PERFORMANCE ASSESSMENT: EXISTING LITERATURE AND DIRECTIONS FOR RESEARCH[1]

JAMES DEAN BROWN

University of Hawai'i at Manoa

ABSTRACT

The paper begins by distinguishing between task-based testing and performance testing. It then provides a short historical overview of the development of performance assessment. Next, a preliminary overview of the performance testing literature sketches out some of the general trends in that literature including the fact that that it already contains many (a) overviews and discussions of language performance testing, (b) many papers and books providing guidelines for developing and doing performance testing, and (c) numerous articles on the place of performance testing in language curriculum. The paper then restricts itself to looking at the most productive recent areas of research on performance testing with the goal of addressing the following key questions (that also serve as main headings in the paper):

1. What happens when actual performance tests are developed and validated?
2. What are the characteristics of performance tests?
3. What are the benefits of performance testing (why bother?)?
4. What factors are involved in performance task difficulty?
5. How should we best go about scoring performance tests (in terms of scoring criteria and rater effects)?
6. Are performance tests reliable?
7. Are performance tests valid?
8. Are there other issues about performance tests that need to be researched?

The conclusions section then extends each of the main questions into a number of as yet unanswered sub-questions that might prove useful for future research on performance testing.

INTRODUCTION

Task-Based vs. Performance Testing

Many researchers have defined tasks and task-based tests, and they have done so in a wide variety of ways (for fairly recent overviews of these definitions, see Norris, Brown, Hudson, & Yoshioka, 1998, pp. 32-35; Bygate, Skehan, & Swain 2001, pp. 9-10). In this paper, I will take task-based tests to be any assessments[2] that "require students to engage in some sort of behavior which stimulates, with as much fidelity as possible, goal-

[1] The research done in this paper was supported by a grant from the Center for Advanced Study of Language (CASL) at the University of Maryland, College Park, MD 20742-0025.

[2] Note that the terms *testing* and *assessment* will be used interchangeably in this paper.

oriented target language use outside the language test situation. Performances on these tasks are then evaluated according to pre-determined, real-world criterion elements (i.e., task processes and outcomes) and criterion levels (i.e., authentic standards related to task success)" (Brown, Hudson, Norris, & Bonk, 2002, p. 10 after Long & Norris, 2001). This would be what McNamara refers to as the *strong form* of performance testing (see McNamara, 1996, p. 43-45; Brown & Hudson, 2002, pp. 22-24). In task-based tests, success in performing the tasks is a central issue. As Long and Norris (2001) put it, "Task-based assessment does not simply utilize the *real-world* task as a means for eliciting particular components of the language system which are then measured or evaluated; on the contrary, the construct of interest in task-based assessment is performance of the task itself."

In contrast, performance tests will at least initially be defined here as any tests that are designed to elicit performances of the specific language behaviors that the testers wish to assess. Examples of performance assessments that are not necessarily task-based assessments are composition tasks, oral interview tasks, and so forth. They are designed to elicit students' abilities to write or speak, but they are typically scored in terms of the linguistic characteristics of the writing or speaking performances that the test designer feels are important for theoretical and/or pedagogical reasons. Unlike task-based tests, success or failure to complete the task is not necessarily at issue in performance tests.

My father, after serving five years in the US Navy during the World War Two, straightforwardly described the difference between a boat and a ship to me as follows: "a boat can fit on a ship, but a ship cannot fit on a boat." Similarly, task-based testing fits within the definition of performance testing because that definition is broader, but the definition of performance testing does not fit within the definition of task-based testing. Moreover, any discussion of performance testing will necessarily include some discussion of task-based testing, but the reverse will not necessarily be true.

Background

I must begin by acknowledging that performance assessment is not a new issue, at least in general educational measurement circles, where solid work on performance assessment has been going on for decades. This research began to appear in the sixties (e.g., Glaser, 1963; Glaser, & Klaus, 1962), and carried on into the seventies (e.g., Fitzpatrick & Morrison, 1971; Fleishman, 1978), eighties (e.g., Stiggins, 1987; Wiggins, 1989), and early nineties (e.g., Dunbar, Koretz, & Hoover, 1991; Zessoules & Gardner, 1991). As many of the references cited below will attest, this work in general educational performance assessment is ongoing and worthy of the continued attention of language testers.

