PERFORMANCE ASSESSMENT OF ESL AND EFL STUDENTS

JAMES DEAN BROWN, THOM HUDSON, JOHN M. NORRIS
University of Hawai'i at Manoa
WILLIAM BONK
Kanda University of International Studies

ABSTRACT

Thirteen prototypical performance tasks were selected from over 100 based on their generic appropriateness for the target population and on posited difficulty levels (associated with plus or minus values for linguistic code command, cognitive operations, and communicative adaptation, as discussed in Norris, Brown, Hudson, & Yoshioka, 1998, after Skehan, 1996, 1998). These 13 tasks were used to create three test forms (with one anchor task common to all forms), two for use in an ESL setting at the University of Hawai'i, and one for use in an EFL setting at Kanda University of International Studies in Japan. In addition, two sets of rating scales were created based on task-dependent and task-independent categories. For each individual task, the criteria for the task-dependent categories were created in consultation with an advanced language learner, a language teacher, and a non-ESL teacher, all of whom were well-acquainted with the target population and the prototype tasks. These criteria for success were allowed to differ from task to task depending on the input of our consultants. The task-independent categories were created for each of three theoretically motivated components of task difficulty in terms of the adequacy of: (linguistic) code command, cognitive operations, and communicative adaptation. A third rating scale was developed for examinees to rate their own performance in terms of their familiarity with the task, their performance on the task, and the difficulty of the task. Pilot data were gathered from ESL and EFL students at a wide range of proficiency levels. Their performances were scored by raters using the task-dependent and task-independent criteria. Analyses included descriptive statistics, reliability estimates (interrater, Cronbach alpha, etc.), correlational analysis, and implicational scale analysis. The results are interpreted and discussed in terms of: (a) the distributions of scores for the task-dependent and task-independent ratings, (b) test reliability and ways to improve the consistency of measurement, and (c) test validity and the relationship of our task-based test to theory.

INTRODUCTION

This introduction will provide background to the current paper by addressing some basic questions: (a) where did performance assessments come from? (b) what are the advantages and disadvantages of performance assessments? (c) what are performance

University of Hawai'i Working Papers in ESL, Vol. 18, No. 2, Spring 2000, pp. 99-139.

assessments? (d) what constitutes performance task difficulty? And, (e) what is the purpose of this performance assessment project?

Where did Performance Assessments Come From?

Performance assessments first surfaced in the form of machine performance assessments. Such assessments usually aimed at finding out how durably/reliably a certain type of machine would perform in a standard operating environment. When the same sorts of questions were asked about people and their job performance, human performance assessments were born.

Consider what a performance test for an airline pilot might look like. In addition to paper-and-pencil tests, physical examinations, psychological tests, and infinite hours of experience, most of us who fly regularly would also like to have pilots pass a rigorous performance test in a flight simulator as well as be observed regularly at the controls of a real airplane of the sort for which they will be qualified. What types of tasks would you want a pilot to be able to perform? A formal needs analysis (see Brown, 1995, pp. 35-70) would be useful for determining the sorts of tasks pilots should be able to perform. Just as examples, such a needs analysis might reveal tasks like the following that pilots should be able to do: (a) identify the position and function of all instruments in the cockpit of the aircraft for which they are to be qualified; (b) identify all control components on a schematic representation of the aircraft for which they are to be qualified; (c) perform a basic safety check in a simulator (with random problems arising); (d) communicate successfully in a simulator with air traffic control on taking off, on landing, and in a mid-flight emergency; (e) do visual or instrument take-offs and landings under normal and unusual circumstances in a simulator; etc. Paper-and-pencil tests might be sufficient for assessing pilots' abilities with regard to identified needs like (a) and (b) above, but performance assessments would probably be necessary to measure abilities like (c), (d), and (e). We would also like to note that fairly high standards (like perhaps 80%) would be desirable for passing the simple paper-and-pencil tests of things like (a) and (b), but that, at least from a passengers perspective, very high standards (like perhaps 99%) would be desirable for the performance tests in the flight simulator like those in (c), (d), and (e).

Performance assessments of language use may be very similar to those for job performance. Consider for instance the task (d) communicate successfully in a simulator with air traffic control on taking off, on landing, and in a mid-flight emergency. Depending on the sort of scale that is to be used in rating the pilot's ability to communicate successfully, such a task could just as easily serve as part of a job performance test or a language performance test. Any differences would be based largely

on whether success was to be defined in terms of job performance success or language performance success. [For examples of this type of job-related performance testing see McNamara (1990, 1996), which describes an English test for health professionals in Australia, or Teasdale (1996), which describes a language test for air traffic controllers].

What Are the Advantages and Disadvantages of Performance Assessments?

The literature (e.g., Brown & Hudson, 1999; Jones, 1985; Miller & Legg, 1993; Norris, Brown, Hudson, & Yoshioka, 1998; Shohamy, 1995; Short, 1993) indicates that the primary advantages of language performance tests in education contests are that they:

1. compensate for negative effects of traditional standardized multiple-choice testing (like bias, unnaturalness of language, irrelevant content, etc.)
2. simulate authentic language use by measuring students' abilities to respond to real-life language
3. predict students' future performances in real-life language situations
4. counteract negative washback effects of standardized testing
5. can provide strong positive washback effects, especially if such performance assessments are directly related to a specific program and its curriculum

The literature (e.g., Aschbacher, 1991; Brown & Hudson, 1999; Henning, 1996; McNamara, 1995, 1996; Mehrens, 1992; Messick, 1994, 1996; Miller & Legg, 1993; Norris, Brown, Hudson, & Yoshioka, 1998; Shohamy, 1995) also indicates that the primary disadvantages of performance assessments are that they:

1. are difficult to create (requiring a needs analysis, coordination of teachers etc.)
2. take more time to administer
3. lead to logistical problems
4. cause problems of reliability
5. create problems of validity
6. increase risks to test security

For more detailed information on both the advantages and disadvantages of performance tests see Norris, Brown, Hudson, and Yoshioka (1998).

What are performance assessments?

Many definitions for performance tests have been proposed over the years. For instance, Wiggins (1989) argued for extensive use of authentic tests in educational measurement, saying that such tests should: (a) have collaborative elements, (b) be complex and contextualized, (c) assess real-world tasks, and (d) have authentic standards that are clear to students. As we pointed out in Norris, Brown, Hudson, and Yoshioka

(1998), we feel that these are key characteristics for consideration in designing second language performance assessments.

An example more closely related to second language testing is Shohamy (1995); some of the issues that she raised as important to developing performance assessments were the following:

A. Needs analysis
   1. What criteria should be used?
   2. What content and contexts?
   3. Should a task or item pool be used?
   4. How should experts be used?
B. Nature of instrument
   1. Which and how many tasks should be used?
   2. How long should they last?
   3. How often should they be used?
C. Raters
   1. Who?
   2. How many?
D. Integration of skills with content?
E. Student input in selection of content?
F. Methods for accountability
   1. Should self-assessment be used?
   2. Portfolios?
   3. Multiple judgments?

However, in this paper, we will define a performance test simply as any assessment in which the following three conditions are met: (a) students must perform tasks, (b) the tasks should be as authentic as possible, and (c) success/failure or level of performance on the tasks must be rated by judges according to criteria which are explicitly related to the tasks. Thus, performance assessments will typically be based on authentic tasks, which will be judged by raters on the basis of some form of rating scale. Table 1 provides our expansion of the detailed elements included in our basic three-point definition of performance assessment.

Table 1
Characteristics of Performance Assessments
(adapted from Norris, Brown, Hudson, & Yoshioka, 1998)

1. The tasks should:
   a. Be based on needs analysis (including student input) in terms of rating criteria, content, and contexts
   b. Be as authentic as possible with the goal of measuring real-world activities
   c. Sometimes have collaborative elements that stimulate communicative interactions
   d. Be contextualized and complex
   e. Integrate skills with content
   f. Be appropriate in terms of number, timing, and frequency of assessment
   g. Be generally non-intrusive, i.e., be aligned with the daily actions in the language classroom
2. Raters should be appropriate in terms of:
   a. Number of raters
   b. Overall expertise
   c. Familiarity and training in use of the scale
3. The rating scales should be based on appropriate:
   a. Categories of language learning and development
   b. Appropriate breadth of information regarding learner performance abilities
   c. Standards that are both authentic and clear to students
4. To enhance the reliability and validity of decisions as well as accountability, performance assessments should be combined with other methods for gathering information (for instance, self-assessments, portfolios, conferences, classroom behaviors, and so forth)

What Constitutes Performance Task Difficulty?

One issue that arises repeatedly in the literature is that of task difficulty. Over the past few years, we have struggled with this concept of task difficulty, and the concept is crucial to understanding one aspect in our efforts to develop performance tests as described in the present paper.

