the cloze and reading passages part of this section of the test from material likely to be read by teachers on Guam or similar to what teachers on Guam might read. Sources included educational periodicals such as Educational Leadership, NEA Journal, and Phi Delta Kappan; newspapers from Guam and the mainland such as the Pacific Daily News and Education Week; Guam DOE materials, such as memos, publications, research reports, curriculum guides, and

simulated student writing were prepared in three genres in which secondary students are frequently required to write: personal narrative, descriptive writing, and persuasive writing. One sample in each genre was written for each form. Each passage contained errors in about three-fourths of the lines, and no errors in one-fourth of the lines. These items were reviewed by CAL staff and revised where potential problems were found.

To further simulate student compositions, the passages were written neatly in long-hand in the GETEP test booklet. (Note: although these error detection items form part of the writing proficiency test, in the final form of the GETEP they have been placed in the same test section as the reading comprehension items. This allows the examinee to do all the multiple-choice items, other than the listening comprehension items, in a single uninterrupted period of time.

3.2.4 Preparing the Field Test Forms of the GETEP

The final pre-field test version of the multiple-choice parts of the GETEP were formatted on an IBM-XT microcomputer in WordPerfect 5.0 and printed on an HP Laserjet Series II printer. This produced camera-ready copy quality products for the field test administration.

3.3 Field Testing the Multiple-Choice Sections

The four forms of the multiple-choice sections of the GETEP were administered as pairs to two groups of examinees on two different testing dates. Forms 1 and 3 were administered to Group A on October 14, 1989; Forms 2 and 4 were administered to Group B a week later.

About 90% of the members of Group A were current students majoring in a variety of disciplines at the University of Guam enrolled in the Reserve Officers Training Corps (ROTC) program; the other 10% were teachers. Group B consisted of approximately 80% teachers currently employed by the Guam DOE; the other 20% were ROTC cadets. In total, 40 examinees were present for the

Group A administration; however, only 34 completed both exams due to prior commitments which precluded their participation in the entire session. All 38 examinees present for the Group B administration completed both exams. Seven examinees took all four forms (i.e., were present for both Group A and Group B administrations).

68% of Group A were nonnative English speakers while only 55% of Group B were nonnative. Table 3.1 presents the self-reported native language background of the two groups (Group A includes only those who took both forms).

Table 3.1
Native Language Background of Field Test Participants

Group A		Group B	
English	32%	English	45%
Chamorro	53%	Chamorro	29%
Filipino	9%	Filipino	26%
Other	6%	Other	0%

In each test administration, approximately half of the subjects took one of the two forms first, while the rest took the other form first. Table 3.2 presents the number of subjects that completed each form in each group, the mean total score on all items, and the standard deviation of the score distribution.

Table 3.2
Means and Standard Deviations on the Total Test

		Mean	Std Dev
Group A			
Form 1	(n=34)	96.1	11.80
Form 3	(n=34)	101.0	12.70
Group B			
Form 2	(n=38)	106.4	14.93
Form 4	(n=38)	105.3	14.20

From the data presented in Table 3.2, it seems that both groups performed quite well on the test. Group B performed slightly better than Group A, which is reasonable since Group B contained a larger percentage of native speakers. In addition, Form 1 appears to be slightly more difficult than the others. However, given the small size of the sample, the differences in means were not statistically significant.

Table 3.3 presents the means scores and standard deviations for each form by section.

Table 3.3
Mean Scores and Standard Deviations by Sections

Form	Listening	Cloze	Passages	Error Detection
1	40.9 (4.02)	17.7 (3.15)	15.8 (3.40)	21.7 (4.33)
3	39.4 (3.90)	19.3 (2.33)	18.5 (4.53)	23.8 (4.29)
2	41.8 (5.67)	19.9 (3.11)	19.5 (4.92)	25.1 (3.77)
4	42.3 (4.69)	18.8 (4.28)	19.6 (5.16)	24.7 (2.72)

The data presented in Table 3.3 suggest that the Form 1 Cloze, Passages and Error Detection parts were slightly more difficult than those on the other forms, which appear to be equivalent in difficulty.

Table 3.4 presents the mean scores of the seven examinees who took all four forms. Although this is a very small group, its performance can give clues as to the comparability of the GETEP forms.

Table 3.4
Mean Scores and Standard Deviations on Test Sections for the Seven Subjects Who Took all Forms

Form	Listening	Cloze	Passages	Error Detection
1	41.6 (3.60)	17.9 (3.18)	16.3 (3.50)	22.6 (2.82)
3	39.7 (4.27)	19.9 (2.41)	20.7 (4.82)	25.0 (4.04)
2	44.0 (4.69)	20.3 (2.69)	18.3 (4.23)	25.0 (4.12)
4	44.1 (2.41)	18.6 (3.31)	19.4 (3.60)	25.1 (2.19)

Comparing the data in Tables 3.3 and 3.4, we can see that in general the performance of these seven examinees on the different sections and different forms is very similar to those of both Group A and Group B in general (with the exception of the unexplained better performance on the Listening Comprehension section on the second test day). These data suggest that groups A and B were generally comparable and that the test forms were generally equal in difficulty, though the Cloze, (reading) Passages and Error Detection parts of Form 1 may have been slightly more difficult than those parts on the other three forms. This characteristic of Form 1 was kept in mind in making revisions, especially in the Error Detection section. The revision process is discussed in Chapter 6 of this report.

Finally, Table 3.5 presents the KR 20 reliability estimates obtained for the field test forms, by section.

Table 3.5
KR 20 Reliability Estimates for the Field Test Forms by Section

Form	Listening Comp.	Reading Comp.	Error Detection
1	.73	.73	.90
2	.83	.88	.79
3	.65	.80	.79
4	.75	.89	.52

The reliability estimates given in Table 3.5 are encouraging

given the fact that KR 20 reliability estimates are extremely sample dependent. This means that for any given test form, the reliability estimates will be highest when the mean of the sample population is approximately at the mid-point of the possible range of scores and the sample is very heterogeneous, with examinees covering the range from very low to very high scorers. The mean scores of the sample on which the GETEP was field tested were generally well above the mid-point of the range, especially for Listening Comprehension (ranging from 79% to 85% correct on the different forms) and Error Detection (ranging from 72% to 84% correct). The examinee sample, too, contained a majority of native English speakers or Chamorro speakers, who had received their education in English (85% of group A, 74% of group B). In general, there were no examinees with very low scores in the field test sample; the group was rather homogeneous in ability, creating a ceiling effect. In light of the above, the reliabilities obtained on the field test sample are very supportive. Additionally, it must be remembered that these estimates are from the field test form and not the final form. Since field test forms have been revised according to the results of statistical item analyses (see below), it can be expected that the reliability of each section of the final form of the GETEP is even higher than those in Table 3.5. Since the revised, final test forms were not re-administered, it is not possible to present reliability estimates for the final forms here. Ultimately, as indicated above, the reliability of the final forms will depend on the heterogeneity of the operational population. We believe this population is more heterogenous than was the pretest sample.

3.4 Statistical Item Analyses and Revisions to the Field Test Forms

Three types of item analysis were conducted on each multiple-choice item for each section and each form of the GETEP. Item statistics on the item difficulty (p values), item point-

biserial correlations with the total test score, and item discrimination by performance of the top 25% and bottom 25% of the examinees were computed. All items that had a difficulty value greater than .85 (i.e., were very easy), and/or a low or negative point-biserial correlation, and/or a discrimination index below .3 (i.e., were not discriminating well) were examined as candidates for revision. In the vast majority of cases, such items were revised. The procedures used to revise items for each section are described below.

3.4.1 Revision of Listening Comprehension Items

The descriptive statistics presented in Tables 3.3 and 3.4 indicate that there was no real difference in the difficulty of this section across forms within and between groups. Scores on the listening section of the test, however, were quite high considering that there were only 50 items in the section. Mean scores reveal an approximate average score of 80% correct across forms. In addition, a few of the questions on each form were answered correctly by all of the examinees. Thus, it was decided that this section could be shortened without any serious loss in its ability to separate the less able from the more able examinees. The number of Listening Comprehension items in the final form was reduced from 50 to 40; five items were deleted from the first 20 short dialog items and five items from the 30 extended dialog and monolog items.

In selecting items to remove from the first section, first those items answered correctly by all examinees were deleted, and then those answered incorrectly by only one or two examinees were deleted. Such items did not help to discriminate between the higher and lower ability examinees. Despite the removal of the five items from this section, listening items with a large variety of characteristics remained. However, unlike on the field test form, in which the same numbered item on each form had the same item content characteristics as per the original specifications, each of the listening forms in the final version

is slightly different in terms of the item content characteristics it contains. Appendix B contains the original specifications for each item. Appendix C presents the specification numbers of the short dialog items that remain on each final test form.

In removing items from the extended dialog and monolog section, again the goal was to delete items that were answered correctly by all or almost all of the examinees. However, care was taken so that no items were removed if doing so would leave less than four items for any extended conversation or monolog. Care was also taken to ensure that an adequate representation of the three types of items (Main Topic, Supporting Ideas, and Inferences) was kept on each form. None of the original recordings of the extended dialogs or monologues were changed in any way.

Reducing the length of a measure may have a negative impact on its reliability. By examining the performance of the field test examinees on the items remaining in the test, it is possible to get an estimate of what the shortened test's KR20 reliability would have been, had it been administered to the same field test sample. These estimates are presented in Table 3.6.

Table 3.6
Estimated Reliabilities of the Shortened Listening Comprehension Section

Form Number	Field Test Results		Shortened Version Estimates	
	Mean	KR20	Mean	KR20
1	40.9	.73	32.56	.75
2	41.8	.83	32.18	.83
3	39.4	.65	31.92	.69
4	42.3	.75	32.95	.75

From the above statistics, it is clear that shortening this section by removing the easier items did not hurt the its

reliability; for Forms 1 and 3 it even improved reliability. The mean scores across the forms are also more nearly equal, indicating that four forms are more parallel in difficulty. Thus, although each test form now has different specifications for the characteristics of the short dialog items, each test form is of a higher quality in measuring listening comprehension ability.

3.4.2 Revising Reading Comprehension Items

Test specifications for the GETEP included 10 items per cloze passage. However, each passage for the cloze section of the GETEP field test contained 11 or 12 items (i.e., one to two extra items), as it was envisioned that the best 10 items for each passage would be kept in the final form. For most of the texts it was possible to remove items that all or almost all of the examinees got correct. In this way, performance on the other cloze items on the final version will be minimally affected by the removal of these items. Item analyses revealed eight instances in which the options on remaining items needed revision. In most instances this was because there was no one clearly best answer from among the choices given. In one instance, analysis revealed that there were insufficient clues in the passage to restore the word that had been selected for testing. In this one case, a totally new item was created from a word located nearby in the passage.

Again, the descriptive statistics given above for this section indicate that there were no significant differences in its difficulty within or between groups. The revisions undertaken to improve this section worked to make the section slightly more difficult and improve its discrimination and thus its reliability. Since options have been revised on some items (in addition to the deletion of some items), it is not possible to give meaningful estimates of this section's KR20 reliability on the basis of the field test data.

As regards the reading comprehension items for the longer

passages, the item analysis revealed that the majority of the reading comprehension items were of sufficient quality. Nevertheless, there were several items on each form for which the options needed to be revised. In most of these cases, one or two incorrect options were very close to the correct answer. Item analysis revealed that in some of these cases, examinees who were in general high scoring choose an incorrect option that may have been correct given an alternative interpretation of the reading passage. Such items were revised by making these options more clearly incorrect. In a few cases, a correct option was unclear and needed revision to make it more clearly correct. In a few more cases, the stem section of the item needed to be revised to make the item clearer. Only one case involved totally revising both stem and options of the item.

The number of items needing revision was greatest on Form 1 (12), as could be expected from the examinee performance on it compared to on the other forms. On Form 2 only four items needed revision, on Form 3, nine, and on Form 4, seven. These revisions will make the forms of more equal difficulty and better able to discriminate between more able and less able examinees.

Since options have been revised on some items, it is not possible to give meaningful estimates of this section's KR20 reliability on the basis of the field test data.

3.4.3 Revising Error Detection Items

Each field test form of the test contained three error detection passages with 10 items each, for a total of 12 passages and 120 items. Although these items were generally easy, item analysis (in terms of the items' ability to discriminate between more successful and less successful examinees) revealed that the vast majority of passages were adequate as originally written: for six of the passages only one revision of an option was required, for an additional three only two revisions were required. Items requiring revision in this section were generally either too easy or else non-discriminating. When

appropriate, in cases where the error was too obvious, the revision was to correct the error so that the correct answer to the item became option D, "No Error." Non-discriminating items were those where some of the higher scoring examinees either choose an incorrect option when D "No Error" was the correct answer, or choose D when the error was too difficult to find. In these ambiguous cases, options were revised to be more clearly right or wrong. In only three cases were the sentences themselves changed, and then only in a very minor way.

Tables 3.3 and 3.4 indicated that the Error Detection section of Form 1 was slightly more difficult than that of the other forms. Item analysis indeed revealed that one passage on Form 1 needed fairly extensive revision (four items) as it was above average in difficulty, and one passage on Form 2 needed five items revised, as it was originally quite easy. The net effect of the minor revisions to this section is to make the four forms more comparable and improve the section's ability to discriminate between more and less able examinees. Since options have been revised on some items, it is not possible to give meaningful estimates of this section's KR20 reliability on the basis of the field test data.

3.5 Preparation of the Final Forms

The revisions were entered and formatted in Wordperfect 5.0 on an IBM-XT microcomputer and printed in camera-ready copy on an HP Laserjet Series II printer for duplication for the operational testing program. Although no changes were made to the short and extended dialogs and monologs recorded for the field test version, since the number of the items was reduced, the same professional announcer used for the field test version recorded the new item numbers and questions, and the field test version tapes were professionally re-edited to correspond to the new numbering.

4. Development of the GETEP Writing Sample

This chapter describes the development of the writing prompts and scoring guidelines for the GETEP Writing Sample (GWS). It also describes the training of the scorers for the sample and gives results of the field testing of two of the writing prompts.

4.1 Specifications

The Writing Sample on the GETEP (GWS) is a test of productive writing ability based on a 30 minute response to a single prompt. The writing task required by each prompt is similar to the kind of writing that a teacher might have to carry out while teaching on Guam. The GWS is scored on a 1 to 5 scale by two trained raters using a modified holistic scoring procedure.

4.2 Developing Writing Prompts

As described in Chapter 2, Jodi Crandall and John Karl of CAL visited Guam as part of a needs assessment trip to acquire information for developing an English proficiency test for the Guam DOE in November, 1988. One of the questions asked during the needs assessment trip was "What kinds of writing must a teacher on Guam do?" Through talking with a number of educators on Guam, observing classrooms, and speaking with professors in the English Department and the Department of Education at the University of Guam, Crandall and Karl identified a number of writing tasks that a teacher might have to perform on one occasion or another. Based on the information gathered, a list of 12 types of writing assignment tasks and 5 potential topics that should be considered for the direct writing test was developed (see Appendix A for the complete list).