I must also acknowledge the necessary links between performance testing and the ongoing SLA theory development, especially the research going on in task-based learning and syllabus design. Some of the key work in these areas includes articles and books on task-based learning (Long, 1985, 1989, 1997, 1998, forthcoming; Candlin, 1987; Nunan, 1989; Thompson, 1992; Crookes & Gass, 1993; Pica, Kanagy, & Falodun, 1993; Lee, 1995; Robinson, 1995, 2001; Skehan, 1998a & b; Willis, 1996; Skehan & Foster, 1999; Bygate, Skehan, & Swain, 2001; Ellis, 2003) as well as work on task-based syllabus design (Crookes, 1986; Long & Crookes, 1992, 1993; Nunan, 1993; Skehan, 1996; Willis, 1996; Robinson, 1996a, 2001). Certainly, research in this area will continue, and language testers must keep an eye on what is learned along the way.

Some Preliminary Observations About the Performance Testing Literature

One of the things I observed in reading through the literatures on performance testing (including language performance testing), is that this literature already contains many (a) overviews and discussions of language performance testing, (b) many papers and books providing guidelines for developing and doing performance testing, and (c) numerous articles on the place of performance testing in language curriculum.

The literature is rife with *overviews and discussions of performance testing* beginning as early as 1984 and continuing today (e.g., Skehan, 1984; Bailey, 1985; R. L. Jones, 1985; Brindley, 1994; McNamara, 1995, 1997; Shohamy, 1995; Fulcher, 1996; Tulou, & Pettigrew, 1999; Skehan, 2001). Overviews of performance testing in the educational measurement literature have also begun to appear (e.g., Aschbacher, 1991; Linn, Baker, & Dunbar, 1991; Eisner, 1999).

A number of articles have provided *guidelines for developing and doing performance testing* (e.g., Roeber, 1996; Brualdi, 1998; Anderson, 2003; Moskal, 2003). A quick search of the Amazon.com website will confirm that there are also literally hundreds of "how to" books on performance testing for various fields of study have been published in recent years (the ones that happen to be on my shelf at the moment are Herman, Aschbacher, & Winters, 1992; Angelo & Cross, 1993; Kane & Mitchell, 1996; Banks, 1997; Khattri, Reeve, & Kane, 1998). Resources also exist for teachers wishing to see examples of language performance tests (e.g., Wesche, 1987; Brown, 1998), science performance tests (e.g., Shavelson, Baxter, & Pine, 1991), and so forth.

Other work has focused on *the place of performance testing in language curriculum* (e.g., Shohamy, 1992; Brindley, 1994; Robinson, 1996b; Brown & Hudson, 1998; Norris, 2000; Long, & Norris, 2001; Candlin, 2001; Byrnes, 2002). Similarly, concerns have been expressed about the place of performance testing in the more general educational curriculum (e.g., Mehrens, 1992; Haertel, 1999).

Naturally, at the local institutional and classroom levels, any one of these areas could do with more work; however, for the field as a whole, it might be better if we turned to other areas of research that ultimately will prove more useful.

RECENT PRODUCTIVE AREAS OF RESEARCH ON PERFORMANCE TESTING

Much of the remainder of the literature on performance testing can be categorized broadly as attempting to answer certain key questions. Naturally, the approaches to these questions and the answers that were found vary considerably among these papers and books. The key questions that will serve as organizing headings in this paper are:

9. What happens when actual performance tests are developed and validated?
10. What are the characteristics of performance tests?
11. What are the benefits of performance testing (why bother?)?
12. What factors are involved in performance task difficulty?
13. How should we best go about scoring performance tests (in terms of scoring criteria and rater effects)?
14. Are performance tests reliable?
15. Are performance tests valid?

16. Are there other issues about performance tests that need to be researched?

Each of these questions will be discussed in terms of how it as been addressed in the literature to date. Then, in the **CONCLUSIONS** section, I will expand each question into a number of sub-questions that might prove useful for future research on performance testing.

What Happens When Actual Performance Tests are Developed and Validated?

A surprising number of actual language performance test development and validation projects have been reported in the literature. Some of those projects were designed to do testing in different settings or for different purposes. For example, Clark and Gognet (1985) reported on how they developed and validated a performance-based test of ESL *survival skills*. Wesche (1987) reported on the development of an integrated skills (reading, writing, listening, & speaking) English for academic purposes performance test called the *The Ontario Test of ESL*, which contained a general academic English section, plus a discipline-related thematic section (examinees could choose either science or social science). McNamara (1990) investigated the effectiveness of using item response theory to develop and validate an English-for-specific-purposes performance test for health professionals. Shameem (1998) studied the relationship between self-reported language proficiency and performance tests developed for the Indo-Fijian immigrant community in Wellington, New Zealand. North and Schneider (1998) report on the use of Rasch analysis to empirically develop and validate scale descriptors for proficiency assessment based on language performance for English, French, and German in a Swiss National Science Research Council project (see also North & Schneider, 2000).