Based on a review of the literature on different potential sources contributing to task dificulty, Norris, Brown, Hudson, and Yoshioka (1998, p. 59) summarized as follows:

> Task difficulty, then, will be based on assessment of the variable contributions of the processing components suggested by Skehan (1996): code complexity, cognitive complexity, and communicative demand. Such difficulty components seem to offer a principled means for categorizing *ability requirements* and *task characteristics* that

> are inherent in L2 tasks. By identifying these components within a given task, variable sources of difficulty can be estimated [emphasis in the original].

From this perspective, then, task difficulty can be thought of as the demands made by a given task on the abilities brought to the task by an examinee. Thus, it has been posited that different task qualities and conditions may engage learners' language and cognitive processing abilities in differing ways, and that these differences in processing demand may be systematically related to learners' performances on tasks requiring the use of an L2 (e.g., Skehan, 1998). Of course, whether and in what way learner performances (as well as evaluations of learner performances) may be related to the cognitive processes ostensibly engaged by a given task are empirical questions. As we will explain later in the ***Materials*** section of this paper, in order to investigate this relationship between several hypothesized processing demands made by L2 performance tasks and examinee performances on such tasks, we focused in the current project on the three sources listed above: code complexity, cognitive complexity, and communicative demand. Combinations of these three components were used in estimating the likely difficulty of our test tasks, as well as in developing one set of rating scales (the task-independent scales) for judging students' performances on such tasks.

What Is the Purpose of This Performance Assessment Project?

One purpose of this on-going study was to develop a framework for task-based second language performance assessment that could be adapted by language programs around the United States and elsewhere. To that end, we started with the notion of a curriculum that emphasizes developing second language learners' abilities to use language to accomplish real-world tasks (see also Long & Norris, to appear), and then tried to develop a test that would assess the abilities of students in such a curriculum, while at the same time maintaining as much as possible the real-world nature of the tasks we were trying to assess.

A second purpose of this research project was to examine the degree to which combinations of cognitive processing factors, which are ostensibly engaged by tasks to differing degrees, were useful concepts for helping to understand performance on complex communication tasks, as well as for estimating the difficulty in performing individual tasks in such a test. To those ends, we posed the following research questions for the current preliminary phase of the *Assessment of Language Performance* (ALP) project:

1. How adequate are the distributions of scores for the task-dependent and task-independent ratings?

2. To what degree are the performance tests in this project reliable? And, how can the consistency of measurement be improved?
3. To what degree can interpretations be validly based on the performance tests in this project? And, what is the relationship between examinees' performances and the predicted difficulty levels?

The alpha level for all statistical decisions was set at .05, experiment-wise.

METHOD

Participants

In order to begin to address our research questions, as well as to model the process of performance assessment development within an educational context, we developed the *Assessment of Langauge Performance* (ALP) test to be applicable at our university, which is fairly typical of many United States universities and colleges in terms of the range of international students who attend and the kinds of English language learning and language use demands which face them. Initial data collection for this phase of the project took place between December 1997 and February 1998. To begin with, eight participants completed pilot versions of our 13 operational ALP tasks. These participants were: (a) two L1 speakers of English, (b) three advanced L2 speakers of English, who were graduate students in the Department of ESL at the University of Hawai'i at Manoa (UHM), and (c) three L2 speakers of English enrolled in the English Language Institute at UHM. On the basis of observations of their performances and feedback from them, the 13 ALP test tasks, task instructions, and overall test formats were revised in order to make sure that: (a) future examinees would be able to fully understand what was expected of them on each task, (b) task realia were not ambiguous in the ways they related to expected task performances, (c) administration procedures were clear, and (d) an appropriate amount of time was allowed for each task.

The thirteen test tasks selected for this project were divided up into two forms P and Q such that the two forms had six tasks that were different on each and one task in common (for a total of seven tasks per form). The tasks on the two forms were selected to be at the same levels of difficulty as estimated by combinations of cognitive processing factors engaged (see below).

In order to recruit further participants, advertisements were sent to three organizations at UHM: Hawaii English Language Program (HELP), the English Language Institute (ELI), and the Department of ESL (DESL). Thirty-eight examinees volunteered for the first round of data collection at UHM. The participants from each of these organizations

with their different proficiency levels were assigned to either Form P or Form Q such that these three proficiency levels were about equally represented on the two forms (as shown in Table 2) for a total of 19 examinees on Form P and 19 on Form Q. All participants were compensated with a $10.00 library copy card.

Table 2
Distribution of Participants Each Form by Type of Student

TYPE OF STUDENT	FORM Q	FORM P	FORM J
NSs*	1	1	-
DESL NNSs	2	2	-
ELI NNSs	12	12	3
HELP NNSs	4	4	7
TOTAL	*19*	*19*	*10*

*NS = native speaker; NNS = non-native speaker

In order to compare performances by examinees in an ESL setting with those in an EFL setting, 10 additional EFL participants were recruited for the first round of data collection from EFL classes at Kanda University of International Studies in Japan. These volunteers were tested on a version of the ALP that we called Form J (made up of tasks common to forms P and Q which would make the most sense in an EFL context and sampled at the same difficulty levels represented on P and Q). These examinees were also given approximately $10.00 as compensation. The test administrator in Japan, (who understood the distinctions between HELP, the ELI, and the DESL from several years of experience as an ESL teacher at UHM) estimated the general proficiency levels of the Japanese EFL students to be as shown in the last column of Table 2.

Materials

Stage 1: Needs analysis. As Norris, Brown, Hudson, and Yoshioka (1998) stressed, the first stage in developing performance assessments should be to perform a needs analysis with the goal of aligning assessment tasks as closely as possible with students' actual language learning needs and the objectives of a given curriculum. However, because the ALP test was a prototype intended to model some of the processes involved in developing language performance assessments, we did not want to tie the test to a specific institution or curriculum. Instead, we envisioned a population of students for whom these prototypes might be useful or of generic interest: the more-or-less advanced L2 learners of English for purposes typically associated with United States university settings. With this population in mind, we surveyed a range of text books and language teaching materials in order to select a set of more than 100 tasks which might be relevant

enough for us to flesh out with realia, descriptions, task prompts, and explanations of task parameters (as described in Norris, Brown, Hudson, & Yoshioka, 1998). Table 3 shows two example tasks under the Health and Recreation/Entertainment theme.

Table 3
Example Task Descriptions (from Norris, Brown, Hudson, & Yoshioka, 1998)

Theme A - Health and Recreation/Entertainment

*Task **A.1**: Deciding on a movie* Difficulty Index: **5**

- *Prompt*: Read your friend's note describing when he can go to the movies and what kind of film he would like to see. Then listen to the list of movies from your local movie theater. Pay careful attention to the show-times and the brief movie descriptions. Note titles and times that seem appropriate. Now match up your friend's times and preferences with any of the films that fit both. Call your friend and leave a message on his answering machine giving pertinent information about your choices. Finally, suggest one film that seems preferable to you (be sure to state a reason for your preference).
- *Realia/Materials*: Note from the friend (high-code description; logical organization); tape-recorded list of (multiple, varied) movies and times like you get from US theaters ("Welcome to Varsity theater...), with movie-jargon descriptions of different films and possible show-times, well-organized (parallel to friend's note); telephone; answering machine message from the friend (standard--easy code).
- ...

*Task **A.3**: Planning the weekend* Difficulty Index: **2**

- *Prompt*: Several friends are coming to visit you (e.g., in Honolulu) this weekend. Look through the three following lists: arrival and departure times and pre-determined schedule of activities for your visitors, the things they would like to do while in town, and the weekend entertainment section of the newspaper. After comparing these three sets of information, write out a weekend activity schedule that includes all activities that can be matched up from the three sources of information. Start by including all activities that have already been scheduled.
- *Realia/Materials*: Written notes (e.g., from a previous phone conversation) that have arrival and departure times and pre-determined activities (whale-watching at 5:00 p.m. on Sunday); further written notes about their desired activities (what they heard/read about Oahu ahead of time); entertainment section of local newspaper, isolating only highlighted activities for this weekend (don't want this to be a task of searching for information, rather just organizing it); daily planner type schedule pages with days and times from Friday through Monday.
- ...

These sample tasks were organized into themes, theme subdivisions, and tasks. The following are some examples taken from two themes showing how the themes might be subdivided (from Norris, Brown, Hudson, & Yoshioka, 1998):

D. At Work
   1. Filling the empty position
   2. Applying for a job
   3. Those mundane office chores

E. At the University
   1. Application to a university

2. Registration at the university
3. In-class presentation
4. Responding to a lecture and readings

Stage 2: Selecting and sequencing tasks. The second stage in developing test materials was to look for ways to compare and contrast our tasks so we could rationally select and sequence them. As we pointed out above, we started with Skehan's (1996, 1998) three task difficulty components as a framework for categorizing possible sources of task difficulty in our assessment tasks: code command, cognitive operations, and communicative adaptation.