These tasks were examined carefully by CAL staff and subsequently classified into four general question types: 1) comments on or evaluations of school programs or facilities, 2)

personnel, and 4) letters to parents. The complete classification is presented in Table 4.1 below.

Table 4.1
Classification of Guam Writing Prompts
Based on Teachers Input

Comments/evaluations of (defending an opinion/argumentation)

1. staff development needs
2. program of extracurricular activities
3. school lunch program
4. wish list of supplies, equipment, facilities
5. student teacher

Comments on proposed changes in (defending an opinion/argumentation)

1. the school year/school day
2. promotion and retention policies
3. the curriculum
4. teacher evaluation criteria and methods
5. teacher certification or recertification requirements

Memoranda on (description)

1. how to involve parents in education
2. how to work with parent volunteers or teacher aides
3. how to organize professional development activities
4. discipline in the classroom
5. dealing with a student who doesn't speak English
6. your school/class for prospective teacher or student teacher
7. motivating students
8. referring a student to special education
9. requesting permission for a field trip (to the principal)
10. discipline referrals
11. a lesson plan, suggesting ways to implement it (to the substitute teacher)

Notes to parents on (description/argumentation)

1. student problems (identifying the problem, suggest ways in which the teacher is working to help, and ways in which parents might be of help)
2. explaining purpose/function of an extracurricular activity
3. encouraging use of the library/ other facilities
4. how to get/find something in the school
5. encouraging involvement/oversight of homework

Using these four general question types, CAL staff developed 12 writing prompts. These were then revised by the project director (Stansfield), who took them to Guam on his first visit there in October, 1989. He showed the prompts to UOG English professors Gene Bruce and Evelyn Flores. Following extensive discussion, each prompt was revised, as required. These question prompts have been turned over to the Guam DOE for use in the GETEP operational program.

The instructions for the administration of the GETEP Writing Sample appear in Appendix D.

4.3 Field Testing

Once the prompts were finalized, the project director selected two prompts which appeared to be accessible to a general audience. These were administered to two Developmental English classes and two Freshman English classes at the UOG. A total of 67 students responded to these two topics, which dealt with proposing criteria for an excellence-in-teaching award, and how money obtained through an unrestricted grant to the school should be spent. These 67 writing samples were then scored by Kenyon and Stansfield using the Test of Written English (TWE) scoring guide.

4.4 Developing the GWS Scoring Guide and Training Raters

From the start of the project, it was felt that a modified holistic scoring procedure would offer the greatest possibility of reliable scoring. Holistic scoring is based on the reader's overall impression of the communicative writing ability of the examinee. A modified holistic procedure is based on a scoring guide. The scoring guide assists the rater in classifying the writing sample into a single category or score by providing a basic description of ability for each category as well as certain features that exemplify writing at that level. The use of a scoring guide also contributes to the "anchoring" of scores by ensuring that writing samples written in response to different

prompts are graded on a common scale. This anchoring also helps to prevent "drift" among the raters, who may alter their standards slightly over an extended period of time. A scoring guide is accompanied by a set of "benchmarks," which are examinee writing samples that exemplify each point on the scoring guide.

CAL staff reviewed three holistic scoring guides that had previously been developed for grading writing. These guides were those used by the Test of Written English (TWE), General Educational Development (GED) Essay Test, and the National Teacher Examination (NTE). Each guide was judged to be potentially applicable to the population of examinees that could be expected in the GETEP operational program that will be administered by the Guam DOE. Permission to use each scoring guide was obtained from the test publisher that developed it.

In January 1990, during a second trip to Guam, Stansfield trained four educators to score the GWS. The GWS scoring guide was developed as a part of this process. The four educators trained are Margaret Camacho, Julie Sisson, Francis McDonald, and Marie Barretto.[1] All are current or former language arts teachers in the Guam public schools; two are currently administrators.

These educators underwent fourteen hours of training in holistic scoring. The training began with a review of the TWE, NTE, and GED scoring guides. It was decided that the GED guide would not be as useful for this examination as the others. The GED guide places greater emphasis on language than on rhetoric. Also, the GED guide appears to be written with descriptive writing in mind, while the GWS prompts include an element of persuasion.

The group began its training by learning to analyze an essay prompt using the Situation-Problem-Solution-Evaluation method of analysis developed by Hamp-Lyons (1989). The two prompts that

---

[1] Dan Robertson, CAL consultant on this project and professor at UOG, was also trained at this time.

were administered to UOG students were analyzed in this manner. Then, the raters read four benchmark writing samples selected from the ratings assigned by CAL staff using the TWE scoring guide. The rhetorical and syntactic characteristics were discussed. Four additional writing samples were read and scored using the TWE scoring guide, and then discussed. A third group of four writing samples were then read, scored and discussed. All of the writing samples fell within TWE scores 3-6, since all were obtained from university students at an English speaking university. The modal TWE score assigned to the writing samples was 5.

At this point, the raters studied the NTE scoring guide and the published benchmark essays representing the six points on the NTE scale (Educational Testing Service, 1989). The raters then scored four more GWS writing samples using both the TWE and the NTE scoring guides. Their scores were compared and discussed. Subsequently, the two scoring guides were discussed. Then the process was repeated.

Next, the raters were shown a modified TWE scoring guide developed by CAL staff based on the performance of the 67-subject UOG sample. This modified guide contained five score levels. The group discussed the guide and then rated a number of papers using it. Then the group rated four papers, assigning scores using all three guides. This was followed by a discussion of the suitability and appropriateness of each guide to the Guam DOE pool of applicants. The process was repeated. At the end of the first day of training, the raters decided that the modified TWE guide is more suited to their examinee population than is the official TWE guide.

During the second day of training, the raters began by scoring writing samples using the modified TWE guide and the NTE guide. After each set of four writing samples was scored, the inter-rater agreement of the group was determined and the two scoring guides were discussed. Soon it became clear that the group preferred the modified TWE guide. After this was decided,

the group scored two more sets of four writing samples and discussed the modified guide following the scoring of each set of writing samples. During these discussions several changes were made in the guide. The final exercise was the uninterrupted scoring of 20 UOG writing samples using the new GWS guide. Following the scoring, the guide was again discussed and one final minor change was made.

The lowest level of this GWS guide combines TWE levels 1-3 into a single GWS level. This is the level the GWS guide associates with incompetence in writing. Level 2 is approximately equivalent to TWE level 4. GWS levels 3-5 represent higher levels of writing ability, with level 5 approximating level 5 performance on the NTE scoring guide. This guide is presented in Appendix E, together with benchmark writing samples for levels 1, 2, 3, 3.5, 4, and 4.5.

4.5 Inter-rater Reliability

As mentioned above, at the end of the training the four raters independently rated 20 writing samples using the GWS scoring guide. The data produced from this scoring was used in a generalizability study to determine the reliability of an examinee's rating during the operational program. The generalizability coefficient obtained was .80. This is an estimate of the reliability of an examinee's score when the examinee's essay is rated by any two of the four raters trained to score the GETEP Writing Sample. The generalizability study produced an estimate of the standard error of measurement of .43. This may be interpreted to mean that an examinee's true score on the GWS would have a 95% probability of falling within .84 points of the composite score awarded to the examinee on the GWS.

4.6 Standard Setting

Upon completion of the training in holistic scoring and the selection and revision of a suitable scoring guide, the raters had a lengthy discussion of an appropriate score standard.

Nearly all members of the group felt that GWS level 4 should be set as the level of acceptable performance. However, Stansfield brought out that this would normally require an agreement by two raters that a paper indeed merited a four, since each paper would be scored twice. It was further noted that the group's ratings of some of the level four papers included some ratings below level four. These were typically at level 3. As a result, it was decided to recommend 3.5 as the acceptable minimum composite score. Under such circumstances, a passing score would typically require at least one rater to perceive the paper to be a 4. There was unanimous agreement among the group that 3.5 was a fair and appropriate level of writing competence to expect from new teachers seeking certification on Guam. Indeed, given the standard error of measurement (.43) produced by the generalizability study above, an examinee receiving a composite rating of 3.5 would be unlikely to have a true score as high as four or as low as three, since the 67% confidence interval or range encompassed by one standard error of measurement for tha examinee would extend from 3.07 to 3.93. Similarly, the confidence interval around the score of an examinee receiving a composite score of 3.0 would not reach the passing score of 3.5. Applying statistical theory further, we can estimate the less than 16% of the examinees with an obtained score of 3.0 would have a true score of 3.5. When one remembers that nearly all raters felt that 4.0, not 3.5 represented acceptable performance, and constructs a confidence interval on that basis, then one can predict that only about 1% of examinees with an obtained score of 3.0 would have a true score of 4.0, or a writing ability that the group would find acceptable.

5. The Oral Proficiency Interview

In this section, we will describe the training that CAL provided in order to implement the Oral Proficiency Interview (OPI) on Guam.

5.1 Background

Early in the project, CAL decided that an oral proficiency interview (OPI) would be the appropriate measure of speaking proficiency to use in Guam. It might have been possible to develop a semi-direct speaking test of ESL, similar to the ones that CAL has developed in a number of other languages (Stansfield and Kenyon, 1989). However, funding to cover the fairly high cost of developing and validating such a test was not available. In addition, the Guam DOE had been using a semi-direct speaking test for the past three years and the DOE reported that this test did not seem to be doing a satisfactory job of identifying applicants with problems in communication. As a logical alternative to developing a semi-direct test, the OPI had the advantages of high face validity, established construct validity, and acceptability in the field. A great deal of information has been published that describes the oral proficiency interview and numerous other publications provide a detailed explanation of interviewing techniques, the accompanying criterion referenced scale, and the many situations in which the OPI is used (Liskin-Gasparro, 1987; Buck, 1989; Clark, 1978). Thus, it was decided to include an OPI as part of the GETEP.

5.2 Training Interviewers

In September, the project director went to Guam for 10 days. The main purposes of this visit were to train oral proficiency interviewers, to administer the pretests to the first group of pretest examinees, to obtain feedback from local writing specialists on the writing prompts developed by CAL staff, and to consult with DOE personnel. The training for the OPI consisted of

Stansfield met for four days with four speech pathologist/ESL teachers on Guam. They were Art Wheeler, Gerri Diaz, Bonnie Sarempa, and Cathy Cardenas. On the first day, the group met at the DOE. The project director introduced the group to the OPI and discussed its use in teacher testing and in general language proficiency assessment on the mainland. He also introduced them to the skill level descriptions for speaking. The interviewers-to-be were given copies of the ACTFL and the ILR versions of the level descriptions. These were discussed in detail, in order to give the interviewers a clear idea of what tasks a learner could perform at each level. Next, the interviewers listened to a single tape recorded interview. Following the interview, the examinee's performance was discussed at length. The characteristics of the performance were related to the skill level descriptions and an appropriate level was assigned the interviewee.

On the second day of training, the interviewers listened to and critiqued eight other taped interviews provided by CAL. During the morning, the focus of the discussion was on an appropriate rating. During the afternoon, the discussion turned to interviewing technique. At this point, the project director described the four phases of the interview (warm-up, level check, probe, and wind-down) in depth and provided examples of questions for each phase of the interview. A discussion of techniques for dealing with each individual's area of specialization was held also. Then, the group listened to and critiqued more taped interviews, this time focusing on both the rating and the interviewer's technique.

On day three the group met at the University of Guam, in order to interview a number of nonnative English speaking students enrolled in either ESL or regular university classes. The first interview was conducted by the project director. Afterwards, there was a discussion of the characteristics of the examinee's speech. These characteristics were related to the function, content, and accuracy characteristics of the various

skill level descriptions. The interviewers indicated the ability level of the interviewee. The discussion then turned to the interview itself. The project director critiqued his own interview and the speech teachers critiqued it also, based on what they had been told about interviewing techniques. Next, another student was interviewed; this time by one of the speech teachers. This interview was followed by a group discussion of an appropriate rating and a critique of the interview. The process continued in this way for the rest of the day. During th day, the interviewers were introduced to the role of role-plays (situations) in carrying out an interview. They were provided with the situation cards developed by ACTFL. Subsequent interviews included the use of two situation cards per interview. Whenever a scheduled interviewee failed to show, the group listened to and critiqued additional taped interviews provided by CAL.

On day four, the group met again at the DOE, where it conducted additional interviews. The focus of training on this day was techniques for determining the ability to speak on a wide variety of topics (content) and on one's special field of competence. CAL provided the interviewers with a list of topics for examinees at each level, as well as a list of current topics of national and worldwide interest. It was pointed out that topics of current national and worldwide interest may change over time, and methods of identifying current topics (such as checking the local newspaper of a national newspaper, listening to the news on the radio, etc.) were discussed. Through group discussion, topics of local interest were identified. Similarly, topics of interest to local educators were identified, such as current problems of children on Guam, DOE policies, etc. General topics of interest to educators were also identified, such as the relationship between homework and school achievement, discipline in the classroom, school attendance, dropout prevention, vocational education, work-study programs, parental involvement, etc. These topics, it was indicated, could all be used to enter

into a general discussion of education on a fairly high level. A written list of current topics of general interest was provided to the speech teachers.

As "homework" the speech teachers were asked to interview at least three teachers in the public schools, to record the interview, and to indicate their rating of the examinee's ability. These interviews were to be conducted during the subsequent weeks and sent to the CAL office in Washington, DC. It should be pointed out that the speech teachers generally listened as a group to each taped interview after it was conducted. In this way, they were able to provide an insightful analysis and an accurate rating of the interviewee's language proficiency. After the interviews were received at CAL, they were listened to by the project director, who prepared a written critique of each.

During his second trip to Guam, Stansfield gave the tape and the written critique to each interviewer. In addition, he also played several of the interviews to the group for further discussion of both the interviewing technique and the appropriateness of the rating that had been assigned. The interviews played at this session were those that posed special problems for rating because of the atypical nature of the examinee's language proficiency. Such examinees are usually very strong in reference to some criteria related to the skill level descriptions, while unusually weak on others.

Upon meeting with the speech teachers on the second visit to Guam, it was apparent that they were now very conversant with the structure of the interview and with the scale. All had a good feel for the OPI.

CAL feels that these trained interviewers will be able to provide valid, accurate interview ratings in the future. However, a number of things can be done to maintain the skill of the interviewers in carrying out and scoring the interviews. These are listed in the chapter 7 entitled "Recommendations for the Operational Program."

6. Setting the GETEP Passing Scores

6.1 Background

The use of tests for certification and licensing is commonplace in many professions and occupations. During the 1980s, the use of teacher competency examinations has increased substantially in response to the demands of parents, business leaders, and politicians for the improvement of public education. For example, in 1981 only seven states required the National Teachers Examination (NTE). At present, the NTE is required by 35 states. The majority of the remaining states have developed and currently administer their own teacher competency examinations.