Other performance testing projects, while also varying with regard to the settings for which they were designed, further narrowed their focus to the reading, writing, listening, speaking, and pragmatics language skills. For instance, focusing on English for academic purposes *reading*, Robinson and Ross (1996) report on the development of task-based performance test designed for placing students into reading courses in the English Language Institute at the University of Hawai‘i at Manoa.

Related to *writing*, Allaei and Connor (1991) report on developing and using performance tests for assessing ESL writing ability. Two other studies empirically investigated the sorts of writing tasks that are required in American academic degree programs, one by examining the actual writing assignments of students at a single university (Horowitz, 1986), and the other through a wide ranging and carefully designed survey study (Hale, Taylor, Bridgeman, Carson, Kroll, & Kantor, 1996).

With regard to *listening*, Scott, Stansfield, and Kenyon (1996) investigated the validity of a "summary translation" performance test of Spanish language listening ability. Stansfield, Wu, and van der Heide (2000) reported on the development and validation of a job-relevant listening summary translation performance test for selection and placement of speakers of Minnan who were employees of the U.S. Government. Brindley and Slatyer (2002) investigated the difficulty of tasks in ESL listening assessment based on actual performances of adult ESL learners in Australia. O'Sullivan, Weir, and Saville (2002) studied the validity of tasks on UCLES EFL speaking tests by comparing the results to *a priori* and *a posteriori* analyses of the actual speaking task output.

In terms of *speaking*, Stansfield and Kenyon (1992) report on the development and validation of a simulated oral proficiency interview. Douglas and Selinker (1993) studied the relationship between performances of international teaching assistants on a general speaking test and a test designed to be discipline-specific. Fulcher (1996) studied task design and group oral performance tests from the students' points of view. Kenyon (1998) investigated the validity of tasks on performance-based tests of oral proficiency of German, French, and Spanish students at the high school and college levels. Hill (1998) studied the effectiveness of validating an oral English proficiency test through test-takers reactions to and performance on a test performance test for prospective migrants to Australia. Chalhoub-Deville (2001) examined the task-based validity of three oral assessments (an oral proficiency interview, a contextualized speaking assessment, and a video/oral communication instrument). Caban (2003) examined the reliability and validity of a scale designed for rating speaking performances when it was used by 83 raters to score four L1 Japanese ESL students.

With regard to *pragmatics*, Hudson, Detmer, and Brown (1992; 1995) developed and validated a combination of traditional and performance tests of cross-cultural pragmatics for ESL students at the University of Hawai'i. Rose (1994) questioned the usefulness of discourse completion tests of pragmatics in non-Western contexts. Berwick and Ross (1996) examined the effects of cross-cultural pragmatics in oral proficiency interviews. Yamashita (1996a & b) reports on a project in which she translated and adapted the tests developed in Hudson, Detmer, and Brown (1992; 1995) for use in traditional and performance testing of the *cross-cultural pragmatics of English speaking learners of Japanese.* Yoshitake and Enochs (1996) investigated the effectiveness of the self-assessment and role play tasks developed in Hudson, Detmer, & Brown (1992; 1995) for evaluating the pragmatics ability of Japanese speakers learning EFL. Yoshitake (1997) reported a study of the effectiveness of multiple measures (again adapted from Hudson, Detmer, & Brown, 1992; 1995) of interlanguage pragmatic ability for Japanese EFL students. Roever (2001) validated a web-based test of interlanguage pragmalinguistic knowledge for ESL students. And, Brown (2000) provides an overview of the research to that date on pragmatic testing. Brown (2001) reanalyzes and compares the data produced in the work of Yamashita and Yoshitake.

With regard to *integrated skills*, in what Bachman (2002, p. 454) referred to as the "most fully conceptualized, operationalized and researched exemplification of this approach of which I am familiar," Norris, Brown, Hudson, and Yoshioka (1998) as well as Brown, Hudson, Norris, and Bonk (2002) provide two books that examine in detail the development and validation of task-based performance assessments when applied with no constraints on resources or logistics in an English for academic purposes setting. Students in Hawaii and Japan performed seven integrated-skills tasks in one of three sets (Forms P, Q, or J). The tasks were based on the needs analysis and item specifications developed in Norris, Brown, Hudson, and Yoshioka (1998). The results of this testing process and the rating of the performances using task-dependent and task-independent scales are reported in Brown, Hudson, Norris, and Bonk (2002) (see also, Brown, Norris, Hudson, & Bonk, 1999; Brown, Hudson, & Kim, 2001; Norris, Brown, Hudson, & Bonk, 2002).