Table 4
Assessment of Language Performance Revised Task Difficulty Matrix

COMPONENT	EASY>>>>>>>DIFFICULT	EASY>>>>>>>DIFFICULT
Code Command	Range - +	Variety of Inputs/Outputs - +
Cognitive Operations	Organization of Input/Output - +	Availability of Input/Output - +
Communicative Adaptation	Mode - +	Response Level - +

After considering the kinds of characteristics and performance conditions represented in our original pool of tasks, we decided to adapt this framework in order to estimate the likely difficulty of the 13 test tasks we had selected for this project. In adapting the framework, we tried to focus on a minimal set of easily identifiable task characteristics that would likely be associated with each of the three cognitive processing factors listed above. Thus, we wanted to explore one very general way of looking at combinations of task characteristics that might be useful in estimating the kinds of cognitive processing demands that tasks place on L2 users during performance. We posited that with the help of such a framework, differing levels of performance on different tasks might be better understood and generalizations about examinees' abilities on a range of related tasks might be based on observations of performance on a small set of tasks.

Table 4 shows the revised task difficulty matrix (from Norris Brown, Hudson, & Yoshioka, 1998), in which we associated: (a) code command with two kinds of characteristics (code range and the variety of different input and output sources involved in accomplishing the task); (b) cognitive operations with two other characteristics (the organization of the input or output involved in a task and the availability of input for informing the language act involved in a task); and (c) communicative adaptation with two additional characteristics (the language mode(s) needed to accomplish the task and the immediacy required in the responses to information presented in a given task). We posited that tasks would prove more difficult for examinees when they involved greater combinations of these characteristics. [For more detailed information on these variables, see Norris, Brown, Hudson, and Yoshioka, 1998.]

Table 5
Interrater Correlations for Task Difficulty Estimates

Component	Subcomponent	Correlation
Code command	Range	.68
	# of input sources	.77
Cognitive operations	Input/output organization	.75
	Input availability:	.62
Communicative adaptation	Mode	.88
	Response level	.94

Next, two experienced ESL teachers (well-acquainted with the target population for the current study) independently applied the difficulty estimation system to brief descriptions of our collection of more than 100 tasks by assigning pluses and minuses for each of the six task characteristics (indicating that the characteristic was or was not present within a given task in sufficient degrees to make performance demands on the examinee). Interrater correlations for the sums of their pluses on each task for each characteristic (which are shown in Table 5) ranged from a low of .62 for cognitive operations (input availability) to .94 for communicative adaptation (response level). Such moderate to strong correlations generally indicate that the difficulty estimates were reasonably consistent. Table 6 shows how pluses and minuses were assigned by the two raters and the difficulty estimates that resulted for seven out of more than 100 tasks.

Table 6
Example Task Difficulty Ratings

Component:		CC>>>	>>>>>>		CO>>>>	>>>>>>		CA>>>>	>>>>
Characteristic:		Range	#Input Sources		In/Out Organiz.	Input Avail.		Mode	Resp.
Task Themes	Diff. Index								
A.1 Deciding on a movie	5	+	+		-	+		+	+
A.2 Choosing the appropriate film	4	+	+		-	+		+	-
A.3 Planning the weekend	2	-	+		+	-		-	-
A.4 Getting directions to the party	2	-	-		+	-		-	+
A.5 Using the dating service	5	+	+		+	+		-	+
A.6 Giving medical advice	4	+	-		-	+		+	+
A.7 Be careful with medicine	1	-	+		-	-		-	-

Etc.... (to include all tasks listed in the Appendix of Norris, Brown, Hudson, and Yoshioka, 1998, pp. 151-226)

Table 7 shows the 13 tasks we selected for our prototype performance tests and the difficulty estimates for each task, as well as whether they were predominantly aural or visual. Any disagreements between the two ESL teachers on the difficulty estimates for these 13 tasks had been resolved through discussion and adjustments to the tasks. Notice that only tasks which received two pluses or two minuses for each of the three processing factors were used. We did this to increase the probability that, if our theoretical difficulty estimates did translate into actual differences in performance difficulty among examinees, these differences would be detectable.

Table 7
Difficulty Matrix for Tasks on Three Experimental ALP Forms J, P, and Q

Component:			**CC**	**>>>>>>**		**CO**	**>>>>>>**		**CA**	**>>>>>>**
Characteristic:		**Diff. Index**	**Range**	**In/Out Sources**		**In/Out Org.**	**In/Out Avail.**		**Mode**	**Resp. Level**
Tasks **AURAL**	**VISUAL**									
A9 15 min	**F9** 15 min	**6**	+	+		+	+		+	+
E21 15 min	**F7** 10 min	**4**	+	+		+	+		-	-
B20 10 min	**B20** 10 min	**4**	+	+		-	-		+	+
E20 5 min	**E22** 10 min	**4**	-	-		+	+		+	+
A20 10 min	**C14** 10 min	**2**	+	+		-	-		-	-
F5 10 min	**A18** 10 min	**2**	-	-		+	+		-	-
C15 5 min	**A21** 5 min	**2**	-	-		-	-		+	+
XXX*	XXX	**0**	-	-		-	-		-	-

*Note: no zero-level tasks were included for this phase of the investigation.

Table 8 shows the distribution of the tasks across the three ALP test forms (P, Q, & J) with the combination of factors in the first column, the difficulty level in the second column and the tasks on each of the three forms in the columns that follow (with the task number and primary input for each). Notice that we counterbalanced the tasks by estimated difficulty as well as with input that was predominantly aural versus visual on all three forms.

Table 8
Task Sampling for ALP Forms P, Q, and J

Components	Difficulty	Form P	primary input	Form Q	primary input	Form J	primary input
CC, CO, & CA	6	A9	Aural	F9	visual	F9	visual
CC & CO	4	F7	visual	E21	aural	F7	visual
CC & CA	4	B20	mixed	B20	mixed	B20	mixed
CO & CA	4	E22	visual	E20	aural	E20	aural
CC	2	A20	aural	C14	visual	A20	aural
CO	2	A18	visual	F5	aural	F5	aural
CA	2	C15	Aural	A21	visual	A21	visual

Stage 3: Rating scales. We developed two different types of scales for rating examinees' performances in this project: task-dependent and task-independent.

First, our approach to developing the *task-dependent rating scales* was to remove ourselves from the process as much as possible in order to simulate the conditions under which such scales might be developed in actual language programs. In short, we hired three potential stakeholders for this kind of assessment to act as informants about the kinds of criteria that should be applied in judging whether or not a task had been successfully accomplished: one ESL teacher, one advanced ESL learner, and one non-ESL teacher (who had considerable experience with international students). This task accomplishment criteria team worked through several stages, both individually and as a group, and eventually negotiated amongst themselves what would constitute the criteria by which each individual assessment task would be judged in the real world (see Norris, forthcoming for considerably more details). These were then transformed into rating rubrics like the ones shown in Tables 9a and 9b, and such rubrics were to be used in judging an examinees' performance on each particular task that appeared on the test.

Table 9a

Example Task-Dependent Rating Scale for Task B20

B20	inadequate		able		adept
descriptors	Examinee chooses the wrong hotel, OR examinee writes the fax in a manner that would cause serious confusion on the part of the boss concerning which hotel to use OR examinee writes the fax in a pragmatically inappropriate manner, which would result in future difficulties for examinee's relationship with boss.	Examinee performance contains some elements from the *inadequate* descriptor and some elements from the *able* descriptor.	Examinee produces a fax message recommending the Plaza Inn and provides some form of correct rationale for the choice (based on the parameters set by the boss, that is, distance, pool availability, and price). An *able* performance will not necessarily list the exact hotel specifications from the hotel brochure for the Plaza In (that is, examinee need not give exact distance and price).	Examinee performance contains some elements from the *able* descriptor and some elements from the *adept* descriptor.	Examinee's fax message recommends the Plaza Inn and provides appropriate rationale (but does not necessarily list exact hotel specifications). Examinee produces a pragmatically and stylistically appropriate fax message (demonstrating understanding of relationship relative to that set by boss on the answering machine message).
Rating	**1**	**2**	**3**	**4**	**5**

Table 9b

Example Task-Dependent Rating Scale for Task F05

F05	inadequate		able		adept
descriptors	Examinee incorrectly fills out change of address form such that any essential elements (listed in the *able* descriptor) are not processable by the post office (this might include illegibility, incorrect placement of information, absence of information, etc.	Examinee performance contains some elements from the *inadequate* descriptor and some elements from the *able* descriptor.	Examinee fills out change of address form according to information given by John, minimally including with correct spelling and correct locations (see form for details) --name --new address --old address --starting date --signature and printed name (either John Harris or examinee's own name).	Examinee performance contains some elements from the *able* descriptor and some elements from the *adept* descriptor.	Examinee correclty fills out change of address form with ALL applicable information given by John on the answering machine message (see form for details).
Rating	**1**	**2**	**3**	**4**	**5**

Second, our approach to developing the *task-independent rating scale* was to use our original task difficulty estimation procedures for generalizing across performances. Thus the task-independent rating scale was designed to help raters estimate each student's general level of language performance across a variety of tasks that in turn involved a range of abilities in code command, cognitive operations, and communicative adaptation (see Table 10).