Passing scores on professional and occupational examinations, including teacher competency examinations, are designed to answer, in a rational way, the question, "Exactly how good is good enough?" The process one follows in answering this question is often referred to as "standard setting." Standard setting involves bringing together a group of judges to determine what is acceptable performance. This determination should be made on an absolute basis as opposed to a norm-referenced basis. That is, those (judges) involved in setting standards must determine whether the examinee is safe and effective generally, not whether he or she appears to be safe and effective in comparison with other examinees. The purpose of such a determination is to protect the public from incompetent practitioners.

6.1.1 Judges

The determination of a passing score requires the collection of judgements. These judgements must be made by individuals (judges) qualified to make such decisions. Scores set by persons who do not have a good knowledge of the ability being assessed will have neither validity nor reliability. Research has shown

the judges can affect the passing score that is established. Because of this, information on the background and qualifications of the judges should be gathered and included in the report.

Typically, judges should be masters of the subject. However, there are other criteria to consider as well. It seems appropriate to select judges that represent, to the degree possible, the different groups of individuals who are concerned about competency, as well as those likely to be affected by the outcome. Thus, a panel of judges that is going to set a passing score on a teacher competency test should include at least one practicing teacher, since teachers will be affected by the outcome. Similarly, parents might be included, as might school administrators, since they will be affected by the outcome. Also, state education agency officials or school board members who have ongoing responsibility for reviewing and enforcing the standards set should participate. Ultimately, the standard setting panel should be both competent and heterogenous.

Size is another consideration when selecting a panel of judges. If the panel consists of only two judges, then the reliability of the passing score may be low, even if both judges are competent raters. That is, another panel of two may produce a different passing score. If the panel is adequately large, then the reliability of the composite score should be acceptably high. Livingston and Zeiky (1982) recommend a minimum of five judges. The number of judges can be larger, but past a certain point the addition of a new judge to the panel is unlikely to affect the outcome significantly. Similarly, as the size of the panel grows the addition of another judge is unlikely to affect the reliability of the composite passing score set by the panel.

6.1.2 Defining Minimal Competency

This is the least well developed procedure discussed in the literature on occupational testing. After the judges have been brought together, it is important to explain to them the purpose of their meeting, the procedures they are going to follow, and

the effect the outcome will have on examinees. During this process, the minimally competent examinee is defined. Typically, in the literature, the judges are asked to imagine a group of borderline examinees. These examinees are exactly on the border between incompetent and competent. This borderline group is identified as the minimally competent group.

6.1.3 Standard-Setting Methods

The approach that one uses to set a passing score will depend in part on the type of test that is being considered. On a performance-based test the examinee's performance is scored according to certain criteria by trained raters. The judges then examine the samples of performance of a number of examinees with different levels of ability and decide individually whether each examinee's performance is satisfactory or not. These ratings of adequacy are then compared with the scores assigned by the trained raters. The passing score becomes the point at which there is a high degree of agreement among the judges that a given performance level is adequate. At a very minimum, the passing score would be set at the score level at which 50% of the judges felt that the performance sample was adequate. However, at this minimum level, the passing score is also considered inadequate by 50% of the judges. Thus, a higher percentage of agreement is typically used to set a passing score. This may be 70% or 80%, for example, depending on the level of agreement that the judges desire to attain. The important point to remember here is that on a performance test, the judges rate samples of examinee performance as being either minimally adequate or not. The judges do not rate the test itself.

On a multiple-choice test, at first glance an approach to standard setting would be to simply declare that an examinee must answer a certain percentage of test items correctly (for example, 80%) in order to be considered competent. Being able to answer four out of five questions correctly does suggest some degree of competence. However, the process does not take into account the

difficulty of the items or the test. On a difficult test, a score of 80% correct could represent a high degree of competency, while on a very easy test the same percentage of correct answers might represent less than adequate performance. Because the percentage of items that test takers answer correctly depends in part on the difficulty of the test, the passing score on a test should be established in a way that considers the difficulty of the test and the items that comprise it.

There are a number of different ways of setting a passing score that have been described in the literature on professional and occupational testing. The three most frequently used methods are those of Nedelsky (1954), Angoff (1971), and Ebel (1972). The Nedelsky method requires that a panel of judges determine which distractors (wrong answers) on a multiple-choice test could be eliminated by a minimally competent examinee. The process requires each judge to make as many decisions as there are distractors on the test. Therefore, a 100 item, 4 option multiple-choice test would require each judge to make 300 decisions. Closer consideration of the Nedelsky method caused CAL to question the validity of the method for standard setting purposes. First, CAL staff felt that judges are not generally capable of determining the effectiveness of distractors. They would not have the appropriate background to make such judgements and it did not seem feasible to teach the judges to do this. Indeed test developers are not confident of their own ability to do so, and for that reason, prefer to pretest items. Furthermore, there seems to be a basic psychometric flaw in the Nedelsky method. The method assumes that all incompetent examinees identify certain distractors as being incorrect. In reality, all distractors on a well designed test attract some examinees from a variety of ability levels. Since the assumption does not hold up in practice, it was decided that the Nedelsky method would not be appropriate for setting a passing score on the GETEP.

The Ebel method involves the judges in even more complex

procedures than the Nedelsky method. The Ebel method involves two preliminary stages. First, the judges must classify each item into four categories of relevance: essential, important, acceptable, and questionable. Next the judges must classify the items into three difficulty categories: easy, medium, and hard. This two-stage process creates a matrix containing a total of 12 categories. The items belonging to each category are identified and the judges then consider the items in each category as a whole and attempt to predict the percentage of items in that category that would be answered correctly by the minimally competent examinee. Then, the percentages for each of the twelve categories are averaged to determine the passing score for the test. Besides the additional time it requires, the principal problem with Ebel's method arises in the determination of the percentage that will be answered correctly in each category. Here, the items are not considered individually, but as a group. If the category involves more than a few items, it is questionable whether such an abstract judgement can be made. If asked, a judge would probably have considerable difficulty describing what he or she is rating; that is, what the test taker is supposed to know or be able to do.

The Angoff method is similar to Nedelsky's method, but it is easier for judges to use, less time consuming to carry out, and can be used with items that do not employ a multiple-choice format. Under the Angoff method, the judges consider the whole question (the item stem and its distractors) and indicate the probability that a minimally competent examinee would be able to answer the item correctly. These probabilities are then summed across all items and all judges and the average probability becomes the passing score for the test. Thus, if the average probability were 66%, then 66% would become the passing score for the test.

6.2 Procedures Used in This Study

6.2.1 Selection of Judges

In this study, judges were selected by the Guam DOE, following recommendations made by CAL. In a letter to the Guam DOE (see Appendix F) CAL indicated that the judges should possess the following traits: they should all be competent English speakers; they should all be familiar with the teaching situation; they should all have seen teachers whose language proficiency was adequate and teachers whose language proficiency was inadequate. The DOE was also advised to select judges that represented the different points of view that might be prevalent on the island and the different groups that would be affected by the standard set. Although a dozen people were invited to participate, ultimately only eight showed up for the rating sessions. These judges, their positions and the groups they represented, are indicated below.

Julie Sisson	currently secondary language arts teacher, and assistant principal
Ernestina Cruz	former secondary social students teacher, Federal Programs Administrator for the Guam DOE
Carmen Rodriguez	currently elementary and special education teacher
Beth McClure	high school librarian and former elementary school teacher, school board member and Guam AFT representative
Nerisa Shaffer	program evaluator in the Division of Special Education for the Guam DOE; native speaker of Hiligaynon
Evelyn Salas	former social studies teacher and guidance counselor; Certification Officer for the Guam DOE

Patrick Artero	former PE teacher, National Recruiter for Guam DOE and representative of the Governor's office
Allene Yamashita	former early childhood teacher, Associate Superintendent for Curriculum and Instruction for Guam DOE

While these judges brought diverse perspectives to the group, all were involved with education in Guam and had excellent English language proficiency, which they could apply in judging the difficulty of items. They had experience teaching a variety of subjects and levels, and many were also parents with children enrolled in the schools.

The judges completed a background questionnaire. In response to a question about native language, four indicated that their native language was English, two indicated that both English and Chamorro were their native languages, one indicated that Chamorro was her native language, and one indicated that Hiligaynon (a Filipino language) was her native language. This same judge indicated that she also used Tagalog at home. All of the judges who spoke a language other than English indicated that they also use that language and English at home.

Between them, the eight judges had a total of 44 years classroom teaching experience at the K-12 level. The amount of teaching experience at this level ranged from one to 11 years. Five of the eight judges had experience as a teacher trainer. Together, those five had a total of 11 years experience as a teacher trainer. Five also had experience observing teachers in the classroom in a variety of capacities. These included serving as a consultant, a resource teacher, a program evaluator, a school administrator, a DOE administrator, or a guidance counselor. Seven of the eight judges were female and the average age of the judges was 38 years. Given their qualifications and the variety of their backgrounds, these judges can be considered an appropriate group to participate in the standard setting process.

6.2.2 Approach to Standard Setting

As indicated earlier, the speaking and writing sections of the GETEP are criterion-referenced, performance-based tests of productive communication skills. In previous sections we have described the productive speaking and writing sections of the GETEP along with the procedures followed to train raters. In addition, it was noted that the educators trained in Guam to score these tests recommended a passing score of 2+ on the speaking test which uses the scale used by the Interagency Language Roundtable and the American Council on the Teaching of Foreign Languages. It was also noted that the teachers trained to score the GETEP Writing Sample recommended a composite score of 3.5 as being acceptable performance. In order to set a passing score on these two sections of the test the following procedures were followed.

6.2.2.1 Writing Sample

For the GETEP Writing Sample (GWS), Stansfield explained to the judges the procedures followed in pretesting two GWS essay questions at the UOG. He noted that the UOG student essays were used to train the essay raters in holistic scoring. He also described the procedures followed by the essay raters in arriving at the GWS scoring guide, and they were given a copy of the GWS scoring guide. The judges were then shown a number of benchmark essays based on the results of the uninterrupted scoring of 20 essays that occurred at the end of the training of GWS raters. These benchmarks were chosen on the basis of a high degree of agreement among the five people trained to score the essays. The judges were shown one benchmark at GWS level 1, one at level 2, two at level 3, two at level 3.5, two at level 4, and one at level 4.5. The judges were told that the essay raters had recommended 3.5 as the passing score. After a good deal of discussion of the problems exhibited by the writing in the benchmark essays, the judges voted on which level of writing skill they believed should be established as acceptable

per ɔrmance on the GWS. This process took up the first day of the meeting of the panel of judges.

6.2.2.2 Oral Interview

For the speaking test, Stansfield showed the panel of judges who participated in the standard setting the ACTFL and ILR skill level descriptions. He discussed the nature of examinee ability at each point on the scale. The panel of judges was told that the speech teachers who had been trained had recommended that a score of 2+ be established as passing. However, the panel of judges was told that it was free to raise or lower that score. Stansfield then played to the panel portions of tape recorded interviews given on Guam by the speech teachers, and one interview recorded in the U.S. These examples of examinee speech represented several points on the scale to the panel. The panel heard one tape at level 1, one at 1+, one at level 2, three at level 2+, two at level 3, and one at level 3+. This process and the discussions that followed took the most of the second day of the meeting of the standard setting panel. Toward the end of the second day, following extensive discussions, each member of the panel voted on what he or she considered to be minimally adequate performance.

6.2.2.3 Multiple-Choice Sections

Because of the problems noted in the discussion of the Nedelsky and Ebel methods, the Angoff method (sometimes referred to as Angoff's method 1) was chosen as the standard setting procedure to be used in this study. This procedure was explained to the judges carefully toward the end of the second day of their meeting. Judges were told that their task was to decide on the probability that a minimally competent teacher would be able to answer each item correctly. Because this probability might be vague to some judges, the judges were told that another way of looking at their task was to estimate the percentage of minimally competent teachers who would be able to answer the item

correctly. The judges were given rating sheets (see Appendix G) for each form of the test. These rating sheets contained numbers progressing in integers of 5, ranging from 25 to 100. These numbers represent the percentage or probability of an item being answered correctly by minimally competent teachers. The judges simply circled the probability of their choice.

The judges were also given a set of pages containing the item difficulty values for the groups that participated in the pretesting. It was explained that these item difficulties may be of some assistance in making the appropriate judgement. Judges were told that the pretest group that took forms 1 and 3 was composed mostly of ROTC students and about 10% teachers. For forms 2 and 4, the percentages were about 80% teachers, most of whom had been enrolled in the DOE's remedial English language program for teachers, and 20% were ROTC students. Neither of these groups, it was pointed out, represented the minimally competent group that the judges were to consider in rating the items. It was noted that a number of the pretest items had been modified following the item analysis conducted at CAL. For some items, one or more distractors were changed, or the key was modified. These items and the changes they underwent were indicated on the pages distributed to judges and judges were told that for such items the difficulty may now be quite different from what it was on the pretest, even if the same group were to take the item again.

Subsequently, Stansfield handed out form 1 of the Error Detection test. The panel of judges read the directions and discussed the sample items. They then progressed to a discussion of the items on the first set of 10 items on the test. The difficulty of these items was discussed and the judges were asked if they had observed teachers who made errors like those found on the test in their own writing. The judges agreed that they had seen such teachers in the classroom.

The judges were then asked to use the rating sheets to rate the difficulty of each these first 10 items for a minimally

competent teacher, referring, if they so desired, to the item difficulty statistics that were obtained from the analysis of performance on the pretests. Stansfield then asked each judge to state his or her rating for item 1 aloud. The range of difficulties was then discussed and judges were asked to justify their rating of the item. This produced a high degree of consensus about the difficulty of the item and whether or not a minimally competent teacher should be able to identify this particular error in student writing. The process was then repeated for each of the first 10 items. Through this process, each judge was able to internalize the procedure of judging the difficulty of the item, and then relating that difficulty to what the minimally competent teacher can do. He or she was also able to compare his or her ratings with those of the other judges. This allowed each judge to become aware of other relevant factors that others were considering when making the judgement, and ultimately, to take these into account also. Thus, judges made educated judgements about difficulty or probability for each item.

The judges were then told to rate the remaining items on the Error Detection part of form 1. When all had finished, the judges were given the Error Detection part of forms 2, 3, and 4, and asked the group to rate these items also. Approximately every 45 minutes he would stop the group and ask each judge to indicate how he or she had rated a specific item. The judges ratings were then discussed as described above. This ensured that the judges continued to receive feedback about the appropriateness of their ratings in comparison with the way others viewed the ability of a minimally competent teacher to carry out the task represented by the item.

This procedure was followed also for the listening comprehension and reading comprehension parts of the GETEP. In introducing these parts of the test, Stansfield also discussed the types of items they contained, what each type is purported to measure, and the relative difficulty of these different types of

items. After discussing these parts and comparing and justifying their ratings, the judges rated each of the four forms, with occasional intermissions to discuss a single item as a group.

6.3 Findings

6.3.1 Description of the Minimally Competent Teacher

Throughout the standard setting process, there was a good deal of discussion of the minimally competent teacher. This discussion frequently focused on the spoken language skills of teachers who are not minimally competent, as well as the spoken language skills of nonnative English speaking teachers who are minimally competent. In addition, some judges indicated in writing on the judges background questionnaire the nature of these two groups of nonnative English speaking teachers. The judges oral and written comments are summarized below.