One other area in which actual performance tests have been developed is in the area of computer-based performance tests. In the literature outside of applied linguistics, Swezey, Hutcheson, and Swezey (2000) demonstrate the development of second-generation computer-based team performance assessment technology. In language teaching circles, N. Jones (2001) showed how can-do statements can be used to link and equate computer-based tests across languages. Roever (2001) reported on the development and validation of web-based test of interlanguage pragmalinguistic knowledge. And, van den Branden, Depauw, and Gysen (2002) discussed the development and validation of a computerized task-based test of L2 Dutch in a vocational training setting.

Clearly then, a number of studies have been done that report on the development and validation of performance tests in different settings, or for the testing of specific skills like reading, writing, listening, speaking, and pragmatics. Other tests have been developed to test integrated skills or to examine the efficacy of using computers as a medium for performance testing.

What are the Characteristics of Performance Tests?

A number of the authors in the literature describe what they believe to be important characteristics of performance tests. Interestingly and perhaps naturally, the aspects they choose to describe differ widely. For instance, in general education literature, Quellmalz (1991) lists the following:

1. Significance [cognitive components; metacognitive components (e.g., planning, self-monitoring, evaluation, and reflection); and dispositional components (e.g., attitudes, perseverance flexibility)]
2. Fidelity [Contextual components; Complete performance assessment (full range of process)]
3. Generalizability [of a task to a range of other tasks]
4. Development appropriateness in terms of getting representative sample of achievement milestones
5. Accessibility of criteria for educators, students, parents, community
6. Utility

These are meant by the author to be characteristics of sound criteria for evaluating performances.

Also in the general education literature, Wiley and Haertel (1996, p. 75) list the following:

1. The environment, or circumstances in which the task will be performed (including physical environment, timing, tools, equipment, physical resources, and information to be made available)
2. Any communications directed to the person performing the task (including delineation of the goal and evaluation criteria, circumstances under which task will be performed, and the tools that might be used to perform the task)

These are described by the authors as criteria for task specifications to be used in designing performance tests.

Khattri, Reeve, and Kane (1998, pp. 37-41) list the following five "dimensions":

1. Time demands
2. Applied problem-solving skills demands (cognitive skills)
3. Metacognitive demands (degree to which Ss' have awareness of their own thinking and problem-solving skills)
4. Social competencies (interpersonal skills necessary for successful task completion)
5. Student control (the degree of judgment a student must exercise in defining and performing the task successfully)

These are the five "dimensions of task specification" again to be used in designing performance tests.

In the language assessment literature, Bachman and Palmer (1996, pp. 47-57) (see also Wu & Stansfield, 2001) list the following "task characteristics":

1. Characteristics of the setting (physical characteristics, participants, time of task)
2. Characteristics of the test rubrics (instructions, structure, time allotted for scoring, scoring method)
3. Characteristics of the input (format, language of input)
4. Characteristic of the expected response (format, language of expected response)

5. Relationship between input and response (reactivity, scope of relationships, directness of relationships)

While the authors call these "task characteristics," they appear to be task specifications in the test design sense of that phrase.

Skehan's (1998b, p. 95) list of task characteristics also have direct bearing on task-based performance testing:

1. Meaning is primary in the task
2. Task includes a communication problem that must be solved
3. Task has some sort of relationship to real-world activities
4. Task completion has some priority
5. Assessment of the task is in terms of successful outcome

These characteristics are clearly characteristics of task content.

Douglas (2000, pp. 50-71) lists the characteristics of tasks for language for specific purposes tests:

1. Characteristics of the rubric [specification of objective, procedures for responding; structure of the communicative event (number of tasks, relative importance of tasks, distinction between tasks), time allotment, evaluation (criteria for correctness, rating procedures)]
2. Characteristics of the input [prompt (including the features of the LSP context setting based on participants, purposes, form and content, tone, language, norms of interaction, genre, problem to be addressed), and language of input including format (visual, audio, vehicle of delivery, length) and level of authenticity (situational and interactional)]
3. Characteristics of the expected response [format (written, oral, physical), type of response (selected, limited production, extended production), response content (nature of language, background knowledge), and level of authenticity (situational, interactional)]
4. Characteristic of the interaction between input and response [reactivity, scope, directness]
5. Characteristics of assessment [construct definition, criteria for correctness, rating procedures]

His characteristics generally take the same form as those listed in Bachman and Palmer (1996) (i.e., they appear to be task specifications in the test design sense of that phrase), but they are considerably different in the details in that they attempt to also delineate the content of the tasks.

Chalhoub-Deville (2001, pp. 214-217) describes the following "characteristics":

1. Learner centered properties: go beyond regurgitation; encourage individual expression; and activate background knowledge and experience
2. Contextualization: task should embedded in on-going discourse; utterances in task should create discourse; and task should be embedded in meaningful situations
3. Authenticity: direct relationship between language use, instructional activities, and assessment; and relationship between test language and language of the real world

She labels these "task development characteristics," which seem to be mostly about what the content of the task will be like.