Table 10
Task-Independent Rating Scale

Task-Independent Performance Components	Holistic Performance Rating				
	inadequate		able		adept
CODE COMMAND	1	2	3	4	5
justification					
	inadequate		able		adept
COGNITIVE OPERATIONS	1	2	3	4	5
justification					
	inadequate		able		adept
COMMUNICATIVE ADAPTATION	1	2	3	4	5
justification					

Such task-independent scales were intended to be used to make judgments about an examinees' overall abilities with respect to these processing factors after a rater had observed an examinees' performances on a full set of test tasks.

The components of task-independent performance were described as follows in the rater instructions for this task-independent scale:

1. *Code Command:* For this component, consider the performance of the student in terms of the linguistic code relevant to the tasks found on the ALP. You should bear in mind not only the manifestations of linguistic code apparent in student productive responses, but you should also consider the qualities of linguistic code found in the input on various tasks (which must be received and processed by the student). Under the concept of code should be understood the structure of the language relevant to the tasks, including: vocabulary, morphology, and syntax, as well as pragmatics, non-verbal communication, etc. To what extent is the student in command of the code necessary for accomplishing tasks like those found on the ALP?
2. *Cognitive Operations:* For this component, consider the performance of the student in terms of the operations required by tasks found on the ALP. Once again, you should bear in mind receptive as well as productive reflections of such operations. Cognitive operation should be understood to involve the manipulation of task elements towards the accomplishment of the task, and includes: accessing appropriate information, organizing or re-organizing information, handling multiple stages within tasks, completion of necessary aspects of tasks, etc. To what extent is the student capable of executing the cognitive operations necessary for accomplishing tasks like those found on the ALP?
3. *Communicative Adaptation:* For this component, consider the performance of the student in response to the range of communicative demands made by tasks found on the ALP. Obviously, such demands occur in both receptive and productive directions when utilizing the language. Communicative adaptation should be understood to involve a student's capacity to marshal and utilize linguistic and cognitive resources in appropriate ways across a range of communicative demands found in tasks, including: time constraints, multi-skill requirements (e.g., production as well as reception of varying sorts), task-imposed stress, etc. To what extent is the student capable of adapting to the range of communicative movements necessary for accomplishing tasks like those found on the ALP?

Raters were thus expected to assign task-independent ratings for each of the three performance components (code command, cognitive operation, and communicative

adaptation) based on their overall impressions of each student's performances on the full set of ALP tasks. These ratings were meant to reflect the processes by which examinees attempted tasks as well as the language they produced. However, the ratings were not supposed to be based on the number of tasks successfully accomplished, or the task-dependent ratings for individual tasks. Instead, the ratings on the task-independent scale were supposed to represent the raters' overall perceptions of the examinees' abilities to perform language tasks like those found on the ALP. Raters were told to assign scores from one to five as follows:

1. *Inadequate:* A rating of inadequate indicates that the student seems generally incapable of coming to terms with the particular performance component (code, cognitive, communicative) on tasks like those found on the ALP.
2. Student performance contains some elements from the *inadequate* descriptor and some elements from the *able* descriptor.
3. *Able:* A rating of able indicates that the student seems generally capable of coming to terms with the particular performance component on tasks like those found on the ALP.
4. Student performance contains some elements from the *able* descriptor and some elements from the *adept* descriptor.
5. *Adept:* A rating of adept indicates that the student seems quite capable of coming to terms with the particular performance component on tasks like those found on the ALP; additionally, the student seems to have little to no difficulty in accomplishing such tasks in terms of the component.

Stage 4: Self-rating scale. Immediately after completing the last of the performance tasks on the ALP test, participants were required to complete self-rating sheets (see Table 11). These sheets contained three questions with space for self-ratings on a scale of one to three for each of three questions on each of seven tasks (again, see Table 11). Self-ratings were solicited in order to provide another perspective on the perceived difficulty of tasks, examinees' impressions about task performance and accomplishment, and the relationship between familiarity and task success.

We found that some examinees had trouble combining the two ideas represented in the first question (about task familiarity), perhaps because of the wide range of informational and task-oriented demands across the various ALP tasks. Hence in the future, we would probably divide the first question into two questions: (a) How familiar were you with the information on the task, and (b) How familiar were you with what you were asked to do on the task.

Table 11
Example Self-rating Sheet

==========

ALP Post-test Questionnaire

ALP form: ID#:

How familiar are you with the different items on this test (how well do you know them; have you done them before)?

	Very familiar	Somewhat familiar	Not familiar
Item 1	3	2	1
Item 2	3	2	1
Item 3	3	2	1
Item 4	3	2	1
Item 5	3	2	1
Item 6	3	2	1
Item 7	3	2	1

How well did you do on the different items?

	I did very well	I did okay	I did not do well
Item 1	3	2	1
Item 2	3	2	1
Item 3	3	2	1
Item 4	3	2	1
Item 5	3	2	1
Item 6	3	2	1
Item 7	3	2	1

How easy or difficult were the different items?

	Easy to do	Possible, but not easy	Difficult to do
Item 1	3	2	1
Item 2	3	2	1
Item 3	3	2	1
Item 4	3	2	1
Item 5	3	2	1
Item 6	3	2	1
Item 7	3	2	1

Procedures

The procedures for administering the ALP in Hawai'i and Japan were exactly the same for all three forms. We followed four basic steps: preparing to administer the ALP, administering the ALP, interviewing after the ALP, and scoring the ALP.

Preparing to administer the ALP. The ALP directions asked the proctors to:

1. Check the general equipment and supplies necessary for administering the overall test and make sure that they were present and in working order: video camera, tape player, tape recorder, pencils, English-language dictionary, task prompts, and realia.
2. Check the realia for each task and make sure that everything listed in the task checklists was present (also make sure there were adequate numbers of copies of those realia that the examinees would end up writing on, taking notes on, etc.).
3. Read through all of the task administration guidelines.
4. Set up a testing center. Two tables were needed: (a) one table for the tape-playing and -recording machines (on either side of where the student sits), the dictionary, pencils (in order to erase mistakes, etc.), and anything else the student would need, and plenty of space for placing and working with realia, and (b) a second table for spreading out the realia necessary for each of the tasks on the test, as well as the other forms that would have to be completed (all in chronological order according to the test). The video camera would also need to be set up unobtrusively in a corner, focused on the student's position.
5. Once the testing center was ready, run a pilot test on a volunteer who would provide feedback. Work through the entire test-administration process so as to be prepared for the actual administrations.
6. Be prepared to keep notes on the test sessions. Notes were to be numbered with the student ID numbers, and data were to be collected on any observations the proctor might have about the administration, as well as: the gender of student, time on task and time on test, number of listenings (when there was taped input), any problems that occurred during the administration.

Administering the ALP. In order to administer the ALP, the proctors received the following guidelines:

1. Make sure that you already have audio and videotapes labeled (with the form, date, and student ID #) and ready in their machines before examinees arrive.
2. Have the examinees fill out the AGREEMENT TO PARTICIPATE and BACKGROUND INFORMATION forms.
3. Start the video recorder rolling while they are completing the forms in step 2.
4. Give the student their copy of the test form and turn to the instructions page. Read through these together. Make sure that they understand what's going to happen and that they should ask if they do not understand something in the instructions;

the time is not strictly enforced, but they should try to complete tasks within the suggested times. If they do not have any questions, proceed to the next page (or answer questions, then proceed).

5. Work through each of the seven tasks, following the guidelines. Each task has distinct administration procedures, depending on the realia and task involved. Make sure to keep anything the examinees have written on (in order, attached to their background information sheet) and be sure you have recorded all tasks that need audiorecording.
6. If examinees look completely lost, try to encourage them to move ahead with the task up to a point where they finish it or it becomes obvious that they won't be able to. Prompt them (when they've reached the suggested allotted time for the task) by telling them that they have a few minutes left to finish the task. Stop them if they are not progressing or are using excessive amounts of time.
7. Tell the examinees when they have three tasks left on the test, etc. (this seems to keep them going to see the end of the road).

Interviewing after the ALP. After the ALP administration, proctors helped the examinees work through the self-ratings. The proctor provided the student with the self-rating sheet (shown in Table 11). The student was then told to look at the task prompt pages for each task and was reminded with a few short phrases about the task. They were then asked by the proctor to think about each question and rate the task on a three-point scale (see Table 11). In order to minimize confusion, the proctor actually circled the self-ratings for the examinees. Examinees were told to pick only one point on the scale. The whole process took about five minutes per examinee, although in some cases it took longer when examinees had a great deal to say about their performances (responses were also tape-recorded).

Scoring the ALP. The ALP task-dependent and task-independent ratings were completed by three university-level ESL teachers hired for the purpose. These teachers took part in quite minimal rater training, which consisted mostly of getting to know the tasks and the rubrics written by the criteria team, but did not involve rater norming. We decided to minimize training in order to simulate real-world constraints that are often placed on the teachers who have to implement these kinds of assessments. We also wanted to investigate just how the raters interpreted and utilized the scales and rubrics (see Tables 9a, 9b, & 10). After training, the raters applied both the task-dependent and task-independent rating scales in judging all of the products collected from examinees'

performances (including recordings, written products, notes, etc.). The examinees' final scores for each of the two rating scales were then based on the average of the three raters' judgments.