It was noted that the group that was not minimally competent required repetition in interacting with students and other teachers. The frequency of the repetition required contributed to a high degree of frustration on the part of students and colleagues. Another problem ascribed to the linguistically incompetent teacher is a heavy accent that includes the frequent mispronunciation of words. The difficulty of understanding such persons sometimes causes frustration to the listener. Some listeners tire and quit listening. In addition to a heavy accent, such teachers may exhibit problems in sentence structure, i.e. dropping of word endings, incorrect tense, and lack of subject verb agreement, and inadequate vocabulary or incorrect use of words in context. These teachers are unable to speak with fluency or to organize their thoughts to express more complicated concepts.

At the kindergarten and elementary school level, poor pronunciation on the part of the teacher results in problems in teaching students to read. As part of the reading process, the teacher must teach the correspondence between sounds and letters. If the teacher mispronounces a sound, when teaching a letter that

corresponds to it, the student may learn the wrong correspondence. For example, if the teacher pronounces the word "pin" but writes the word "fin," the student may develop an incorrect and counterproductive reading skill.

The listening comprehension skills of such teachers was criticized. It was noted that they often misunderstand information given to them or do not follow directions given by administrators.

Classroom management problems seem to be one outcome of a lack of adequate English skills. The students, it was pointed out, loose respect for the teacher as an authority on which they can rely. This can lead to less than optimal learning, which in turn can lead to discipline problems. The situation leads to parental complaints. When students do not understand the teacher, they may ask for repetition or clarification. Some teachers become annoyed or defensive under such circumstances. They may perceive such students as obnoxious. When discipline problems occur, the teacher may become a "dictator" in order to maintain control of the classroom. Or, the teacher may surrender control and let the students run the classroom.

It was also noted that inadequate language skills affect the teaching behavior of the teacher. Some teachers with this problem employ a minimum of oral communication during instruction, relying instead on writing on the blackboard or having students do an unusual amount of work with "ditto" handouts.

The reading and writing skills of such teachers were also mentioned. It was pointed out that such teachers are not able to correct student writing, especially at the secondary school level, and they are not able to write well themselves. They provide incorrect models of written language to the students. When writing notes to parents they commit basic errors in syntax or word choice. These notes can generate concern among parents about the competency of the teacher. They can also generate complaints from parents to school administrators.

It was felt that the minimally competent teacher, in contrast, may exhibit a few of these characteristics, but not a majority of them. The minimally competent teacher miscommunicates infrequently. Such miscommunication does not detract from the learning process. While the minimally competent teacher may exhibit a foreign accent and sometimes make errors in sentence construction, he or she usually has a good vocabulary. These teachers usually speak English fluently, depending on the topic being discussed, and are not hesitant to initiate class discussion. Less than exemplary language skills are aided by good organizational skills. They are able to use humor in the classroom and show enthusiasm for their students. The minimally competent teacher recognizes his or her speech problems and strives to improve.

6.3.2 Writing Sample

The judges voted to set 3.5 as the minimum acceptable score on the GETEP Writing Sample. As indicated earlier, there was unanimous agreement on this score level.

6.3.3 Oral Interview

Seven of the panel members voted that level 3 be the minimum standard on the oral interview. One panel member abstained, although in previous discussions she argued in favor of level 2+ as the minimum standard. Even if one considers this abstention a negative vote on a 3 and a vote in favor of a 2+, a teacher with a rating of 2+ would be viewed as having adequate oral language ability by only 12% (1/8) of the judges. On the other hand, it would appear that a teacher with rating of 3 would be viewed as having adequate oral language ability by 100% of the judges. (A judge who believes that a 2+ is adequate would consider a 3 to be more than adequate.) Given this high degree of consensus, level 3 speaking skill should become the passing score on the oral proficiency interview portion of the GETEP.

6.3.4 Multiple-Choice Sections of the GETEP

For the multiple-choice sections of the GETEP, the judges ratings showed a high degree of consistency. These results are depicted in Table 6.1.

In Table 6.1, the raw score (number of right answers) is indicated by the first number to appear after the name of the section. The number in parentheses to the right of the raw score is the percent of correct answers required on that section of the test. This score is the composite or average score obtained by analyzing the ratings of all eight raters. Averages can be expressed by either the mean or the median. Either figure can be chosen, although the median is less subject to the influence of an extremely severe or an extremely generous rater. However, in this study, there was no great difference between the mean and the median scores, especially since the normal procedures for rounding to a whole number were followed, which means that scores ending with the decimal .50 or greater were rounded up to the next whole number. Without rounding, the mean and median scores were generally closer to each other than Table 6.1 indicates.

The data on total scores indicates that the four forms were perceived as being about equal in difficulty for the minimally competent teacher. The total score (in whole number raw scores) varied by form and by measure of central tendency (mean or median) between about 77% and 81% correct.

Table 6.1
Mean and Median Judges Ratings by Raw Score and Percent of Correct Answers for four GETEP Forms by MC Section

Form 1	Mean		Median	
Listening	33	(.825)	33	(.825)
Reading	36	(.72)	37	(.74)
Error Detection	23	(.767)	23	(.767)
Total	92	(.767)	93	(.775)
Form 2				
Listening	32	(.80)	33	(.825)
Reading	39	(.78)	39	(.78)
Error Detection	25	(.833)	25	(.833)
Total	96	(.80)	97	(.808)
Form 3				
Listening	33	(.825)	34	(.85)
Reading	38	(.76)	39	(.78)
Error Detection	24	(.80)	24	(.80)
Total	95	(.792)	97	(.808)
Form 4				
Listening	34	(.85)	34	(.85)
Reading	38	(.76)	39	(.78)
Error Detection	25	(.833)	25	(.833)
Total	97	(.808)	98	(.817)

The GETEP assess four communicative skills: listening, speaking, reading, and writing. Everyone interviewed in Guam felt that it would be most appropriate to report to examinees whether they had passed each section of the test. In this way, if an examinee passes one section, he or she will not have to take that section again. The writing teachers and the judges felt it would be most appropriate to combine the writing part

(Error Detection) of the MC test with the GETEP Writing Sample (GWS), and weigh each portion equally in order to obtain a total score for writing. Since the Error Detection portion contains 30 items and the GWS allows for a maximum rating of 5, in order to weigh the two equally, it is necessary to multiply the GWS score by 6. Thus an examinee who obtains the maximum score of 5 on the GWS will receive a weighted score of 30. The GWS passing score, which was recommended unanimously to be 3.5, would then be a converted score of 18.

Using the mean number of right answers from Table 7.1 as the basis for determining the passing score on the writing (Error Detection) part of the four MC forms, and combining it with the weighted score of 18 on the GWS, we see that the total number of points required to pass the writing skills portion of the test is as follows.

Form 1	23 + 18 = 41
Form 2	25 + 18 = 43
Form 3	24 + 18 = 42
Form 4	25 + 18 = 43

It is not possible to give a precise estimate of what the reliability of this composite score may be since the two subtests (Error Detection and the GWS) were given to different samples of examinees. However, it should be noted that reliability of a composite score is higher than the reliability of the individual components and higher than the average of the components. (For example, the reliability of a composite score based on two tests having .80 reliability is .89.) Therefore, given the KR 20 estimates for the Error Detection section ranging from .52 to .90 and the inter-rater reliability for the GWS at .80, it can be assumed that the reliability of the composite on the various forms of the writing test would range from about .75 to .92.

Thus for the four forms (using the mean rating) we are left with the following passing scores by English communicative skill.

	Speaking	Listening	Reading	Writing
Form 1	Level 3	33	36	41
Form 2	Level 3	32	39	43
Form 3	Level 3	33	38	42
Form 4	Level 3	34	38	43

In order to pass the test, the examinee would have to pass all portions of the test. However, if the examinee failed only a single portion, he or she would only have to take that portion again.

Two studies were carried out in order to determine the reliability of the judges ratings. In the first, all ratings were used to determine the reliability of the judges ratings for each of the 480 items. The average inter-rater reliability across forms was .46 (based on the 28 possible pairing of judges) and the average reliability of the mean scores for each item (based on eight judges) was .87. A generalizability study was also performed on the data, which produced an average generalizability coefficient across the four forms of .88. Both of these statistics indicate a high reliability of the mean scores for each item and are very good, given the hypothetical nature of the task. Indeed, they are much better than what is typically found in the literature. For example, Brennan and Lockwood (1980) of the American College Testing Program applied the Angoff procedure to five judges who rated a 126 item four-option test for licensing purposes in the health sciences. The average inter-rater reliability was only .187. The .87 reliability of the mean score for each item obtained here is also very good, given the diversity of background and experience among the judges. This means that if another group of eight judges were to go through the same procedure, there would be about a .87 to .88 correlation between their mean item ratings and the mean item ratings for this group.

The second study focused on the reliability of the composite scores (passing scores) on the multiple-choice tests. For this

study, each rater's mean rating for each of the twelve tests (4 forms times 3 sections equals 12 tests) was calculated. This mean rating represents the passing score that was indicated by each rater's ratings. The degree of agreement across raters and forms between these passing scores was then assessed. The average inter-rater reliability was .54 (based on the 28 possible pairings of judges) and the reliability of the composite scores based on eight judges was .89. Again, a generalizability study was also performed on the data, also giving a coefficient of .89. These are very good, given the hypothetical nature of the task and the diversity of background and experience among the judges. This means that if another group of eight judges were to go through the same procedure, there would be a .89 correlation between their composite ratings (passing scores) and the composite ratings for this group.

6.4 Discussion

The impressive inter-rater reliabilities obtained in this study contrast markedly with those reported in the literature. This is probably due to the many provisions taken to enhance the reliability of ratings.

The selection of judges was an important factor. Although judges were selected to represent a variety of backgrounds and orientations, all judges were known to be highly competent speakers of English who were generally capable of judging the difficulty of items. In this sense, the selection of judges was limited to subject matter expert. Judges also had experience teaching and observing both competent and incompetent teachers. This allowed them to relate the item to the hypothetically minimally competent teacher and the to teaching situation in a psychometrically adequate way.

Another important factor was the training the judges received. The project director spent from two to three hours introducing the judges to each new section of the GETEP. Items, and the tasks involved in answering them, were analyzed and

discussed. This enhanced the judges sophistication in making ratings.

"Drift" (the development of deviant standards over time) was overcome by interrupting the judges every hour to rate one of more items as a group and discuss the ratings. Judges were free to adjust or change their original rating after hearing the group discussion. The project director remained with the judges throughout the entire rating process (four full days) in order to answer any question that might arise.

Another important step was the provision of some baseline data to judges on the difficulty of items for another group of examinees, albeit a different group. In this case, there was considerable discussion as to how the pretest group may be similar or different from a group of minimally competent teachers.

Having the judges circle a probability represented on paper seemed to offer a number of advantages over having the judges write the probabilities on their rating sheet. In the first place, the probabilities listed facilitated the choice. The judges had to choose among the same 15 easily identifiable probabilities. (Counting by fives, there are 15 numbers between 25 and 100.) Had they been asked to write their probability ratings, they would have been free to choose among all 75 numbers between 25 and 100. This would have made the selection of probabilities more difficult and more time consuming. The use of 15 probability options instead of 75 allowed for adequate discrimination of item difficulty, without introducing more options into the decision than could be reliably utilized by the raters. Although Livingston and Zeiky (1982) recommend having judges write their probabilities on paper, since this permits the use of all numbers as probabilities, the use of a reasonable but limited number of options seemed to work very well in this study.

It should be noted that the judges also were very positive about the procedures. At the end of the process, after all tests had been rated, each judge completed an evaluation of the

standard setting process (see Judges' Questionnaire, Appendix H). The data gathered indicated that seven of the eight judges felt that the passing scores set for speaking proficiency and for the writing sample were "about right." All indicated that they felt that the standard setting procedures used were appropriate and that the judges were highly qualified professionals who took the process seriously. Seven of the eight raters felt that the provision of item difficulty data from the field test administration was useful.

This data suggests that the extensive procedures employed here to ensure reliable and valid ratings had a positive affect on the outcome.

7. Recommendations for the Operational Testing Program

This chapter proposes next steps and procedures for the implementation of the GETEP on Guam and in other locations. It includes many suggestions which we believe will have a beneficial effect on the program.

7.1 Administration and Scoring of the GETEP on Guam

Listed below are a number of recommendations and next steps concerning the implementation of the GETEP on Guam.

7.1.1 Appointment of a GETEP Program Director

As soon as possible, a DOE employee should be made permanent director of the GETEP program. A teacher certification test program requires a great deal of attention. This attention would best be provided by a person who would be responsible for its smooth operation. Initially, and during the first two years, considerable effort should go into the implementation of this test program. An effort of this magnitude will require a program director to supervise the operation.

7.1.2 DOE to Issue Pass/Fail Scores

The Guam DOE should report scores to examinees on the basis of a pass or fail on each of the four sections of the test. Thus, if an examinee fails in only one skill, he or she will have to retake only that one section. Exact scores on any section should not be reported.

7.1.3 Examinee Handbook

An examinee handbook should be created by the Guam DOE. This is a relatively simple matter. It would contain the instructions and the sample items for each section of the test, along with a description of the OPI and the scale. The ACTFL version of the skill level descriptions, which stops at the Superior level (level 3), could be used. It would also contain a copy of the instructions for the GETEP writing sample, a sample

prompt, a copy of the scoring guide, and perhaps an example of a successful essay at the 3.5 level. In this way, examinees will be provided with all relevant information on the test. This information will permit the examinees to familiarize themselves with the test format prior to taking the test. It will also allow them to better understand their GETEP scores.

In addition to a description of the test, the examinee handbook should describe briefly operational procedures and other DOE policies relevant to the test. This would include the purpose of the GETEP, when it is offered, the policy on re-examination, identification documents required, number of pencils to be brought to the test center, etc.

7.1.4 Manual for Administering the GETEP

A manual for administering the GETEP should be written. This would cover basic information for the test center supervisor, such as the acceptability of identification documents, selecting a suitable room, checking the sound equipment prior to the test, counting the test booklets and answer sheets before and after the test, etc. It should also include instructions for administering the GETEP Writing Sample.

7.1.5 GETEP Answer Sheet

CAL has drafted a GETEP answer sheet that can be used on National Computer Systems (NCS) optical scanners. This is presented in Appendix I. Both CAL and the DOE have an NCS scanner. The answer sheet has been sent to Dr. Jeff Shaffer of the Guam DOE. The answer sheet can be sent by the DOE to NCS, which will print a supply of machine-readable GETEP answer sheets identified as such. A scanning routine and database for the answer sheet will have to be developed by the Guam DOE.