Norris (2001, p. 171) says that assessment tasks for university-level EAP students should:

1. Be of general interest to a broad range of university-level L2 English users
2. Represent several content areas
3. Not be highly discipline-specific
4. Engage the examinee in a variety of complex, skills-integrated L2 activities
5. Retain real-world fidelity to the greatest extent possible

These are what he calls "task selection criteria," and again these criteria seem to focus primarily on the content of the tasks.

Norris, Brown, Hudson, and Yoshioka (1998, p. 8) list the following:

1. Examinees must perform tasks
2. The tasks should be as authentic as possible
3. Success or failure in the outcome of the tasks, because they are performances, must usually be rated by qualified judges

These are three characteristics of what distinguishes performance assessments from other types of tests, but they are focused primarily on content.

Elsewhere, Norris, Brown, Hudson, and Yoshioka (1998, pp. 9 – 10) point out that "performance assessments will typically be based on tasks, which will be judged by raters on the basis of some form of rating scale.

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 scale 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)"

These are clearly criteria for rating of performances.

What can we make of all these different approaches to describing the criteria, dimensions, characteristics, and specifications of tasks in performance testing? First, all of these lists are clearly very much in a descriptive mode with regard to performance testing, and perhaps not much in agreement about what needs to be described or how any given category of interest should be described. In terms of *what needs to be described*, note that the above authors chose to describe the following:

1. Sound criteria for evaluating performances (Quellmalz, 1991)
2. Criteria for task specifications to be used in designing performance tests (Wiley & Haertel, 1996, p. 75)
3. "Dimensions of task specification" used in designing performance tests (Khattri, Reeve, & Kane, 1998, pp. 37-41)
4. "Task characteristics," which I classified as task specifications in the test design sense of that phrase (Bachman & Palmer, 1996, pp. 47-57)
5. Characteristics that the content of tasks should have (Skehan, 1998b, p. 95)
6. Characteristics that generally take the same task-specification form listed in Bachman and Palmer (1996), but also delineate the content of the tasks (Douglas, 2000, pp. 50-71)
7. Task development characteristics (which seem to be mostly about what the content of the task will be like) Chalhoub-Deville, 2001, pp. 214-217)
8. Task selection criteria, which are focused primarily on the content of the tasks (Norris (2001, p. 171)
9. Characteristics of what distinguishes performance assessments from other types of tests, but focused primarily on content (Norris, Brown, Hudson, & Yoshioka, 1998, p. 8)
10. Criteria for the rating of performances (Norris, Brown, Hudson, & Yoshioka (1998, pp. 9-10)

I think these various lists actually fall into three categories: task specifications, task content, and criteria for scoring performances. In addition, the Douglas paper appears to mix the last two categories. An outline of the above list reorganized into these categories looks like this:

1. Task Specifications

a. Criteria for task specifications to be used in designing performance tests (Wiley & Haertel, 1996, p. 75) #2
b. "Dimensions of task specification" used in designing performance tests (Khattri, Reeve, & Kane, 1998, pp. 37-41) #3

c. “Task characteristics,” which I classified as task specifications in the test design sense of that phrase (Bachman & Palmer, 1996, pp. 47-57) #4

2. Task Content

a. Characteristics that the content of tasks should have (Skehan, 1998b, p. 95) #5

b. Task development characteristics (which seem to be mostly about what the content of the task will be like) Chalhoub-Deville, 2001, pp. 214-217) #7

c. Task selection criteria, which are focused primarily on the content of the tasks (Norris (2001, p. 171) #8

d. Characteristics of what distinguishes performance assessments from other types of tests, but focused primarily on content (Norris, Brown, Hudson, & Yoshioka, 1998, p. 8) #9

3. Criteria for Scoring Performances

a. Sound criteria for evaluating performances (Quellmalz, 1991) #1

b. Criteria for the rating of performances (Norris, Brown, Hudson, & Yoshioka, 1998, pp. 9-10) #10

A Mix of 1 & 2 Above

a. Characteristics that generally take the same task-specification form listed in Bachman and Palmer (1996), but also delineate the content of the tasks (Douglas, 2000, pp. 50-71) #6

In addition, the ways the authors approach *how any given category of interest should be described* also differ considerably. Consider the two lists (#s 1 & 10) produced for *Criteria for Scoring* (Quellmalz, 1991; Norris, Brown, Hudson, & Yoshioka, 1998, pp. 9-10). A quick comparison of those two lists will reveal very little overlap. The same is true for lists classified under *Task Specifications* and those classified under *Task Content*. Clearly, task specifications, task content, and criteria for scoring performances are three categories of major interest for defining the characteristics of performance tests. Research (especially research based on actually designing, administering, scoring, interpreting, and validating performance tests) would be useful if it could define useful and generalizable task specifications, task content characteristics, and criteria for scoring performances.