RESULTS

This **RESULTS** section will serve as a technical report of the descriptive statistics, reliability statistics, correlational statistics, and implicational statistics found in the current phase of the project. The **DISCUSSION** section, which comes next, will explore these results in more lay terms by showing how they are directly related to the research questions posed at the beginning of this paper.

Descriptive Statistics

The descriptive statistics in this study were examined from three perspectives: task-dependent ratings for each task, task-independent ratings for each difficulty factor, and task-independent ratings for each form.

Table 12
Descriptive Statistics: Task-Dependent Ratings by Task

Task	*N*	Mean	*SD*	Min	Max	Skew
E20	29	3.56	0.99	1.67	5.00	-0.31
A21	29	2.86	1.06	1.00	5.00	0.38
B20	48	2.92	1.30	1.00	5.00	0.00
F05	29	2.66	1.62	1.00	5.00	0.41
F09	26	1.63	1.14	1.00	5.00	2.02
E21	16	2.85	1.31	1.00	5.00	0.10
C14	16	2.04	1.25	1.00	5.00	1.40
E22	19	2.40	1.65	1.00	5.00	0.66
C15	19	3.25	1.19	1.33	5.00	-0.07
A18	19	3.02	1.61	1.00	5.00	-0.22
A09	17	2.18	0.98	1.00	4.00	0.66
F07	27	2.31	1.45	1.00	5.00	0.72
A20	27	3.56	1.24	1.00	5.00	-0.47

The *task-dependent ratings for each task* are presented in Table 12. Notice that, because of the small numbers of examinees involved in this project, the results for forms P, Q, and J were combined wherever possible (i.e., statistics are reported on all examinees who completed each task). Hence, different numbers (*N*) of examinees are represented for each question, ranging from 17 examinees on task A09 to 48 on task B20 (the anchor task which appeared on all three forms of the ALP). The mean task-dependent performance

ratings on the different tasks ranged considerably from 1.63 for task F09 to 3.56 for tasks A20 and E20. This was our first indication that some tasks may have been considerably easier or more difficult for the examinees. The standard deviations also differed considerably across tasks from .98 to 1.62 indicating that some tasks spread the examinees out more than others. The minimum and maximum statistics suggest that all the tasks except E20 and A09 were utilizing the entire range of possible scores from 1 to 5. Finally, the skew statistics in the column furthest to the right indicate that the distributions for tasks F09 and C14 may have been considerably skewed in a positive direction, a further indication that performances on these tasks were generally rated much lower than for other tasks.

Table 13
Descriptive Statistics: Task-Independent Ratings by Subcomponents

Holistic Category	*N*	Mean	*SD*	Min	Max	Skew
Code Command	48	2.63	1.05	1.00	5.00	0.57
Cognitive Operations	48	2.76	1.09	1.33	5.00	0.88
Communicative Adaptation	48	2.62	1.12	1.00	5.00	0.59

Descriptive statistics are shown in Table 13 for the *task-independent ratings for each difficulty factor*. These statistics indicate that all 48 examinees were rated according to the three factors, they are all reasonably well-centered, all have a standard deviation of about 1.00, all utilize the full range from 1 to 5, and are not particularly skewed (i.e., have skew statistics of less than 1.00).

Table 14 shows the descriptive statistics for the *task-independent ratings for each form*. Naturally, the *N*-sizes reflect the number of examinees who took each form. The mean for form J (the EFL sample) is lower than both forms P and Q (the ESL sample) and the standard deviation, minimum, and maximum statistics all indicate that the examinees taking form J performed in a considerably more homogeneous manner than those taking forms P and Q. Finally, the skew statistic indicates no skewing in the distributions when the scores are broken down by form.

Table 14
Descriptive Statistics: Task Independent by Test Forms

Form	*N*	Mean	*SD*	Min	Max	Skew
P	19	2.93	1.05	1.42	4.43	0.12
Q	19	2.88	0.83	1.67	4.81	0.80
J	10	1.99	0.46	1.38	2.76	0.58

Reliability Statistics

Reliability, or the degree to which a test is measuring consistently, was studied from several perspectives. Notice in the first three columns of numbers in Table 15 that correlation coefficients were calculated for each possible pair of raters using each of the component scores (CC, CO, & CA) and task scores. With the exception of raters 1 and 2 on task A09 where the correlation was a relatively low .39, the remaining correlation coefficients range from moderate to very high (i.e., .60 to .99).

Table 15
Interrater Reliability Statistics

Task	Pearson R_{xy}			Spearman Brown $R_{xx'}$		Intraclass	Rater agreement within 1 point		
	r1/r2	r1/r3	r2/r3	Based on average R_{xy}	Based on lowest R_{xy}		r1/r2	r1/r3	r2/r3
CC	.76	.78	.81	.91	.90	.94*	.92	.88	.75
CO	.78	.65	.84	.91	.85	.90*	.88	.90	.92
CA	.66	.64	.86	.90	.84	.88*	.85	.85	.96
E20	.76	.83	.85	.93	.91	.93	1.0	1.0	1.0
A21	.66	.84	.78	.91	.85	.90*	.76	.97	.86
B20	.83	.65	.66	.89	.85	.88	.92	.83	.83
F05	.76	.87	.78	.93	.90	.92	.86	.93	.86
F09	.91	.79	.89	.95	.92	.95*	1.0	.88	.96
E21	.61	.68	.94	.92	.82	.90	.81	.81	1.0
C14	.70	.84	.86	.93	.88	.92*	.88	.94	.94
E22	.91	.88	.91	.96	.96	.96	.89	.84	.95
C15	.84	.98	.82	.97	.93	.96	.95	.89	1.0
A18	.95	.77	.74	.94	.90	.93	1.0	.89	.84
A09	.39	.66	.82	.85	.66	.83	.76	.94	.94
F07	.68	.60	.62	.82	.75	.82	.85	.81	.85
A20	.99	.89	.90	.98	.96	.97	1.0	.96	.96

*significant *F* for between raters means comparisons (at $p < .05$)

In the fourth column of numbers an estimate is given based on the Spearman-Brown prophecy formula using the average (based on the Fisher *z* transformation) of the three correlation coefficients discussed in the previous paragraph. This is an estimate of the reliability with which an average rating on a performance can be interpreted when the three raters' scores are taken together. Notice these three-rater reliability estimates range from .82 to .98, indicating that the average ratings for performances on the 13 individual tasks, and on the three task-independent factors, ranged from fairly high to very high in reliability.

The fifth column presents the same three-rater information, but the Spearman-Brown prophecy formula was applied to the lowest of the three correlation coefficients for each task instead of to the average of the three. These more conservatively slanted estimates of the reliability of performance ratings on the tasks (when the three raters' scores are taken together) range from .75 to .96, indicating again that performance ratings for the individual tasks and the three performance factors ranged from moderately high to very high in reliability.

The column of intraclass correlations shows results similar to those in the previous two columns, but adds the information that ratings on all three of the difficulty factors (CC, CO, & CD) and on three of the tasks produced means that were significantly different (i.e., those with asterisks). Thus, even though these six scales are reliable in the sense that they are producing scores that are very similar in the ways they rank the examinees, it appears that at least one of the raters is significantly higher or lower than the others on these six.

Finally, the last three columns in Table 15 provide agreement coefficients for each of the pairs of raters. These agreement coefficients represent the percent of agreement exhibited by raters in judging examinee performances within one point on each of the scales. Thus, the .92 value for raters 1 and 2 on CC indicates that they agreed within one point of each other 92% of the time. Generally, the degree of agreement was very high, though it ranged from a high of 1.00 to a low of .76 across the various ratings.

Correlational Statistics

In calculating Pearson product-moment correlation coefficients for this study, the minimum pairwise comparison for form P was $N = 17$; for form Q, it was $N = 16$; and for Form J, it was $N = 10$. Because of the small sample on form J, very few of the coefficients were found to be statistically significant; hence, most of the correlation coefficients could not be interpreted as representing non-probabilistic fluctuations from a correlation of zero (although the observed associations might hold across larger sample

sizes). As a consequence, the correlational analyses reported here will focus on forms P and Q. In future research, we plan to gather more data on form J as well as on forms P and Q. We assume that the current trends will be similar to what we have found here, although clearer and more consistently interpretable.

Table 16 presents the correlation coefficients for forms P and Q separately for various combinations of the following categories of scores:

1. The overall ratings for code command (CC), cognitive operations (CO), and communicative adaptation (CA),
2. Examinees' overall self-ratings for familiarity with the tasks (FAM), task performance (PERF), and task difficulty (DIFF), and
3. The total task-dependent scores on each form.