The answer sheet records the following information:

-Name of examinee
-Social Security number
-Birth date
-Sex
-Test center (up to 99 centers may be encoded)
-Test form
-Native Language (English, Chamorro, Filipino, Chinese, Korean, or Other)
-Grade level taught (Elementary, Middle or Junior High, High School)
-Location of university from which examinee graduated (Guam, Hawaii, U.S. Mainland, Philippines, Other)
-Location of exam (Guam, Hawaii, U.S. Mainland, Other)

The use of this answer sheet will allow the easy determination of the average test scores from each of the test centers and from each of the general locations.

Once the GETEP is administered on island, it can be scanned at the DOE and scores can be reported. A score reporting procedure will have to be established.

7.1.6 Printing of the GETEP

Camera ready copy of the GETEP has been provided to the Guam DOE. This must now be used to print an adequate supply of the test for examinees. It would be best if the test booklet could be saddle-stitched (copied on to 11 by 17 inch pages, folded, and stapled in the middle). This would create a test booklet and make it more difficult for examinees to tear out pages of the test. The GETEP Writing Sample can be copied on to a four sided test booklet, consisting of a single sheet of 11 by 17 inch paper. Side one would contain the directions for the test. It would also contain examinee identification information. Side two would be the prompt and space for making notes. Sides three and four would be the pages on which the examinee would write the essay.

The tape for the listening comprehension section has been recorded in a professional recording studio. The tape for each form may be copied on Guam. Each cassette should be labeled and the form number should be prominently displayed.

7.1.7 Test Security

Test security is the most important characteristic of the GETEP operational program. Without it, all other efforts are wasted. Every measure possible should be taken to ensure the security of the test. Each test booklet should be numbered and counted before and after each administration of the test. Each test tape should also be numbered and checked similarly. The Manual for Administering the GETEP should describe in depth the provisions that should be taken to ensure security.

7.2 Administration and Scoring of the GETEP on the Mainland

CAL recommends the establishment of test centers in major cities on the mainland. The test centers should be located in regionally central cities such as Boston, Chicago, Washington, DC, Orlando, Atlanta, New Orleans, Austin, Albuquerque, Denver, Salt Lake City, San Francisco, Los Angeles, Seattle, etc. About 25 such test centers should be established, so that prospective teachers would not have to travel more than 300 miles to be tested. These test centers would be used to test prospective applicants on the U.S. mainland.

Currently, such applicants are typically tested by recruiting teams. However, this process poses a burden on the recruiting team, who also have to test people that have not had previous contact with the Guam DOE. In addition, CAL believes that the DOE should rely less on recruiting teams in the future, since monies may not always be available to cover the cost of travel and salaries for such trips. Thus, additional or alternate structures for testing applicants should be put in place. The creation of regional test centers appears to be a reasonable alternate structure for testing applicants.

CAL would be pleased to assist the Guam DOE in creating regional test centers. At each of these test centers, a trained oral proficiency interviewer would be on call. CAL has a list of 500 such interviewers. When a prospective teacher on the mainland writes the Guam DOE about a position on the island, the

DOE could contact CAL. CAL would in turn contact its consultant at each test center. Most typically, this would be a professor at a local university who is a trained oral proficiency interviewer. CAL would also contact the applicant and tell them to contact the interviewer to arrange for testing. CAL would then send the interviewer the GETEP test materials. The CAL consultant would administer the oral proficiency interview to the applicant, followed by the GETEP Writing Sample, and then the 120 item multiple-choice test. The consultant would provide CAL with the score on the OPI, and return the test materials, the GETEP Writing Sample and the answer sheet to CAL. CAL would then scan the answer sheet and use two of its trained staff to score the writing sample. The scores on the various portions of the test could then be sent to the DOE. When necessary, FAX and courier mail can be used to facilitate rapid turn around.

CAL believes that such a system can be operated at a reasonable cost. The system would ensure the availability of the GETEP throughout the mainland.

7.3 Administration and Scoring by Recruiting Teams

Currently, recruiting teams are sent to the mainland to recruit teachers to come to Guam. These teams have done an excellent job of finding teachers and filling the DOE's needs for certified personnel. Typically, the teams attend a recruiting fair that is also attended by a large number of teachers looking for jobs. Such fairs seem to be an efficient way of identifying teachers.

At the fair, there is a need to administer the English language proficiency test immediately, in order to determine whether the applicant has met all requirements for certification on Guam. This could be done by including one of the four trained oral proficiency interviewers and two of the four trained essay raters in the recruiting team that travels to the mainland. In such a case, these language testers could administer the productive skills sections of the GETEP and the multiple-choice

section also. However, this procedure takes competent staff away from the island (and away from home) for an extended period of time. It also costs the DOE considerable money in terms of salary and travel expenses.

A solution to the problems posed by sending staff from Guam is for CAL to assist with the testing at recruiting fairs. CAL could contact a local oral proficiency interviewer and have this person agree to be present at the fair in order to interview prospective teachers. If necessary due to the size of the fair, two interviewers could be provided. The interviewer would be available to test applicants during the day before the fair, during the fair, and on the day after the fair. Similarly, if two interviewers were used, in addition to interviewing, they could administer the writing sample and the multiple choice portion of the test. The multiple choice portion could be scored using a hand scoring stencil. It would require about 12 minutes per examinee to score the test in this way. The writing sample could be FAXed to CAL, where it could be scored immediately by two trained raters, and the scores phoned back to the DOE staff at the recruiting fair. Or, CAL could send one or two staff members to the fair to administer and score the tests. CAL believes that this procedure would work quite well and would facilitate the job of the other DOE staff at recruiting fairs.

7.4 Maintenance of a Database

CAL believes that the Guam DOE should maintain a database on the GETEP. This database would contain all test data for all examinees. The data would permit the DOE to determine who had taken the test previously, which form they had taken, and their score. Most of the data in the database would be entered automatically by scanning the examinee's answer sheet. Scores on the GETEP Writing Sample (GWS) and the OPI would have to be key entered, unless they were entered on the multiple-choice answer sheet by clerical personnel, and then put in the database by the scanner. Tests scored on the mainland by CAL could also be put

in a database. The data could be put on a diskette and sent to Guam periodically where is could be easily merged into the main database on the island.

7.5 Research

The Guam DOE should utilize the database to conduct research on the GETEP. Presently, the answer sheet contains data on the examinee's native language, location of the test center, the location of the IHE that the examinee graduated from, age, sex, intended teaching level (elementary or secondary), and test form. It would be useful to utilize this data to analyze the quality of the test, to gain a more complete understanding of the factors that are related to an examinees score, and to understand the impact of the test on examinees and the schools.

After a year of implementation of the GETEP, it would be useful to obtain summary information from the database on the proportion passing within each ethnic group on the island, the proportion passing who are graduates of the UOG and institutions located elsewhere, the proportion passing within each native language group, and the proportion passing by sex. This data would permit the Guam DOE to determine exactly how various groups of examinees were performing on the test. Such information could be very useful in gaining an understanding of the impact of the GETEP or any other test on prospective teachers and on the teacher population in the schools. If it were necessary to adjust the passing scores set initially, the data could be very useful in that process.

The database could also be used to determine the reliability of the test for the operational test population, or to determine if any items exhibit an alarming degree of item bias. CAL believes that research should be conducted on any test used for teacher certification purposes. Just as important, however, is the need for the research to be interpreted and reported in an unbiased and responsible way.

7.6 The Operational Speaking Test Program

The following is a set of recommendations for the operational implementation of the OPI section of the GETEP.

1. Whenever an interview is conducted on Guam or elsewhere, the interview should be tape recorded for later reference.

2. Each interview should be administered and scored by a single rater. On the basis of that score, a pass or fail will be reported to the examinee. Whenever an examinee is assigned the score immediately below the passing score (level 3), the tape should be listened to by a designated chief rater or an assistant chief rater. The chief rater and the assistant chief rater would have the authority to change the rating, if he or she does not agree with the first rating. The revised rating could be either higher or lower than the first rating.

3. Interviewers need confirmation of their ratings in order to continue to rate reliably (in the same way as others). Thus, occasionally, taped interviews should be listened to by a second interviewer, who can provide feedback on the rating technique and on the rating. This feedback process helps both interviewers to maintain the same standards.

4. If possible, additional exposure to the OPI should be provided the interviewers as a program of continuing professional development. Perhaps it would be possible to send one of more of them to the mainland for advanced training (called recalibration training) by ACTFL. Such two-day training sessions are usually held prior to the ACTFL annual convention in November.

5. For interviews conducted off-island, a preliminary decision about the suitability of the applicant should be made on the basis of a single rating. If the applicant continues to remain interested in a position and scores immediately below the passing score, but passes all other portions of the test and otherwise seems to be a good candidate, then the tape can be verified later by a second rater on Guam.

6. The anonymity of the second rater should be protected, to the degree possible.

7.7 The GETEP Writing Sample

As indicated in Chapter 4, the GETEP Writing Sample (GWS) should be scored by two raters. Under these circumstances, it is quite reliable. However, in order to increase efficiency, it may be possible to do only a single rating of papers whose first rating was at score levels 1, 2, or 5. A paper assigned a 5 by the first rater would have to be assigned a 1 by the second in order to fall below the passing score level of 3.5. The probability of this happening are practically nil. Similarly, a paper rated a 2 by the first rater would have to be rated a 5 by the second rater in order to pass. The probability of this happening is also extremely small (about 1%). Thus, it can be assumed that papers whose first rating is either a 1, 2, or 5 are safely above or below the passing score level. Since most papers will be at the 3 and 4 levels, this policy would make it unnecessary to rescore only about 1/3 of all papers. Although this does represent some cost savings, the savings is not great since a single rating takes only about 3 minutes.

More important than the above, however, is the need to develop new essay prompts within a few years. Even though prompts should remain secure after being administered, some examinees will write down the prompt and give it to friends after the test administration. Thus, although prompts may be reused after several years, the supply of prompts must continue to grow. Otherwise, examinees will simply practice writing essays on the dozen prompts used at different administrations and the test will not provide a valid sample of a prospective teacher's writing ability. In writing new prompts, the essay raters should refer to the information in sections 4.1 and 4.2 of this report. Reference to the discussion of the characteristics of an essay prompt in the volume by Hamp-Lyons (1989) will also be helpful.

It is also important to create a supply of new benchmarks, using the papers on which the two raters agreed, after each administration. These benchmarks will be useful in training new raters for the operational program. At present only four raters

are available. Once they acquire experience in the operational program, they will be able to train additional raters.

REFERENCES

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement. Washington, DC.

Brennan, R.L., & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Buck, K. (Editor). (1989). The ACTFL Oral Proficiency Interview tester training manual. New York: American Council on the Teaching of Foreign Languages.

Clark, J.L.D. (1978). Direct testing of speaking proficiency: Theory and application. Princeton, NJ: Educational Testing Service.

Ebel, R.L. (1972). Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice Hall.

Educational Testing Service. (1986). TWE scoring guide. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1989). The PPST essay guide. Princeton, NJ: Educational Testing Service.

Hale, G.A., Stansfield, C.W., Rock, D.A., Hicks, M.M., Butler, F.A., & Oller, J.W. (1988). Multiple-choice cloze items and the Test of English as a Foreign Language (TOEFL Research Report No. 26). Princeton, NJ: Educational Testing Service.

Hamp-Lyons, E. (1989). Preparing for the Test of Written English. New York: Newbury House.

Liskin-Gasparro, J. (1987). Teaching and testing for oral proficiency: A familiarization kit. Boston: Heinle and Heinle.

Livingston, S.A. & Zeiky, M.J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Stansfield, C.W. & Kenyon, D.M. (1989). Development of speaking proficiency tests for the less commonly taught languages. Final report to the U.S. Department of Education. Washington, DC: Center for Applied Linguistics. Forthcoming in ERIC.

APPENDIX A

COMPLETE NEEDS ASSESSMENT REPORT

REPORT OF A NEEDS ASSESSMENT TRIP TO GUAM TO ACQUIRE INFORMATION FOR DEVELOPING AN ENGLISH PROFICIENCY TEST FOR THE GUAM DEPARTMENT OF EDUCATION

JoAnn Crandall and John Karl
Center for Applied Linguistics
Washington, DC

During the week of November 14-18, 1988, we visited the island of Guam as part of our effort to develop an English language proficiency assessment instrument which is both specific to and relevant for Guam educators. During the week, we met with Guam Department of Education officials, University of Guam faculty, principals and teachers from a sample of elementary and secondary schools, and Department of Defense administrative and educational personnel. We also observed classes at four elementary, one middle and one high school. These schools were selected because they provided a representative sample of schools, teachers, and students on the island. The teachers who were interviewed and observed at each school represented the range of teachers in Guam in terms of ethnolinguistic background, gender, and experience; thus, they provide a good picture of the range of variation in the English language used in schools on the island.

The purpose of the interviews and classroom observations was to develop a clear understanding of the ways in which teachers routinely use English in their instruction, in their communications with parents, and in their interactions with other educators, and to collect, wherever possible, samples of reading materials and writing assignments which could be used to develop relevant and appropriate test items. During the visit, as well, we discussed the teacher education program at the University and plans for reformulation; the types of tests which have been used with both students and teachers in Guam, focusing especially on the use of the current BESTE test for assessing teachers' English language skills; and the kinds of professional development programs which might be provided for those who fail the upcoming test, or more broadly, which might be offered to all educators seeking to expand their teaching skills.

Throughout the visit, we were afforded a most gracious reception. We were fortunate in the cooperation provided by the Department of Education, especially in the preparation which preceded each school visit. Because of the efforts of Anita Sukola, Acting Director, the Associate Superintendents of Elementary and Secondary Education, and the Certification Officer, we were greeted warmly and were granted lengthy group interviews during which teachers and principals provided us with examples of the ways in which teachers use English with students, parents, and other teachers and education personnel and offered suggestions on testing procedures. We were also welcomed into classes, where we could observe firsthand how teachers use English in class. We were given a great deal of support by John Shaver, Consultant to the Department of Education and by various military personnel also involved in education. Evelyn Salas,

the Certification Officer, accompanied us to each of the schools and introduced us to the principals. Her planning and the prior preparation by Janette Yamashita, Associate Superintendent of Elementary Education and Beth Montague, Associate Superintendent of Secondary Education, enabled us to gather a great deal of information in a very short time. Because of this degree of cooperation and support, we were able to accomplish our objectives, even though we only had one week to do so.

The information gleaned from the interviews and observations proved to be remarkably consistent. That is, there was great consensus on the ways in which teachers use oral and written English as they go about their work. As a result, we were able to collect a good number of characteristic reading materials and writing assignments, and to identify a substantial number of examples of oral English use which can form the basis of the proficiency test. Moreover, because of the agreement among the many educators about appropriate and expected English use, we can also be assured that we can create an English proficiency test which will be viewed as both valid and appropriate as a measure of the Guam teacher's ability to teach and to function as a professional in English.