What are the Benefits of Performance Testing (Why Bother?)?

In education circles, Wiggins (1989) argued that authentic educational measurement (including performance assessment):

1. Involves collaborative elements
2. Is contextualized and complex
3. Measures real-world tasks
4. Establishes standards that are clear to students and authentic

Moss (1992, pp. 229-230) felt that performance assessments help "document and encourage critical, creative, and self-reflective thought."

Herman, Aschbacher, and Winters (1992, p. 48) pointed out that using clear scoring criteria to evaluate student performances on tasks is beneficial because they:

1. Help teachers define excellence and plan how to help students achieve it
2. Communicate to students what constitutes excellence and how to evaluate their own work
3. Communicate goals and results to parents and others
4. Help teachers or other raters be accurate, unbiased, and consistent in scoring
5. Document the procedures used in making important judgments about students

Miller and Legg (1993) argued that performance assessment can:

1. Counterbalance the negative effects of washback from standardized tests (i.e., the negative effects of standardized testing on classroom teaching practices and curriculum).
2. Help promote multi-faceted approaches to information gathering for decision making

Khattri, Reeve, and Kane (1998, pp. 26-27) stated that "initiators and supporters of performance assessments claim that the assessments influence and inform instruction, because they:

1. Help teachers and other educators conduct a comprehensive evaluation of students' achievement, including students' strengths and weaknesses …
2. Support instruction and curriculum aimed at teaching for understanding by providing good pedagogical templates

3. Help teachers and other educators better assess students' understanding of procedural knowledge, which is not so easily judged through traditional assessment methods."

In language education, R. L. Jones (1985) felt that the greatest advantage of performance assessment was that it can measure students' abilities to accomplish real-life language tasks.

The "main features" Shohamy's (1992, pp. 517-518) diagnostic feedback model using performance testing included were as follows:

1. Faculty and assessment team should work collaboratively
2. The tests should tap language ranging from achievement to proficiency
3. The information provided should be diagnostic
4. Conclusions should be drawn at the local school level
5. The school should translate the conclusions and any interpretations and/or decisions made on the basis of test interpretations
6. Assessment should be repeated so that change can be monitored over time

All of these features can be taken to be advantages of performance testing.

Short (1993) felt that that non-standardized assessments (i.e., including performance assessments) offer more accurate accounts of student abilities than traditional tests do.

Shohamy (1995, pp. 188-207) felt that performance assessments:

1. Are more valid than traditional tests for predicting students' abilities to use language in real-world situations
2. Can offer positive washback effects on instructional
3. Test language in context along with other skills

Based on their review of the literature and their own experiences, Norris, Brown, Hudson, and Yoshioka (1998, pp. 15-16) summarized the benefits of using performance assessments:

1. Performance assessments can compensate for the following negative aspects of standardized testing (note that Mehrens, 1992 argued against these):
   a. Negative washback (i.e., any negative effects of testing on teaching, curriculum development, or educational policies)

   b. Lake Wobegon effects (i.e., teaching too closely to the test, or excessive focus on the teaching of abilities needed to score well on standardized tests)
   c. Bias in testing
   d. Irrelevant content
      i. Delimited domain coverage in multiple-choice tests (lack of inclusiveness of test content)
      ii. Lack of relationship with curriculum goals and objectives
   e. Multiple-choice tests measure ability to recognize only and cannot measure higher order thinking skills
   f. Multiple-choice tests lack obvious and real-world criteria for selection and scoring
2. Performance assessments can have positive washback effects by:
   a. Providing diagnostic information in functional, or task-based curriculums
   b. Supplying achievement information in functional, or task based curriculums
   c. Documenting critical thought, creativity, and self-reflection
   d. Encouraging critical thought, creativity, and self-reflection
   e. Aligning classroom assessment and instructional activities with authentic, real-life activities
   f. Showing students' strengths and weaknesses in detailed and real-world terms
3. Performance assessments approximate the conditions of real-life tasks so they can:
   a. Measure abilities to respond to real-life language tasks
   b. Create more accurate assessments of student's knowledge and ability than traditional multiple-choice tests
   c. Predict students' abilities in future, real-world situations
   d. Be more valid than traditional tests in terms of predicting students' abilities to use language in the future real-life situations

Brown, Hudson, Norris, and Bonk (2002, p. 6) state that performance assessments can be used for:

1. Establishing whether or not learners can accomplish specific target tasks that are directly related to learner, curricular, or professional objectives.