As shown in Table 16, the intercorrelations for the task-independent factors on Form P are all high with .93 for CC with CO, .97 for CC with CA, and .94, CO with CA. Similarly high levels of association were observed on form Q, with correlations of .87 for CC with CO, .91 for CC with CA, and .96 for CO with CA. In fact, as shown in Table

Table 16
Intercorrelations for Task-independent Subscores and Self-ratings

	CC	CO	CA	FAMIL	PERF	DIFF
Form P						
CC	1.00					
CO	.93*	1.00				
CA	.97*	.94*	1.00			
FAMIL	.75*	.61*	.74*	1.00		
PERF	.78*	.74*	.78*	.83*	1.00	
DIFF	.70*	.67*	.68*	.79*	.91*	1.00
P TOTAL	.97*	.94*	.94*	.70*	.77*	.70*
Form Q						
CC	1.00					
CO	.87*	1.00				
CA	.91*	.96*	1.00			
FAMIL	.64*	.74*	.78*	1.00		
PERF	.75*	.74*	.79*	.73*	1.00	
DIFF	.66*	.64*	.67*	.73*	.83*	1.00
Q TOTAL	.90*	.95*	.96*	.72*	.75*	.61*
Form J						
CC	1.00					
CO	.34	1.00				
CA	.78*	.74*	1.00			
FAMIL	.27	-.08	-.06	1.00		
PERF	.56	.53	.57	.21	1.00	
DIFF	.37	.54	.61	.45	.56	1.00
J TOTAL	.62	.83*	.90*	-.15	.58	.49

* $p < .01$ (one-tailed)

16, all of the correlation coefficients for the task-independent factors and the three self-ratings are statistically significant and reasonably high, indicating that the subscores for CC, CO, and CA are all related to each other and to the self- ratings for familiarity, performance, and difficulty—all of which supports to some degree the validity of basing interpretations on the task-independent and self-ratings scores from performances on forms P and Q.

Table 17 presents the correlations for each of the task-dependent ratings on forms P and Q and the same overall task-independent ratings and self-ratings described for Table 16. With the exception of tasks A09 and A18, all other correlation coefficients for form P are statistically significant and at least moderately high. These results lend support to the validity of basing interpretations about examinee performance ability on ratings from tasks B20, E22, C15, F7, and A20 on form P, and to a lesser degree to task A18.

Table 17
Intercorrelations for Task-dependent Subscores and Self-ratings

	CC	CO	CA	FAMIL	PERF	DIFF
Form P						
B20	.68*	.65*	.67*	.59*	.55*	.60*
E22	.78*	.86*	.79*	.55*	.61*	.58*
C15	.76*	.71*	.75*	.56*	.61*	.55*
A18	.74*	.74*	.74*	.48	.61*	.52
A09	.49	.49	.53	.35	.24	.29
F07	.92*	.86*	.83*	.57*	.68*	.57*
A20	.86*	.77*	.79*	.60*	.72*	.59*
P TOTAL	.97*	.94*	.94*	.70*	.77*	.70*
Form Q						
B20	.68*	.67*	.76*	.52	.40	.37
E20	.85*	.67*	.77*	.34	.54*	.45
A21	.83*	.78*	.80*	.58*	.77*	.71*
F05	.36	.41	.45	.27	.55*	.35
F09	.70*	.89*	.83*	.72*	.52	.31
E21	.80*	.80*	.75*	.69*	.67*	.53
C14	.48	.55	.53	.51	.22	.22
Q TOTAL	.90*	.95*	.96*	.72*	.75*	.61*
Form J						
E20	.73*	.24	.57	-.02	-.04	-.04
A21	.74*	.70	.87*	.26	.48	.61
B20	.22	.34	.58	-.08	.06	.58
F05	.32	.65	.50	-.45	.53	.07
F09	.51	.08	.61	-.03	.41	.60
F07	-.02	.45	.16	-.39	-.03	-.28
A20	.34	.76*	.62	.01	.71	.47
J TOTAL	.62	.83*	.90*	-.15	.58	.49

* $p < .01$ (one-tailed)

For form Q, the picture is a little less clear. With the exception of tasks F05 and C14, the correlation coefficients of tasks with task-independent ratings for form Q are statistically significant and at least moderately high. In addition, eight of the correlations between individual tasks and the three categories of self-ratings were statistically significant. These two sets of results taken together lend support to the validity of basing interpretations of examinee performance ability on ratings from task A21 and to a lesser degree from tasks B20, E20, A21, F05, F09, and E21 on form Q.

The correlation coefficients shown in Table 18 show the degree of relationship between task-dependent ratings for each of the 13 tasks and the total scores for all tasks taken together. They are reported separately for forms P and Q. These coefficients indicate the degree to which each task discriminates among ability levels in the same way as the total scores (which are presumably a better estimate of the examinees overall abilities than is any single task). All of these coefficients are statistically significant and moderate to high, except A09 on form P and F05 on form Q. These results suggest that we might want to further evaluate and revise these two tasks for future versions of our tests.

Table 18
Correlations for Task-dependent Scores and Totals

Form P (minimum pairwise *N* = 17)	
B20	.7307*
E22	.8448*
C15	.7386*
A18	.7524*
A09	.5398
F07	.9347*
A20	.8824*
Form Q (minimum pairwise *N* = 16)	
B20	.6661*
E20	.7390*
A21	.8217*
F05	.5181
F09	.8466*
E21	.8108*
C14	.5988*
Form J (minimum pairwise *N* = 16)	
E20	.4507
A21	.8012*
B20	.5346
F05	.6846
F09	.3552
F07	.4537
A20	.8361*

*$p < .01$ (one-tailed)

Implicational Statistics

Implicational scaling is a statistical technique that allows investigation of the degree to which data form a hierarchical, or implicational, scale. The data from Form P of the ALP in Table 19 will serve as an example. Notice that tasks are labeled across the top, one in each column, and persons are labeled down the left side, one for each row. Note also that the table itself is filled with ones and zeros. A one indicates a person who passed a task at a particular criterion level, and a zero indicates a person who failed. In this particular table, 60% (or a rating of 3 on the 5-point scale) was used as the cut-point for passing (a rating of *3* was also the point at which task performances were described by criteria informants as having the minimal elements necessary for being considered successfully accomplished). Thus, in this table the ones represent tasks that examinees passed at with a rating of *3*, while the zeros indicate tasks on which they received a rating below *3*.

Table 19
Example Implicational Scale - Form P

FORM P:	A20	B20	A18	C15	F07	E22	A09	TOTAL
9	1	1	1	1	1	1	1	7
19	1	1	1	1	1	1	1	7
11	1	1	0	1	1	1	1	6
16	1	1	1	1	1	0	1	6
18	1	1	1	1	1	1	0	6
36	1	1	1	1	1	1	0	6
41	1	1	1	1	1	1	0	6
10	1	1	1	0	1	1	0	5
14	1	0	1	1	1	0	0	4
21	1	1	1	0	0	0	0	3
24	1	1	0	1	0	0	0	3
12	1	0	1	0	0	0	0	2
23	1	0	1	0	0	0	0	2
28	1	0	1	0	0	0	0	2
31	1	0	0	1	0	0	0	2
25	0	1	0	0	0	0	0	1
32	0	1	0	0	0	0	0	1
38	0	0	0	1	0	0	0	1
39	0	0	0	0	0	0	0	0
SUM	15	12	12	11	9	7	4	
DIFF	2	4	2	2	4	4	6	
ERRORS	0	3	5	4	0	1	0	13
POSSIBLE								133
CR	0.9023							
p	0.7900	0.63	0.63	0.58	0.47	0.37	0.21	
q	0.2100	0.37	0.37	0.42	0.53	0.63	0.79	
MMR	0.6541							
PI	0.2481							
CS	0.7174							

Notice also that the ones in each row have been added up and the results put in the column furthest to the right labeled TOTAL, and the ones in each column have been

added up and the results put in a row near the bottom labeled SUM. Next the rows were ordered from the highest TOTAL at the top to the lowest TOTAL at the bottom, and the columns were sorted from the highest SUM on the left to lowest SUM on the right. Notice that, after this sorting, the ones tend to be to the left and up, while the zeros tend to be to the right and down. Theoretically, this is the general shape that a hierarchical scale should produce. If there is a true hierarchical scale, all examinees who passed only one task would have passed task A20, and any examinee who passed just two tasks would have passed A20 and B20, and so forth. From the other end of the scale, we would expect any examinees who passed A9 to pass all and only the tasks to the left of it, and all those who passed E22 to pass all and only the tasks to the left of it. In short, the pattern of ones should form a line like the one drawn in Table 19 that more or less forms a triangle of ones. The problem in real-world data is that there will turn out to be errors, that is, zeros in the field of ones and ones in the field of zeros. Clearly, such cases show up in Table 19, where zeros are found to the left and above the line and ones to the right and below it. In each column, in a row labeled ERRORS, we counted the number of ones and zeros that did not fit the pattern.