We also discussed possible formats for the proficiency test and provided the teachers with an opportunity to respond to the test which they had recently taken (the BESTE) and to offer suggestions on the types of tests which they believed would be appropriate and valid. In our proposal, we suggested that the assessment instrument might consist of an oral proficiency interview, a writing sample, a multiple-choice listening comprehension measure, and a multiple-choice or cloze reading measure, but that the final form of the test would depend upon the results of the needs assessment. There was substantial agreement on all components, with the exception of the reading test. Unfortunately, the BESTE written exam had a cloze component which created a great deal of controversy, some of it based on lack of familiarity with the cloze procedure and some on the types of reading passages which were included. If we are to use a cloze-type procedure, we will need to consider the use of multiple choice cloze passages and, of course, to be certain that the passages are drawn from typical educational reading materials. We might also want to include multiple-choice error correction exercises within the writing test since many people feel that a writing test should include both editing tasks and production.

Test administration, test scoring, and test security measures were discussed as well. There was a consensus that multiple forms of the test be developed, since test security could be a problem. Administration and scoring procedures were not finalized. There was some feeling that the oral interview should be administered live but that the interview should be taped either for scoring or for obtaining a corroborative score. The writing test might consist of a selection of prompts for two writing assignments, again to be scored by Department of Education personnel trained in holistic scoring or to be scored by another party. In general, we determined that the test should take two to three hours: 20 minutes for the oral interview, one hour for the writing test, and 30 minutes each for the listening and reading tests.

What follows is a daily report of our activities. In addition, we have provided a list of typical reading materials, writing assignments, and listening and speaking activities which could form the basis for items in the English language proficiency test.

SCHEDULE OF ACTIVITIES OF NOVEMBER 14-18, 1988

Monday, November 14

7:00 am: Breakfast Meeting with John Shaver, Department of Defense Consultant to the Guam Board of Education. We discussed our visit and outlined the activities of the week.

9:30 am: Meeting at Department of Education with
Anita Sukola, Acting Director of DOE
Janette Yamashita, Associate Superintendent for Elementary Education
Isabel Montague, Associate Superintendent for Secondary Education
Bill Pesch, Legal Counsel for DOE
Evelyn Salas, Certification Officer
John Shaver, DOE Consultant

We discussed project goals and procedures; identified some potential problem areas; agreed upon a schedule of activities for the week; identified schools to be visited, reflecting the ethnic, experiential, gender, and age mix of teachers on the island; and arranged schedules for school visits with Evelyn Salas, Jan Yamashita, and Beth Montague, who, in turn, discussed our objectives and made other prep ons with the school principals.

2:30 pm: Meetings at University of Guam wit
Robert Underwood, Chair, Edu .on Department
Joyce McCauley, Chair, English Department, and Director, Teacher Institute (for those who have failed the BESTE)
Tom Tinkham, English Professor
Dee Johnson, English Professor

Dr. Underwood reviewed the pre-service education program at the University of Guam, where many teachers on the island received their teacher education, and outlined current plans for reformulating the teacher education curriculum and program. He also provided information on the educational system in Guam and on sociolinguistic and other relevant factors of both student and teacher populations on the island. We outlined the goals of our visit and our expectations concerning the test and also discussed types of professional development activities which could b. provided to teachers having difficulty with the new test or even more broadly, to any interested teacher.

Drs. McCauley, Tinkham, and Johnson outlined the current Institute they are providing at the University for teachers who failed the BESTE. We also had an opportunity to observe some of these classes. We discussed the goals and potential parameters of the new test and other types of inservice programs which might be offered to teachers on the island.

Tuesday, November 15

7:45 am: Visit to John F. Kennedy High School. Met with Gayle Hendricks, Principal, and several teachers who agreed to be interviewed and observed. Observed classes.

We discussed our purpose in being there, solicited suggestions for general format or item types for the test, and identified a number of potential materials or situations which would be appropriate for use in an English proficiency test for Guam educators. The teachers were not opposed to testing; they were opposed to tests which were not reflective of the kinds of skills they possess or the kinds of situations in which they use English. They suggested typical communication activities they are involved in with students, with parents, and with other educators. Their specific suggestions are contained in the list which follows. After our group discussion, we visited classes.

11:50 am: Visit to Piti Middle School. Met with Edward Sablan, Principal, and several teachers who agreed to be interviewed and observed. Observed classes.

We followed the same basic procedures as were used at JFK High School, meeting first with the principal, then with the principal and teachers, and finally observing classes.

3:30 pm: Meeting at Naval Air Station with

Lt. Colonel Randy Prier, Air Force/Government of Guam Liaison Officer and ex-officio member of the Guam Board of Education
Lt. Commander John Alexander, Navy Representative to Government of Guam Affairs
Ms. Barbara Askey, Navy Education Specialist

We discussed the purpose of our visit; reviewed some of our initial findings concerning this test, the BESTE, and the Institute; and discussed potential benefits which might accrue from the new test (in terms of professional development activities). We also mentioned the high degree of cooperation we were finding among DOE, the principals, and the teachers we had talked with.

4:30 pm: Meeting at Department of Education with
Dr. Jeff Shafer, Director of Testing
Dr. Art Wheeler, Developer of the BESTE test

We discussed the development of the BESTE test, reviewed each component, and discussed the scoring procedures. We were informed that the oral portion of the BESTE had been developed by speech therapists and the written portion by English teachers. We also requested a representative sample of the audiotapes of the oral section and a full version of the test. Art Wheeler will collect these (identifying a representative range of oral scores/skills) and send them to us at a later date.

7:30 pm: Attended island-wide meeting of the PTO (Parent Teacher Organization).

Wednesday, November 16

7:30 am: Visit to Merizo Elementary School. Met with Tomas S.N. Barcinas, Principal, and some teachers. Observed classes. Met with another group of teachers.

11:00 am: Visit to Wettengel Elementary School. Met with a group of teachers. Observed classes. Met with other teachers and with Angelita P. Camacho, Principal.

Thursday, November 17

7:30 am: Visit to Andersen Elementary School. Met with Acting Principal, Rose Mary Lamela. Observed classes. Met with teachers.

10:30 am: Visit to M.U. Lujan Elementary School. Visited classes. Met with large group of teachers (Principal was out that day).

2:30 pm: Meetings at University of Guam with
Dr. Florence Riegelhaft, Visiting Professor of Linguistics/Bilingual Education
Dr. Mary Spencer, Director of Project BEAM Multifunctional Resource Center for Micronesia

We discussed our visit to the island, the teaching situation and teacher preparation program, the sociolinguistic nature of the island's population, and some of the activities of the teacher education department and resource center. We also identified a number of ways in which we can work together during the coming year.

Friday, November 18

9:00 am: Breakfast Meeting with John Shaver.

We reviewed our findings and discussed the briefing meeting which was to follow in the afternoon.

10:30 am: Meeting at Headquarters with Admiral Johnson.

We briefly reviewed our activities for the week, our findings, and our suggestions for the parameters and item types for the test. We also discussed the teacher education aspect of the testing project, suggesting a positive role for professional development (which would be open to any teacher wishing to participate) rather than a program reserved only for those who had had difficulty with the test.

12:00 pm: Luncheon Meeting with Evelyn Salas, Jan Yamashita, and John Shaver.

We discussed the results of the week's interviews and observations and the contents of the briefing which would be given to the Department of Education. We also discussed materials which we had not been able to obtain which would be helpful for test development. Evelyn Salas collected a number of these and will also follow up on future requests.

3:00 pm: Meeting at Department of Education with
Anita Sukola, Isabel Montague, Jan Yamashita, Evelyn Salas, Bill Pesch, and John Shaver.

We reviewed the week's activities and presented our findings. We agreed that the test would consist of four components, an oral interview administered live but taped for subsequent confirmation of score; a writing sample; a listening comprehension test; and a reading test. We identified sample item types for each test, drawn from the list we developed over the course of the week. We also discussed the need for multiple forms of the test and for security measures and agreed that CAL would submit a proposal augmenting the current plan for two forms of the test for an additional two. In addition, we discussed some possible professional development models and activities and suggested that CAL would be interested in working collaboratively with the University of Guam in designing and implementing the program. We agreed to submit some ideas for this professional development program in writing to the Department of Education for further consideration.

SUGGESTIONS FOR WRITING ITEMS

Some types of writing assignments which might be used include:

1. Notes to parents about student problems such as discipline problems, attendance problems, potential failures in a subject, etc. This note could identify the problem, suggest ways in which the teacher is working to help, and ways in which parents might be of help.

2. Note to parents explaining the purpose or function of an extracurricular activity, perhaps requesting parental permission for a student to participate.

3. Referrals to special education with justification for the decision.

4. Discipline referrals to an administrator such as the Assistant Principal.

5. Field trip request to the Principal.

6. Memo to a substitute or to an aide outlining a lesson plan and suggesting ways of implementing it.

7. An assignment which is written at the appropriate level of their students, with grade appropriate vocabulary and syntax.

8. A sample student essay question that teachers would correct.

9. A lesson plan for one day's instruction.

10. Directions for students on how to use the library and a discussion of the importance of developing good library habits.

11. Directions for students or parents on how to get/find something in the school (the cafeteria, assembly hall, clinic, etc.)

12. A letter to parents requesting that they make certain their child does his homework every night in which they describe or justify the importance of such assignments.

The following are topics which could serve as prompts for extended writing. Teachers agreed that any of these would be appropriate for a teacher English proficiency test and that all should be able to write on these.

1. Extensions or changes in the school year
2. Extensions or changes in the school day
3. Teacher testing
4. Staff development needs
5. Promotion and retention policies
6. Parental involvement - limitations; how to involve
7. How to work with parent volunteers or teacher aides
8. The middle school concept
9. The integrated curriculum
10. Professional development activities - how to organize these
11. A useful professional development activity which you have participated in and why it was useful; how you applied it
12. Extracurricular activities - what to include; what to omit
13. Discipline - how to attain and maintain it
14. Changes in the curriculum
15. A successful lesson/lesson plan
16. School lunch program (nutrition, efficiency, etc.)
17. A wish list of supplies, equipment, facilities, etc.
18. How to deal with a student who doesn't speak English
19. A description of your school or class for a prospective teacher or student teacher
20. Teacher evaluation criteria and methods
21. Strategies for helping a student who is having difficulty mastering a concept
22. Motivating a student
23. Motivating a lesson
24. Teacher certification or recertification requirements
25. Strategies for re-teaching something

SUGGESTIONS FOR READING MATERIALS FOR TEST ITEMS

1. Board of Education policy statements
2. Memos from the Department of Education, the Associate Superintendents, the Principals, etc. that teachers also read
3. Daily bulletins read to the students
4. Articles from The Union
5. Articles from Now
6. Curriculum manuals from DOE
7. School handbooks - look especially at policy and procedures for absences, fire drills, immunization, grievances, etc.
8. Job announcements
9. Schedules (lunch, yearly calendar, professional inservice days, etc.)
10. Announcements of professional development workshops (see Now especially)
11. Typhoon policy or procedures
12. DOE policy statements on leave, political activity, etc.
13. Teacher's editions
14. Memos from the Certification Officer
15. Lesson plans
16. Criteria for field trips/activities
17. Messages taken by office staff or aides (especially telephone messages)
18. Notes from parents
19. Texts used in classes

SUGGESTIONS FOR
ORAL INTERVIEW ITEMS/LISTENING COMPREHENSION ITEMS

1. Explain educational policy to parents (attendance, discipline, grading, etc.)
2. Explain a homework assignment to students
3. Teach a concept that is important in your class - for a few minutes - or explain how you would teach it
4. Describe a student problem to parents
5. Explain to a student or to a parent how you arrived at a grade
6. Participate in a professional meeting with colleagues
7. Explain your educational objectives to students or parents
8. Imagine it's the first day of school; review your expectations and plans for the year.
9. Read and explain written instructions for standardized tests such as the SRA or the Guam BSMT
10. Explain to students how to do a particular task or assignment
11. Narrate something which happened to a student to an administrator (for example, an accident or a problem)
12. Place a telephone call to parents about a student's problem (absences, discipline, potential failure, etc.)
13. Imagine you are in a parent-teacher conference; begin by discussing what you have done thus far in the quarter, how a sample student is doing, and how the parent(s) could help that student
14. Explain fire drill procedures or typhoon procedures
15. Give directions to a student to the nurse's office
16. Give directions for a test
17. Explain grading procedures to students or parents or an administrator
18. Ask students to explain what they are reading or doing
19. Teach and re-teach a difficult concept or term
20. Give directions for a homework assignment
21. Give directions to a parent volunteer or teacher aide
22. Explain to a new teacher how you go about organizing for a successful class
23. Give directions for a book report
24. Answer questions about terms, homework assignments, test items, etc.
25. Follow-up on parental inquiry (for example, when parent calls and says that his/her child has said that she/he needs to bring $50 to school)
26. Explain lunch room policy to a new student
27. Explain typhoon or fire drill policy to students or aides or a new teacher
28. Talk to a student or the class after a student has laughed at another student's mistake
29. Describe a particularly successful class
30. Introduce a new unit or lesson
31. Evaluate a student's work
32. Describe your objectives for a field trip to a principal or parent
33. Introduce a film, filmstrip, or video to your class
34. Correct a student's wrong answer to a quiz

SOME SAMPLE GRAMMATICAL FEATURES OBSERVED IN GUAM ENGLISH SPEAKERS

1. Plural /s/ added where not necessary - clothings; luggages; equipments; furnitures; machineries

2. Have you ever eaten chicken? Yes, I have ever.

3. Just only - as in I have just only five copies

4. Adverb placement -
   I want you to check also your work.
   You should take only the test.
   Everyday you're sleeping, Chris?
   I agree also with Juan.

5. Articles -
   are additional burden to the teachers
   noun is

6. was/were - There was no words from DOE to dispel this belief.

7. Share us

8. Tense sequences - She'll tell us what she had/have seen.

9. Tags - He has grown up, isn't it?
   They have more farm equipments, isn't it?
   There are different ways of running the farm, isn't it?

10. Question formation -
    Now this picture, it was taken about how many years?
    O.K. You need also what?

11. If I reach my, let's say, my fifty years old.

12. An ethnic group is a group of people who are having the same languages.

13. We have not too many buses.

14. Agreement - This is the instructions.

SAMPLE SPELLING PROBLEMS

1. swiming

SAMPLE VOCABULARY FEATURES

1. paper toilet - for toilet paper
2. off the lights
3. play basket - for basketball
4. list down the things
5. unfamiliarity with American idioms such as:
   - don't pull the wool over my eyes
   - don't pull my leg
   - row upon row

SOME GENERAL TESTING CONCERNS/RECOMMENDATIONS

1. Make sure that there is enough context in the test so that distinctions are clear. * The reef is beautiful vs. The wreath is beautiful.

2. Make items relevant to teaching situations. * Directions for wrapping a present vs. Directions to students or parents or others relevant to schoolwork.

3. Use a live administrator for the oral test. Mechanical difficulties, problems in becoming accustomed to a microphone, differences in the output from the headphones, etc. can interfere with the test.

4. Make the vocabulary relevant to teaching.

5. Choose a testing site which is not noisy. Don't give it at the school in an office or during recess.

6. Avoid cloze tests. If using reading passages, base them on things that teachers read. If using a cloze, modify it and provide multiple choice answers.