2. Evaluating various qualities of learners' language ability, including specific qualities such as accuracy, complexity, or fluency of second language production, as well as more holistic qualities such as general proficiency, communicative competence, etc.
3. Making interpretations about particular aspects of language learning that are (or are not) occurring within language classrooms and programs.

Brown and Hudson (1998; 2002) discuss the advantages of a number of different item and test formats. In Brown and Hudson (2002, pp. 74-78), they state that the advantages of any productive language test are as follows:

1. Guessing is not a major factor
2. They measure productive language use
3. They measure the interaction of receptive and productive skills

More specifically, they point out that performance tests can:

1. Simulate authentic language use
2. Correct for negative aspects of traditional standardized multiple-choice tests
3. Predict future real-life performances
4. Contribute positive washback

They further note that in performance- or task-based curriculum, performance tests can

1. Add a personal aspect to assessment
2. Can be integrated into and become a part of the curriculum
3. Assess learning processes

Table 1
Summary of Benefits of Using Performance Assessments

Content
Assess only relevant content
Measure productive language use
Can measure the interaction of receptive and productive skills
Add a personal aspect to assessment
Measure abilities to respond to real-life language tasks
Assess language ranging from achievement to proficiency
Test contextualized and complex language
Test more than multiple-choice recognition
Test higher order thinking skills
Assess learning processes
Assess students' understanding of procedural knowledge
Scoring
Use only real-world criteria for selection and scoring
Help teachers or other raters be accurate, unbiased, and consistent in scoring
Mediate rater bias effects in testing
Score Interpretations
Minimize guessing as a major factor
Provide diagnostic information in functional or task-based curriculums
Supply achievement information in functional, or task based curriculums
Assess students' knowledges and abilities better than traditional multiple-choice tests do
Encourage and document critical thought, creativity, and self-reflection
Demonstrate students' weaknesses *and* strengths in detailed and real-world terms
More accurately predict students' abilities to use language in future real-life situations
Encourage control of score interpretations at the local classroom and school levels
Curriculum Development
Can be integrated into and become a part of the curriculum
Align assessment and instructional activities with authentic, real-life activities
Test in harmony with curriculum goals and objectives
Help teachers define excellence
Help teachers plan how to help students achieve excellence
Support instruction and curriculum by providing good pedagogical templates
Help teachers conduct *comprehensive* evaluation of students' achievement
Create positive washback effects on instruction
Avoid factors leading to Lake Wobegon effects
Decision Making
Counterbalance the negative effects of washback from standardized tests
Document the procedures used in making important judgments about students
Help promote multi-faceted approaches to information gathering for decision making
Support drawing of conclusions at the local classroom and school levels
Encourage control of decision making at the local classroom and school levels
Assess continuously and repeatedly so that change can be monitored over time
Communication
Involve faculty and assessment team in collaborative activities
Establish standards that are clear to students and authentic
Help teachers communicate to students what constitutes excellence
Help teachers communicate to students how to evaluate their own work
Help teachers communicate goals and results to parents and others

In short, teachers, researchers, and testers need to realize that performance testing can have many benefits above and beyond those that can be garnered from traditional tests (see Table 1), including benefits that will affect content, scoring, score interpretations, curriculum development, decision making, and communication among all the stake holders involved in the educational process.

What Factors are Involved in Performance Task Difficulty?

Robinson (1995) points to the need for understanding task difficulty or complexity when he says that "Research into one area, the relative difficulty or complexity of second language tasks, is necessary as input to pedagogical decisions regarding the grading and sequencing of tasks for the purposes of syllabus design" (pp. 100-101). Any issue that affects the grading and sequencing of language teaching is inevitably related to the selection, grading, and sequencing of items on the related language tests. Consequently, this task difficulty issue has become a prominent one in language performance testing as well.

As a starting point, let's once again consider Robinson (1995) (for an overview of earlier research on this topic, see Norris, Brown, Hudson, & Yoshioka, 1998, pp. 39-51). He sets out to understand the issues involved in task complexity by examining what he calls:

1. Referential complexity
2. Structural complexity
3. Processing complexity

He also considers issues that he labels lexical load and memory demands. He finds that "…differences in measures of learner production are a consequence of differences in the cognitive load imposed by the tasks—a claim generalizable to tasks of many different types. The greater the attentional, memory, and reasoning demands of tasks, the greater the cognitive load they impose on the learner" (p. 130).

Robinson, Chi-chien, and Urwin (1996), study the effects of three other factors on the difficulty of speaking writing and listening tasks:

1. Cognitive load
2. Planning time
3. Prior information

Skehan (1996) proposes a framework for the implementation of task-based instruction, which implies that the following three factors are important to grading and sequencing of tasks:

1. Accuracy
2. Complexity
3. Fluency

Turning now to language performance tests, Kenyon (1998) studied the degree to which foreign language students assessed the difficulty of tasks in the same way they were sequenced by the descriptors in the ACTFL guidelines (1986). With the exception of one task, he found, on average that student ratings of task difficulty corresponded to the ACTFL hierarchy of difficulty.