One question that implicational scaling allowed us to address was whether the number of such errors was so high that they threaten the integrity of our task performance ratings or not. Two statistics have been developed which help with this interpretation:

1. The *coefficient of reproducibility* (*CR*), shown toward the bottom left of the table, tells us the percentage of adherence to a scale, or the percentage of ones and zeros that are not errors. In Table 19, the *CR* turned out to be .90225564, so about 90.22% of the ones and zeros were within the pattern that we are calling a hierarchical scale. Looked at another way, about 9.78% of the ones and zeros were errors, or outside of the scale. Typically, the *CR* should be .90 or higher to be interpreted as a positive indication of a scale (Guttman, 1944, 1950).
2. The *coefficient of scalability* (*CS*), shown at the very bottom left of the table, tells us the degree to which the data indicate a progression that is scale-like. A minimum *CS* of .60 is necessary to interpret a scale as implicational (Dunn-Rankin, 1983, p. 107). In Table 19, the *CS* turned out to be .7174, so that scale for Form P can be considered implicational.

If both the *CR* and *CS* reach their respective .90 and .60 minimums, the scale can be considered implicational, which was the case in Table 19 for performance ratings on Form P of the ALP.

Table 20
Example Implicational Scale - Form Q

FORM Q:	E20	B20	A21	F05	E21	F09	C14	TOTAL
5	1	1	1	1	1	1	1	7
8	1	1	1	1	1	1	0	6
20	1	1	1	1	1	1	0	6
2	1	1	0	1	1	1	0	5
27	1	1	1	0	1	0	1	5
40	1	1	1	0	1	0	1	5
15	1	0	1	1	1	0	0	4
29	1	1	1	1	0	0	0	4
37	1	0	1	1	1	0	0	4
1	1	1	1	0	0	0	0	3
3	1	1	1	0	0	0	0	3
13	1	0	1	1	0	0	0	3
22	1	1	0	1	0	0	0	3
26	1	1	0	1	0	0	0	3
17	1	1	0	0	0	0	0	2
33	0	0	0	1	0	0	0	1
35	1	0	0	0	0	0	0	1
30	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0
SUM	16	12	11	11	8	4	3	
DIFF	4	4	2	2	4	6	2	
ERRORS	1	3	1	7	1	0	2	15
POSSIBLE								133
CR	0.8872							
p	0.8400	0.63	0.58	0.58	0.42	0.21	0.16	
Q	0.1600	0.37	0.42	0.42	0.58	0.79	0.84	
MMR	0.6917							
PI	0.1955							
CS	0.6341							

Tables 20 and 21 show further example scales at the 60% cut-point (rating of "3") for Forms Q and J respectively. Notice that the *CS* for both forms reaches the .60 minimum and that the *CR* for Form J reaches .90, but that the *CR* for Form Q does not.

Table 22 summarizes the implicational scaling statistics for forms P, Q, and J, using 40%, 60%, and 80% cut-points. Notice the *CR* statistics for all three cut-points on form P reach the .90 level and all three *CS* statistics reach .60. The same pattern emerges for form J. However, none of the *CR* statistics reach .90 for form Q, and only two of the *CS* statistics reach .60. These results support the hierarchical structure of the tasks in forms P and J, and thus their validity for informing interpretations about examinee abilities to accomplish tasks like those found on the ALP test. Thus, it seems that an examinee who scores high on the test overall will accomplish most of the tasks, while an examinee who

Table 21
Example Implicational Scale - Form J

FORM J:	E20	A20	A21	B20	F05	F09	F07	TOTAL
5J	1	1	1	1	0	0	0	4
1J	1	1	1	0	0	0	0	3
4J	1	1	0	0	1	0	0	3
6J	1	1	1	0	0	0	0	3
2J	0	1	0	0	0	0	0	1
3J	1	0	0	0	0	0	0	1
7J	0	0	0	1	0	0	0	1
8J	1	0	0	0	0	0	0	1
9J	0	0	0	0	0	0	0	0
10J	0	0	0	0	0	0	0	0
SUM	6	5	3	2	1	0	0	
DIFF	4	2	2	4	2	6	4	
ERRORS	2	0	1	1	1	0	0	5
POSSIBLE								70
CR	0.9286							
p	0.6000	0.50	0.30	0.20	0.10	0.00	0.00	
q	0.4000	0.50	0.70	0.80	0.90	1.00	1.00	
MMR	0.7857							
PI	0.1429							
CS	0.6667							

Table 22
Implicational Scale Statistics All Forms

FORM / STATISTIC	******** 40%	CUT-POINT 60%	******** 80%
FORM P			
CR	0.9023	0.9023	0.9624
MMR	0.7143	0.6541	0.6767
%IMPROV	0.1880	0.2481	0.2857
CS	0.6579	0.7174	0.8837
FORM Q			
CR	0.8947	0.8872	0.8947
MMR	0.7444	0.6917	0.7218
%IMPROV	0.1504	0.1955	0.1729
CS	0.5882	0.6341	0.6216
FORM J			
CR	0.9714	0.9286	0.9857
MMR	0.8571	0.7857	0.9143
%IMPROV	0.1143	0.1429	0.0714
CS	0.8000	0.6667	0.8333

scores lower on the test overall will accomplish only the easier tasks, and so forth. Unfortunately, the same cannot be said for Form Q at this preliminary stage in the study. We intend to gather further data on this form in order to further investigate the scalability of performance ratings on the tasks found on Form Q.

DISCUSSION

How adequate are the distributions of scores for the task-dependent and task-independent ratings?

First, Table 12 indicated that the individual tasks with all data combined ranged considerably in average performance ratings, from 1.63 to 3.56, and the variance also ranged considerably as indicated by standard deviations that ranged from .98 to 1.62. However, within those parameters, we can also say that only two of the distributions for the individual tasks (F09 and C14) were somewhat positively skewed. Second, Table 13 showed that the task-independent ratings for each processing difficulty factor were all reasonably well centered, had a standard deviation of about 1.00, used the full range of possible scores from 1 to 5, and were not skewed. Finally, Table 14 revealed that the mean of the task-independent performance ratings for form J (the EFL sample) was considerably lower than for both forms P and Q (the ESL samples), and also that performance ratings for examinees taking form J were considerably more homogeneous than for examinees taking forms P and Q. However, once again, the skew statistic indicated no skewing in the distributions. Thus, though differences in average performance rating (i.e., apparent task difficulty) and dispersion did surface for different tasks, performance difficulty factor ratings, and forms, only two skewed distributions surfaced. These distributions on our initial sets of performance rating data suggest that the ALP seems to be eliciting performances which are rated at widely varying levels of task accomplishment, which is exactly what we would expect from the different general proficiency levels found in the three groups of examinee volunteers.

To what degree are the performance tests in this project reliable? And, how can the consistency of measurement be improved?

First, Table 15 revealed moderate to high consistency for the task-independent factor ratings (CC, CO, & CA) as well as for the task-dependent ratings for each of the individual tasks, whether examined through interrater correlation (adjusted for three-rater reliability), intraclass correlation, or rater agreement. All of these reliability estimates take the point of view of examining the reliability of the ratings within each component or

task, and the result is that the ratings are mostly highly reliable, or at least moderately reliable. These initial findings suggest that the raters are able to utilize the different scales with acceptable levels of consistency, especially given the fact that they did not receive any practice or norming in the use of the scales.

Naturally, given that (all other factors held constant) more test items tend to test more reliably than fewer test items, the reliability of any combinations of these items will probably be even more reliable than the individual items taken separately as they were in Table 15. Thus, when larger samples are gathered in future research, it would be interesting to examine the degree to which the reliability of the three task-independent scores (CC, CO, & CA) would be improved by combining them into an overall task-independent score. Similarly, it would be interesting to examine the degree to which the seven tasks taken together as a single ALP performance score on each of the three task-dependent forms would be more reliable than individual tasks. Indeed, with larger samples it will be possible to study these issues by using generalizability theory to investigate the relative contribution of persons and tasks to the overall consistency of the test variance. Such an analysis will even make it possible to estimate what the test consistency would have been with fewer or more tasks. Thus we will be able to report what the maximal number of tasks would be for best constructing a test that would lead to reliable interpretations about examinee performances.

To what degree can interpretations be validly based on the performance tests in this project? And, what is the relationship between examinees' performances and the predicted difficulty levels?

Relatively high correlations between the task-dependent ratings, task-independent ratings, and self-ratings provide some initial support for using average task-dependent ratings as a basis for interpretations about examinee abilities to accomplish the individual ALP tasks. In addition, initial indications also support the use of the task-independent ratings for making judgments about overall examinee abilities with the kinds of tasks represented on the ALP test. Thus, the ALP approach to performance assessment seems to elicit examinee performances on a wide range of tasks and to differentiate among examinees in terms of levels of performance on these tasks, and evaluations of these performances based on three different scales from three different perspectives all seem closely related. Implicational scale statistics provide further support for the use of the ALP tests in making judgments about examinee abilities with the tasks used in the current study. In general, it seems that when examinees can accomplish or perform well on tasks that seemed more difficult for other examinees, they can also accomplish or perform well

on all tasks that seemed easier. This was especially found to be the case with Forms P and J, where near-perfect implicational scales were found at multiple cut-points, although ratings on Form Q tasks did not result in the same degree of hierarchical structure as that found on the other two forms.