7. Have multiple forms of the test.

8. Present test as an opportunity for professional development, not as a threat.

9. Don't give test at end of school year.

10. Make sections independent. If someone fails only one section, let them repeat only that section.

11. Don't make this test a test for those who fail the BESTE a second time. Keep the tests independent.

APPENDIX B

COMPLETE SPECIFICATIONS FOR THE LISTENING AND READING COMPREHENSION SECTIONS OF THE GETEP

GETEP SPECIFICATIONS

Test Assembly Specifications:

Section I, Listening Comprehension (50 items)
- Part 1: 20 Short Dialogs
- Part 2: 30 Extended Dialogs and Monologs items
  (Each ED or M followed by 4-6 questions.
  Balance number of extended dialogs [3] and
  monologs [3] in each test form)

Section II, Reading Comprehension (50 items)
- Part 1: 20 Cloze
  (2 short passages: 1 practical, 1 education
  related)
- Part 2: 30 items based on passages
  (5-6 passages)

All items will have four options: (A), (B), (C), and (D).

DESCRIPTION OF ITEM TYPES

SECTION I, PART 1.
Short Dialogs (20 items). One item in each category is to be included in each form of the test. Classification: SD+number

1. Overstatements or understatements (e.g. We've already done a million of those problems! Not much, I didn't! [like the test.])

2. Rhetorical questions: (What difference does it make? Who cares?)

3. Exclamatory response: (AF: We'll have to reschedule the film.
SF: Why not the test!)

4. Questioning first speaker: AM: We'll take a break after we finish the next group of exercises.
SM: Can't we take it now?

5. Disagreeing with first speaker (response generally contains a tag question): (But they didn't win, did they?)

6. Limiting or qualifying first speaker: (This is only my third cookie.)

7. Responding to a question with a question (Second question is often a suggestion): AF: Who would be willing to give the first oral report?
SM: What about Jason? He always has his assignments ready early.)

8. Shortened (reduced) responses: AF: Have you handed in any of your homework assignments late this marking period?
SF: Only twice, I think.

9. Indirect answers: SF: May I hand in the assignment tomorrow?
AM: You can finish it during study hall today.)

10. Causatives: (I got Mike to do that part of the assignment.)

11. Commands or requests: (Could you make the paper due on Friday instead? We have play rehearsal tonight.)

12. Common expressions: (e.g. How come? You can say that again!; Beats me!; Come on!)

13. Common expressions (literal or metaphorical): (e.g. kill two birds with one stone; the usual song and dance; throw in the towel; put two and two together)

14. Two or three word verbs: (e.g. team up with; be on to; run up against; put one up to something)

15. Inference: Speaker implies something without saying it specifically.

16. Stress on wh- word (to determine if examinee understands when speaker is requesting either a repetition or new information): e.g. Who? [rising intonation, asking for repetition; falling intonation, asking for new information or answer).

17. Stress on auxiliary word in response: (Oh, she <u>did</u> give us homework [assumption that teacher had not given homework])

18. Stress on word or phrase in dialog to clarify who, what, when, why, how, etc.: It was <u>his</u> book. It was his <u>book</u>. It <u>was</u> his book. [This stress on aux. may occur in first speaker's lines.]

19. Repeating first speaker's statement with change in stress on a sentence or question element: AF: I've seen better films.
SM: Seen <u>better</u> films? I certainly hope so. That one was dead boring!)

20. Phonology (testing final consonants or consonant clusters or the lack of them): (e.g. peace-peas; face-faze; place-plays; loss-laws; mass-mash; catch-cash; crutch-crush; lunge-lunch; bag-back; snag-snack; dug-duck)

SECTION I, PART 2

Extended Conversations and Monologs (30 items total based on 5-6 stimuli). Items attached to the stimulus material should include the following types of questions:

1. Main topic: (e.g. What is the main topic of the conversation [talk, announcement]?) Classification: EC1 or M1

2. Supporting ideas: (See examples in "Instructions for Item Writers.") Classification: EC2 or M2

3. Inferences: (e.g. What does the speaker (teacher, student) imply about _____? Where is this conversation most probably taking place? What probably prompted this announcement? etc.) Classification: EC3 or M3

SECTION II, PART 1

Multiple-Choice Cloze Items (20 items, 2 passages with 10 items each). The first 1-2 sentences and final 1-2 sentences should be left intact. Average number of words between deletions can range from 4-15 with an average of 8-9 words. Points tested should include the following:

1. Reading Comprehension/Grammar (4-5 per passage). These items should focus on logical connectors, or grammatical structures that are determined cross-sententially and are necessary for paragraph cohesion. (See examples in "Instructions for Items Writers.") Classification: RG

2. Reading Comprehension/Vocabulary (5-6) per passage). These items require a lexical choice based on information in other clauses or, preferably, other sentences. (See examples in "Instructions for Item Writers.") Classification: RV

SECTION II, PART 2

Reading Comprehension Passages (30 items total based on 5-6 passages). Passage length can range from 150-250 words. Questions about each passage should include the following types.

1. Main idea: (e.g. What is the main idea of the passage? What is the main purpose of the passage? With what topic is the passage mainly concerned?) Classification: RP1

2. Supporting ideas (asking about facts presented in the passage): (e.g. According to the passage,...?; According to the author,...?; Which of the following...does the author mention as...?; Which of the following...is NOT mentioned as...?; When...?; etc.) Classification: RP2

3. Inferences: (e.g. It can be inferred from the passage that....; The author implies that....; What did the paragraph preceding the passage most probably deal with?; What will the paragraph following this passage most probably discuss?; Where would this passage most probably be found?; What was the author's purpose in writing this passage? What probably prompted the author to write this passage?) Classification: RP3

APPENDIX C

ITEM-SPECIFICATION TABLE FOR THE FINAL LISTENING COMPREHENSION SHORT DIALOG SECTION

Item-Specification Table for the

Final Listening Comprehension Short Dialog Section

This table lists the item number from the short dialog part of the Listening Comprehersion section of each final form of the GETEP along with the original content specification reference number (Spec #). The description of each Spec number is given in Appendix B of this report.

Form 1		Form 2		Form 3		Form 4	
Item	Spec #	Item	Spec #	Item	Spec #	Item	Spec #
1	1	1	1	1	1	1	1
2	2	2	3	2	2	2	2
3	3	3	4	3	3	3	3
4	4	4	5	4	4	4	4
5	6	5	7	5	5	5	5
6	8	6	8	6	6	6	6
7	9	7	11	7	7	7	7
8	10	8	12	8	9	8	8
9	11	9	14	9	10	9	9
10	12	10	15	10	11	10	10
11	13	11	16	11	12	11	14
12	14	12	17	12	15	12	15
13	16	13	18	13	16	13	16
14	19	14	19	14	17	14	18
15	20	15	20	15	18	15	19

APPENDIX D

INSTRUCTIONS FOR THE GETEP WRITING SAMPLE

Instructions for the GETEP Writing Sample

This part of the Guam Educators' Test of English Proficiency will allow you to demonstrate your ability to write in English. You will be given a writing task related to the field of education. You will have 30 minutes to plan, write, and correct your writing. Your writing will be graded on its overall quality.

1. When the supervisor tells you to begin, go to the next page and read the writing task carefully.

2. Give yourself three to five minutes to think about what you are going to write. Making notes may help you organize your thoughts.

3. Write in the format appropriate to the writing task. Write clearly and precisely. Respond to the task in as complete a manner as possible, using examples and supporting points as appropriate. Remember that how well you write is more important than how much you write, but make sure you have covered all the aspects of the writing task appropriately.

4. Check you work. Allow a few minutes before time is called to read over your essay and make minor revisions.

5. After 30 minutes, the supervisor will tell you to stop. You must stop writing and put your pencil down.

APPENDIX E

GETEP WRITING SAMPLE SCORING GUIDE

AND BENCHMARK SAMPLES

Guam Educators Test of English Proficiency
Writing Test
Scoring Guide

5 Clearly demonstrates competence on both rhetorical and syntactic levels, though it may contain occasional minor errors.

A paper in this category
- is well organized and developed throughout
- effectively addresses the writing task
- uses appropriate supporting details in a manner that clearly supports a thesis or illustrates ideas
- shows unity, coherence and progression
- displays consistent facility in the use of language
- demonstrates syntactic variety and appropriate word choice
- can be understood effortlessly in a quick reading

4 Demonstrates competence in writing on both the rhetorical and syntactic levels. It may have occasional errors.

A paper in this category
- is well organized and developed
- adequately addresses the writing task
- uses appropriate supporting details to support a thesis or illustrate ideas
- shows unity, coherence and progression
- demonstrates syntactic variety and appropriate word choice

3 Demonstrates minimal competence in writing on both the rhetorical and syntactic levels, though it may contain errors.

A paper in this category
- is generally organized and developed, though it may have fewer details than does a 4 paper
- may address some parts of the task more effectively than others
-shows unity, coherence and progression, though not to the degree of a 4 paper
- demonstrates some syntactic variety and range of vocabulary
- displays general facility in language, though it may have more errors than does a 4 paper

2 Suggests minimal competence in writing on both the rhetorical and syntactic levels.

A paper in this category
- is adequately to minimally organized
- addresses the writing topic adequately but may slight parts of the task
- inconsistently uses details to support a thesis or illustrate ideas
- demonstrates minimal facility with syntax and usage
- may contain some serious errors that occasionally obscure meaning

1 Suggests incompetence in writing, remaining flawed on either the rhetorical or syntactic level, or both.

A paper in this category may reveal one or more of the following weaknesses:
- inadequate organization or development
- failure to support or illustrate generalizations with appropriate or sufficient detail
- an accumulation of errors in sentence structure and/or usage
- a noticeably inappropriate choice of words or word forms

Writing Prompts for Benchmark Sample

You are a teacher at a local public school. Your school has received a $15,000 grant to improve its facilities. Debate is centering on whether to spend this money to build the computer lab, to air condition classrooms, to better stock the library with books and periodicals, or to purchase athletic equip . You now have the opportunity to present your own views on ho ne money should be spent. Write a letter to Mrs. Blanca Cardenas, principal of your school. Clearly state which way you think the money should be spent and give reasons to support your position.

You are a teacher at a local public school. You have been asked to be on a committee to develop standards for an excellence-in-teaching award. Write a memo to the chair of the committee in which you describe the qualities you think an excellent teacher must possess, and why you think these qualities are so important.

(A)
Score = 1

To: Mrs. Blanca Cardenas:

This is my honor to have this great opportunity to present my own view on how the $16,000 grant should be spent. I would like to give a suggestion to improve its facilities.

I concern that to use this $16,000 grant to better stock the library with books and periodicals. The reason of this is that Guam is a small and lacking for any kind of books supplies, especially the school supplies, books and periodicals. If we may use this grant to full up the library with books and periodicals, at least the students they would like to spent more more times and be at the library and they would be able to learn more things, too. I believe that would help the to improve the education or degree of Guam.

(B)

Score = 2

To: The principal, Mrs Blanca Cardenas

From: Teacher Billy Joe.

I would like to suggest my idea of how to Spend the $15,000 grant that our School received for this School year.

I would suggest that we use the money wisely. I have been thinking about some wise ways to use the money. Since our lil rary doesn't have a good Source for our students to do their researches. I would strongly Commend that we use that money to improve our library. Let's use this money to better stock our library with books and periodicals. I know that is wise Way to use the money and I will encourage the students to study hard. The reference books and all the other articles are very old stuff. The students are Complaining and having difficulties to find what they're asked for in our library. That is why we have found many student asleep in the library. Some are making this out of date books be their excuses not to do ther assignments or researches. There are many of our Students transfering because of the out of date Sources that they are using for their study. researches and studies

Dear Mrs. Blanca Cardenas

Score = 3

In regards to the $15,000 grant our school has recieved to improve the facilities I think we should concentrate on spending this money on the computer lab. on spending this money we should concentrate on spending the money the school has recieved for building the computer lab because it is essential for our students to learn not only from teachers but also from computers. Learning from a teacher is very good but if our students learn from computers it would be rewarding for them. Learning from a computer is fun; it teaches students to think out different forms. The computer has to offer not like if students were learning from teachers in which they would have to stick to one basic styles or option.

Because society is changing and so is our technology, we should let students be familiar with computers. Computers are being used by many people mail especially adults, we should attend the use of computers to students which will benefit them later in the near future where basic knowledge of computers is constantly in demand.

I strongly suggest that we should spend the money on building the computer lab to let our students learn from them and

to familiarize them with computers as it is important in today society.

Thank You for your consideration.

Josephine F Reed

(D)

Score = 3

To the Chairman of the committee,

Herewith are some of the qualities I feel a teacher must possess. First of all a teacher should have the ability to teach in a classroom — that is a college degree. or else they wouldn't be there in the first place. In a person earns that degree they learn various things but the things they will learn in class are very different. the more time they spend teaching - the more they learn and gain new skills. Teacher must also have respect for the students as they should have respect for the teacher Communication comes along with this If a student has a problem she in she should be able to approach the teacher and discuss the problem. Communication is essential in the classroom because in order for a class to be cooperative with each other they must talk with each other I also feel that a teacher should be able to get along with each other. Once in a while they should take with each other and just enjoy each others company. So mainly a teacher who can be eligible for an excellent in teaching award is one who has discipline patience, respect, and good communication skills.

Sincerely, James H Dallara

Score = 3.5

Dr excellent teacher have qualities that blend well together making her an exceptional teacher. The teacher's educational background, creativeness, & personality personality and creativeness are some of the qualities she needs.

The educational background of the teacher is very important. The teacher's main purpose is to teach and so by teaching the knowledge needed, the student are the ones who will eventually suffer from this logic consequence.

In addition to having a solid knowledge of a subject, it is also very important that the teacher be able to communicate effectively. It is utterly useless to be the smartest teacher, yet not able to convey the materials needed to be learned by the students. The communication and clarification are very important for the learning process of students as well as the teacher-student interaction.

An exceptional teacher must also give the students the space & room individually. She should not restrict their own imagination, ambitions, and

inquiries but help the students broaden and approach their interests. This is a good way to give students an idea of what they probably want to be later on in life.

The teacher must also plan various activities in-and-out of class to prevent boredom among students. Field trips and presentations of a book or a play are few the suggestions. The teacher must always challenge the students helping them reach their maximum potential.

Lastly, the personality of a teacher is very important. A teacher must respect her students so that they will respect her back. Showing her concern and willingness to help a student overcome a problem will mean a great deal to the student.

It is vital that a teacher not criticize ~~the~~ a student in front of his peers. During mid-school and high school years, peers play a very important role in his life. By criticizing him in front of his classmates, the

Teacher is throwing the student's self-esteem and he may be scared for life of thinking "he is not worth it"

Even if the students are not a exhibit, a good teacher pays special attention to them. The teacher must always encourage her students that they are all special and possess unique talents. A good teacher will guide her students toward the road of success. It is also important to make her students know that failure can be overcome by being persistent.