Given the low and generally unequal numbers of examinees who attempted each of the 13 ALP tasks, it is difficult at this point to evaluate the extent to which the three cognitive processing factors (which were ostensibly engaged by the tasks in differing degrees) were actually related in any systematic way to the difficulty of tasks as perceived by examinees or to consistent differences in performance ratings on the tasks. Based on both the mean ratings for each task and on the implicational scale data at varying cut-points, it does not seem at this point to be the case that tasks which engage greater combinations of the cognitive factors necessarily result in greater levels of difficulty for examinees or in lower average performance ratings.

CONCLUSIONS AND DIRECTIONS IN FUTURE RESEARCH

In conclusion, findings from this initial phase in our on-going study suggest that effective performance assessment instruments and procedures may be developed following the stages and practices described in this paper and in Norris, Brown, Hudson, and Yoshioka (1998). Prototype test tasks and instruments were developed in a way that maximally maintained fidelity with target communication tasks through the careful simulation of task characteristics and associated realia within testing conditions. Performance rating scales and procedures were developed from multiple perspectives in order to inform the various sorts of interpretations (e.g., about examinee abilities to accomplish particular tasks as well as overall abilities in performing a set of tasks) that are associated with the uses for such L2 performance assessment in many educational contexts. Finally, all test tasks and forms were administered in two different contexts with two different sets of target examinees, and the resulting performance data were rated in consistent and interpretable ways by teachers who had received virtually no training or practice in the use of the two prototypical rating scales.

It remains to be seen to what extent the sampling of L2 tasks according to the cognitive processes ostensibly engaged by variable task characteristics will prove to be a helpful addition to the development and use of the kinds of performance assessments modeled in the current project. Although it seems to be the case that several general characteristics like those identified for the current set of tasks may be associated with some consistency with particular target tasks and particular target examinee populations,

the relationship between combinations of such characteristics and actual examinee performance remains unclear. The use of such a framework for generalizing from performances observed on a small set of tasks to a broader domain of related tasks is as yet not supported.

In order to arrive at more definitive conclusions, we are currently engaged in further collection and analysis of performance data using the same instruments and procedures outlined in this paper. The next phase in this study will result in the collection of 90 examinee performances (30 examinees for each of the three ALP forms), and these performances will again be rated by three raters using the task-dependent and task-independent rating scales. Given this larger data set, additional analyses will be undertaken in order to better understand: (a) to what extent raters are able to consistently utilize the two rating scales in evaluating examinee performances; (b) to what extent performance ratings are able to consistently inform interpretations about individual examinee abilities and the average abilities of examinees sampled from different global proficiency levels; and (c) to what extent the estimates of task performance difficulty (based on the cognitive processing factors identified in the current study) may be systematically related to levels of examinee performance on the 13 ALP tasks. In addition to the analytic approaches reported for the current phase of the study, both multi-faceted Rasch model analyses using FACETS (Linacre, 1998) and multivariate analyses will be incorporated.

REFERENCES

Allaei, S. K., & Connor, U. (1991). Using performative assessment instruments with ESL student writers. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 227-240). Norwood, NJ: Ablex.

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Aschbacher, P. A. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education, 4*(4), 275-288.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University.

Berwick, R. (1989). Needs assessment in language programming: From theory to practice. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 48-62). Cambridge: Cambridge University.

Brindley, G. (1984). *Needs analysis and objectives setting in the adult migrant project.* Sydney: AMES.

Brindley, G. (1994). Task-centered assessment in language learning: The promise and the challenge. In N. Bird (Ed.), *Language and learning*. Papers presented at the Annual International Language in Education Conference (Hong Kong, 1993). (ERIC Document 386 045).

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*(1).

Brown, J. D. (1993). A comprehensive criterion-referenced testing project. In D. Douglas & C. Chapelle (Eds.) *A new decade of language testing: Collaboration and cooperation* (pp. 163-184). Ann Arbor, MI: University of Michigan.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Brown, J. D., & Hudson, T. (1998). Alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675.

Candlin, C. N. (1987). Towards task-based language learning. In C. N. Candlin & D. Murphy (Eds.), *Lancaster Practical Papers in English Language Education: Vol. 7. Language learning tasks* (pp. 5-22). Englewood Cliffs, NJ: Prentice Hall.

Clark, J. L. D., & Grognet, A. G. (1985). Development and validation of a performance-based test of ESL "survival skills." In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 89-110). Ottawa: University of Ottawa Press.

Crookes, G. V. (1986). *Task classification: A cross-disciplinary review.* Technical Report No. 4. Honolulu, Center for Second Language Research, Social Science Research Institute, University of Hawai'i at Manoa.

Dandonoli, P., & Henning, G. (1991). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals 23*(1), 11-22.

Dunn-Rankin, P. (1983). *Scaling methods.* Hillsdale, NJ: Lawrence Erlbaum.

Doyle, W. (1983). Academic work. *Review of Educational Research, 53*(2), 159-199.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

Ferris, D., & Tagg, T. (1996). Academic oral communication needs of EAP learners: What subject-matter instructors actually require. *TESOL Quarterly, 30*(1), 31-58.

Fleishman, E. A. (1978). Relating individual differences to the dimensions of human tasks. *Ergonomics, 21*(12), 1007-1019.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519- 521.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Guttman, L. (1950). The basis of scalogram analysis. In S. A. Stouffer (Ed.), *Measurement and prediction.* Princeton, NJ: Princeton University.

Hauptman, P. C., LeBlanc, R., & Wesche, M. (1985). *Second language performance testing.* Ottawa: University of Ottawa Press.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing, 13*(1), 53-61

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Honeyfield, J. (1993). Responding to task difficulty: What is involved in adjusting the relationship between learners and learning experiences? In M. L. Tickoo (Ed.), *Simplification: Theory and application* (pp. 127-138). Singapore: SEAMEO Regional Language Centre.

Huerta-Macías, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal, 5*(1), 8-11.

Jones, R. L. (1985). Second language performance testing: An overview. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 15-24). Ottawa: University of Ottawa Press.

Lee, James F. (1995). Using task-based activities to restructure class discussions. *Foreign Language Annals, 28*(3), 437-446.

Long, M. H. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 77-99). Clevedon, England: Multilingual Matters.

Long, M. H. (1989). Task, group, and task-group interactions. *University of Hawai'i Working Papers in ESL, 8*(2), 1-26.

Long, M. H. (forthcoming). *Task-based language teaching.* Oxford: Blackwell.

Long, M. H., & Crookes, G. V. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly, 26*(1), 27-56.

Long, M. H., & Crookes, G. V. (1993). Units of analysis in syllabus design. In G. Crookes & S. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice.* Clevedon: Multilingual Matters.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-75.

McNamara, T. F. (1995). Modelling performance: Opening Pandora's box. *Applied Linguistics, 16*(2), 159-179.

McNamara, T. (1996). *Measuring second language performance: A new era in language testing.* New York: Longman.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3-9, 20.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241-256.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12*, 9-15.

Norris, J. M. (1999). Identifying rating criteria for task-based EAP assessment. Unpublished ms. Honolulu, HI: University of Hawai'i at Manoa.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu, HI: University of Hawaii Press. 227 pages.

Norris, J. M., Brown, J. D., Hudson, T. (1999). *Constructing and validating second language performance assessments: Tasks, tests, and rating criteria.* Honolulu, HI: University of Hawaii Press.

Nunan, D. (1989). *Designing tasks for the communicative classroom.* Cambridge: Cambridge University.

Nunan, D. (1993). Task-based syllabus design: Selecting, grading, and sequencing tasks. In G. Crookes & S. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 55-68). Clevedon: Multilingual Matters.

Pica, T., Kanagy, R., & Falodun, T. (1993). Choosing and using communicative tasks for second language instruction and research. In G. Crookes & S. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice.* Clevedon: Multilingual Matters.

Quellmalz, E. S. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education, 4*(4), 319-331.

Scott, M. L., Stansfield, C. W., & Kenyon, D. M. (1996). Examining validity in a performance test: The listening summary translation exam (LSTE)—Spanish version. *Language Testing, 13*(1), 83-109.

Shohamy, E. (1992). Beyond performance testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal, 76*(4), 513-521.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics, 15*, 188-211.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33-42.

Stiggins, R. J. (1988). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan, 69*, 363-368.

Wesche, M. B. (1987). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing, 4*, 28-47.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Zessoules, R., & Gardner, H. (1991). Authentic assessment: Beyond the buzzword and into the classroom. In V. Perrone (Ed.), *Expanding student assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

James Dean Brown
Department of ESL
1890 East West Road
Honolulu, Hawai'i 96822

e-mail: brownj@hawaii.edu