An excellent teacher is hard to find. Possessing all the qualities that be mentioned above require commitment to the teaching profession. A teacher must not only have the knowledge to teach, but she must have the compassion and various techniques of teaching helping her students become a better citizens.

(F) page 1

Score = 4

Dear Sir of Madam:

I feel that the qualities a teacher must possess to be eligible for the excellence-in-teaching award are the following: First, the teacher should be certified and knowledgeable in that subject area. It is essential for the teacher to know and be able to teach the subject area with expertise and ease. The teacher should be clear and consistent when communicating with the students about the subject. And, the teacher should be trained good methodologies for effective teaching. This characteristic would enhance learning if other methods of teaching are unsuccessful. Third, the teacher should possess a personality of being able to relate the students needs or problems. The teacher should be like a friend that students need to talk to, or a person to give advice. The teacher should be dedicated to helping students who need extra instruction about the subject or those who need help of anything at all. Fifth, the teacher should also have good discipline in the classroom. The teacher should be able to control the students in the class by expressing your rules and regulations that a student must follow. Lastly, the teacher should be able to encourage his or her students to do better, as

~~I encourage Barbara~~ this would enable students to prepare better and to set goals for themselves.

I feel that the qualities that I have stated are essential for selecting the excellence-in-teaching award. I hope that the teacher who is eligible will possess all these characteristics and would also be able to pass it on to other teachers who are looking these abilities.

Thank you for your kind consideration.

Sincerely,
Diana Keane

Score - 1

OCT 16 1984

Dear Mrs. Cardenas,

It has come to my attention that the school has received a cash allowance of $15,000 to be used for campus improvement. I firmly believe that the money would be best used for air conditioning, because of the terrible heat coupled with the high humidity. This creates a stressful environment for students to learn in. The heat distracts students from lectures as well as discussions. If air conditioning was installed it would create a comfortable and relaxing environment advantageous for learning. Students would be more receptive to lectures without the distractions caused by the heat (like sweating).

Also it would benefit each student equally since everyone would have air conditioned class rooms. If the money was used for sports equipment only a small percentage would benefit (athletes) rather than the student body as a whole.

Overall air conditioning would spur increased interest on education and benefit everyone including the faculty.

Thank-you for your time and patience,

Henry San Nicolas

HENRY SAN NICOLAS OFECIAR

PROF. SOCIAL SCIENCES

(H) page 1
Score = 4.5

TO: CHAIR OF THE COMMITTEE ON EXCELLENCE IN TEACHING
FR: Peter McIntyre
SUBJ: Qualities of an excellent teacher

Sir/madam

Hereforthe I will relay to you my personal opinion on the Subject of The Qualities of an Excellent teacher. They Are by no means in Any Specific order. I deem These conclusions Through my extensive Schooling and how I perceived an excellent teacher during That time period.

One quality which I greatly Admired was patience. A teacher who could put up with what I disliked out and not "lose his cool" gained my upmost respect. If I could Ginger that particular teacher consistently, He would, in my probability eyes, lose credibility. The teacher was human, not the Superman I could look up to. Sometimes, just for the attention, I would try extra hard to irritate the teacher. Yes, Patience is a virtue.

Another trait which I found most desirable in a teacher was knowledge. A teacher who could Answer Any question I had asked on the subject filled me with respect. The Superman who knew "everything." Some teachers could not answer the most basic questions on the subject lost my respect in Newco Seconds. Knowledge is Power.

All this knowledge though is worthless if the teacher cannot effectively communicate with his pupils. The teacher must be able to get the point across. If the teacher has, says, a speech impediment, students would not learn how to properly articulate words. This is especially important in subjects such as english, science and mathematics. It is an invaluable quality. If the student teacher cannot properly explain the central idea, the student will not learn. Communication binds the world together.

Good moral character is another desirable trait in a teacher. That is not to say he should be an angel, by Heavens NO! But the teacher must set a good example for students to follow. Students will look up to their teacher. Giving the teacher an advantage over students. The teacher should be prompt to class to show adherence to the rules while encouraging students to do the same. Hypocrites are not taken seriously.

Last but not least is discipline. A teacher should not allow any students (Athletes inclusive) to cruise through school. The teacher must ensure that students learn and have the discipline to see that the task is accomplished. Without discipline there is no order.

Although I could elaborate on these and many more qualities I feel desirable in a teacher, I feel the preceding five qualities are sufficient enough to choose a truly excellent teacher.

Thank you
Pete

APPENDIX F

LETTER DESCRIBING REQUIRED CHARACTERISTICS OF INDIVIDUALS COMPETENT TO MAKE JUDGEMENTS ON THE PASSING SCORES FOR THE GETEP

CAL Center for Applied Linguistics

September 27, 1989

Ms. Rita Northway
Department of Education
Government of Guam
Agana, Guam 96910

Dear Ms. Northway:

As promised, I am writing to provide additional information on the procedure we plan to use to determine a passing score on each form of the test. This procedure would be used when I come in November, not during the October trip I proposed to you yesterday.

The passsing score procedure is called Angoff's Method 1. It is commonly used in setting passing scores on ocupational tests. The quality of the results depends in good part on the quality of the judges who participate.

First, the procedure involves selecting and bringing together a group of judges who are very familiar with the practitioners who will subsequently be tested. The judges then discuss the characteristics of competent and incompetently professionals (in this case teachers). Eventually, a consensus is reached as to what is an adequately competent (not outstanding) teacher in terms of English language proficiency. Following this concensus building educational exercise, the judges each read each item on each form of the test. After considering each item, each judge estimates and records next to the item the probability that this adequately competent teacher will be able to answer the item correctly. Although these opinions are reached independently, I will initially, and then occasionally, ask the judges to state the probability they assigned in order to determine that each understands the process and is applying it in a logical, careful, and consistent manner.

Another way of posing the probability question is by telling the judges to estimate how many members of a group of 100 minimally competent teachers would answer the item correctly.

If the item has not been changed following pretesting, I will also provide the judges with information on the proportion of the pretest sample that answered it correctly. However, I would caution them that the pretest sample probably includes a range of competencies, and is not composed solely of people whose competency is only adequate. Thus, although the judges may consider the pretest data, they will have to make their own determination about the difficulty of the item for a sample of borderline or minimally adequate teachers.

After the first item is rated, the different estimates are averaged across the judges. Thus, if the average estimation is .65, then .65 is taken as the experts' rating of the probability

4448 22nd Street, N.W. Washington, D.C. 20037 (202) 429-9292 Fax (202) 429-9766

that a minimally competent group of teachers will answer the item correctly. The same procedure is repeated for each item on the test, and the group-average rating for all items is averaged to get the average rating for the entire test. Thus, if the average difficulty rating for all items on the test were .70, then .70 would be the estimate of how an adequately competent or minimally acceptable teacher would perform on the test. When put into practice, examinees who score below .70 would be considered as not adequately competent, whereas examinees who score .70 or above would be considered to have demonstated adequate or greater than adequate competency.

In order for the procedure to work best, the judges must have experienced or observed teachers with different levels of competency. That is, the judges must have seen teachers whose English is adequate or better and teachers whose English is inadequate. It would also be helpful it the judges have a good sensitivity to the English language. Sensitivity to language is helpful since the judges will have to read each question on the test and make a judgement about its general difficulty. Then, they will have to consider its difficulty for the teacher whose proficiency is only minimally adequate.

All people with a legitimate stake in the outcome should have representation on the panel of judges. Thus, the panel might include someone from the certification office, an ESL consultant with the DOE, a representative of the teachers or teachers union, a school principal, a parent or PTA member who is concerned about English proficiency, and a teacher trainer. Since the DOE is ultimately responsible for setting teacher standards, it may legitimately appoint the panel and name more than one of its staff to it. I believe that six to eight linguistically sensitive judges who have observed many teachers would produce a valid and reliable standard.

The suggested passing score produced by the panel would be presented to the DOE. The DOE could either accept the score, raise it or lower it, depending on relevant internal considerations.

I hope this information about the standard setting process is useful to you. You may want to begin thinking about potential members of the panel of judges. The panel would meet for three days, which should give it adequate time to be trained in the process, then read, consider and rate each item on all four forms of the test.

I could discuss the standard setting process and any other aspect of the test development process with you on my first trip there. Although I haven't heard from you yet, I do hope that it will be possible for you to arrange for me to train 3-5 persons in oral proficiency testing during the dates I indicated in my letter to you yesterday.

I look forward to hearing from you soon so I can begin to make travel arrangements.

Sincerely,

Charles W. Stansfield

Charles W. Stansfield
Division Director,
Foreign Language Education
and Testing

cc: Dan Robertson
Hector Nevarez

APPENDIX G

JUDGES' RATING SHEET FOR THE PASSING SCORE STUDY

GETEP Form _____ Name of Judge _____

ITEM	Probability/proportion of correct response															
1)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
2)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
3)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
4)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
5)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
6)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
7)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
8)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
9)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
10)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
11)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
12)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
13)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
14)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
15)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
16)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
17)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
18)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
19)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
20)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
21)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
22)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
23)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
24)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
25)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
26)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
27)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
28)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
29)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
30)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100

GETEP Form _____ Name of Judge _____

ITEM Probability/proportion of correct response

31)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
32)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
33)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
34)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
35)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
36)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
37)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
38)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
39)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
40)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
41)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
42)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
43)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
44)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
45)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
46)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
47)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
48)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
49)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
50)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
51)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
52)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
53)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
54)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
55)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
56)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
57)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
58)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
59)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
60)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100

GETEP Form _____ Name of Judge _____

ITEM	Probability/proportion of correct response															
61)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
62)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
63)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
64)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
65)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
66)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
67)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
68)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
69)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
70)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
71)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
72)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
73)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
74)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
75)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
76)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
77)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
78)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
79)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
80)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
81)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
82)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
83)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
84)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
85)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
86)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
87)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
88)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
89)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
90)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100

GETEP Form _____ Name of Judge _____

ITEM	Probability/proportion of correct response															
91)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
92)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
93)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
94)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
95)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
96)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
97)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
98)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
99)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
100)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
101)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
102)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
103)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
104)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
105)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
106)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
107)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
108)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
109)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
110)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
111)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
112)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
113)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
114)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
115)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
116)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
117)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
118)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
119)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
120)	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100

APPENDIX H

COMPLETE RESPONSES TO THE JUDGES' QUESTIONNAIRE

Judges' Questionnaire
Summary of Responses

1. Please indicate how you feel about the passing score that was set for the writing sample (3.5).

a. Too severe
b. About right 7/7 (100%)
c. Too lenient

(Note: One judge wrote:) Can't judge criteria, wasn't part of process.

2. Please indicate how you feel about the passing score that was set for speaking (3).

a. Too severe 1/8 (13%)
b. About right 7/8 (87%)
c. Too lenient

3. Do you feel that the standard-setting procedures used in this study were appropriate?

a. Yes 8/8 (100%)
b. No

The selection procedure for the perfect teacher group was faulty, the group wasn't representative of minimally competent teachers.

4. You were provided with statistics showing the proportions of the pretest sample that answered each multiple-choice item on the test correctly. When determining the difficulty of items for the minimally competent teacher, did you find it useful to have this information?

a. Yes 7/8 (87%)
b. No 1/8 (13%)

5. Please make an evaluative comment concerning the panel of judges. (Were the other members highly professional? Well qualified? Did they take the process seriously? etc.)

I feel that the panel of judges contained highly qualified individuals who took their job seriously and performed professionally. Setting standards for teachers is a monumental task that takes a lot of soul searching on the part of the individuals on the panel. They did the job extremely well.

Highly qualified, motivated, enthusiastic, professional and extremely vocal. Also, hard working and task oriented.

Highly professional and very well qualified.

Judges were of high quality material.

I believe that the panel of judges who participated were very professional and well qualified to partake in this venture. I wished the panel of judges had listened to the Listening Comprehension Tape, only two judges did.

All members were highly competent and for the most part acted very professionally.

The panel members were "professionals" - they each took the process seriously. However, the group should have included other members of various language groups on the island. We could have used two other Filipinos as well.

Most of them were highly professional and qualified. I would have liked a definite criteria though for selecting participants. One requirement is that education personnel on the panel have at least five years of teaching experience.

6. What is your opinion of the quality of the GETEP test?

I feel the test ranks a five on a scale of 1(low) to 5(high) in terms of quality. I feel that questions about culture insensitivity will be laid to rest by this test.

I think it is a good measuring device for incoming teachers. The relevance of material to education is terrific.

High quality test, looks excellent.

For being a "customized" test, I would say on a scale of one to five, four. The test developing center did a very good job!

Very good.

In general the quality was good, but I feel a few discrepancies exist.

Quite good. But the district may be biting off more than it can chew come time for implementation or administration of the GETEP.

The direct written test is of high quality. The listening comprehension tapes can be improved more.

7. If you wish to make any comments about the standard-setting process, please write them here.

Although a lot of hard work, I thoroughly enjoyed it and am anxious to see the results and hear how the testees do.

Charles, the CAL's rep, did an outstanding job of facilitating the process and relieved the monotony and tediousness of the process by allowing the judges ample opportunity to voice their opinions about any section of the test.

Organized, well facilitated, good experience.

We worked well together We disagreed, yes, but we were also open-minded.

I think it was fair.

The panel of judges should have evaluated the GETEP prior to the field testing. Because many of the test items are specific only to Guam.

Testing time frame should be shorter to work for nation wide recruitment.

Although item difficulty had already been established and "saving time" was important, I think the panel should have listened to at least two of the four listening comprehension tapes. The pace, style of conversation, pronunciation, etc. all contribute to the clarity of the message. Distracting aspects of the dialogue can impede comprehension.

APPENDIX I

PROPOSED NCS MACHINE-READABLE ANSWER SHEET

Last name	First name	MI

TEST DATE	SS NUMBER	BIRTH YEAR	TEST CENTER	TEST FORM
			2 COLUMNS	1 COLUMN

126

SIDE 1

GUAM EDUCATORS TEST OF ENGLISH PROFICIENCY

1. Sex: (1) Male (2) Female..... ①②

2. Native language: (1) English (2) Chamorro (3) Filipino (4) Chinese (5) Korean (6) Other (specify) [ ] ..... ①②③④⑤⑥

3. Which grade level do you teach? (1) Elementary school (2) Middle or Junior High School (3) High School.... ①②③④

4. What is the location of the university from which you graduated? (1) Guam (2) Hawaii (3) US-mainland (4) Philippines (5) Other [ ] ..... ①②③④⑤

5. Where are you taking this test? (1) Guam (2) Hawaii (3) U.S. mainland (4) Other [ ] ..... ①②③④

Items 1-40

127

SIDE 2

ITEMS 41-60

ITEMS 61-90

ITEMS 91-120

GUAM EDUCATORS TEST OF ENGLISH PROFICIENCY

Center for Applied Linguistics

EXAMPLES	IMPORTANT DIRECTIONS FOR MARKING ANSWERS

Signature

Test booklet number