Based on the three components (accuracy, complexity, and fluency) of Skehan's (1996) framework for implementing task-based instruction, Norris, Brown, Hudson, & Yoshioka (1998, pp. 58-59) defined the components from a task performance perspective:

1. Accuracy would involve the minimum level of precision in code usage necessary for successful communication.
2. Complexity would involve the minimum range of grammatical/structural code required for successful completion of a given communication task.
3. Fluency would involve the minimum on-line *flow* required by a given task for successful (acceptable) communication.

Skehan (1996) also developed a framework of factors that affect the difficulty of a task:

1. Code complexity, that is, the "traditional areas of syntactic and lexical difficulty and range" (p. 52).
2. Cognitive complexity including processing ("the amount of on-line computation that is required while doing a task, and highlights the extent to which the learner has to actively think through task content") and familiarity ("the extent to which the task draws on ready-made or pre-packaged solutions") (p. 52).

3. Communicative stress, which includes time pressure, modalities, scale of relationships (number of participants and relationships), stakes, and control (degree to which learners control the task) (pp. 52-53).

Brown, Hudson, Norris, and Bonk (2002), drawing directly on the work of Skehan, developed a task-independent rating scale that was based on the following task characteristics (p. 31):

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 mental 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 operations 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?

Robinson (2001) further developed his thinking about task complexity in a study that examined the effects of task complexity, task difficulty, and task conditions on task production and concluded that "the complexity of tasks does exert a considerable influence on learner production. …sequencing task on the basis of their cognitive complexity is to be preferred over sequencing decisions based on task difficulty or task conditions" (p. 51).

Skehan (2001) also developed his thinking about task difficulty specifically in the context language performance assessment by examining the relationships among five task characteristics and his earlier framework of accuracy, complexity, and fluency. He concluded:

> Recalling that task fulfils an important mediating function which shapes the nature of the performance which will be the basis for the ultimate rating of the candidate score, we can see that the task itself is hardly a constant in this equation. The five task characteristics which have been explored show that systematic (and potentially uncontrolled and undesirable) effects are likely to follow from any task selection decision. In other words, there may be significant consequences when one task is chosen rather than another. … Different candidates, in other words, might be disadvantaged or advantaged by the particular task that they might have taken as part of their test, and so their performance may not be directly comparable to the other candidates. (p. 182).

This is of course an issue that Rasch analysis, especially multi-faceted Rasch analysis, is ideal for addressing.

Wigglesworth (2001) examined the effects on performance of two task characteristics (task structure; familiarity of the activity) and two task conditions (native vs. non-native interlocutor; planning time). His results were inconclusive (showing no significant effects), but naturally, the author went on to over-interpret the "trends."

Elder, Iwashita, and McNamara (2002) examined the degree to which task complexity is related to students' perceptions of task difficulty and to their performance on those tasks. Generally the results failed to confirm the existing research on task difficulty and showed little if any relationship between examinees' perceptions and actual task difficulty. Similarly, Iwashita, McNamara, and Elder (2001) used various forms of analysis (including multi-faceted Rasch analysis) of the candidates' discourse on a performance test and their scores on that same test, and examined the degree to which different task characteristics (perspective, immediacy, adequacy, & planning time) and performance conditions (i.e., different levels of cognitive demand) were related to different levels of Skehan's accuracy, complexity, and fluency. Generally speaking their results failed to confirm the findings of existing research.

O'Sullivan (2002) suggested the importance of considering not only the characteristics of the task, but also those of the test-taker and the interlocutor (and their relationship: acquainted or not acquainted) in examining oral performance test results.

Brindley and Slatyer (2002) explored the effects of task characteristics and conditions (speech rate, text type, number of hearings, live vs. audio-recorded input source, and item format) on students' performances on listening tasks. They concluded that:

> …the complexities of the interactions between task characteristics, item characteristics and candidate responses that we have encountered in this study suggest that simply adjusting one task-level variable will not automatically make the task easier or more difficult. Given the complexities, we must conclude that the rather simplistic notion of 'difficulty' as reflected in item difficulty statistics is of limited usefulness in understanding what happens when an individual candidate interacts with an individual item. (p. 390)

Bachman (2002) reflects on the complexity issues is task-based language performance assessment. He calls for the integration of tasks and constructs (a sort of integration of the traditional forms of evidence-based construct validity and content validity as best I can understand it) as a solution to the problem, but offers no evidence that such integration will make any difference in untangling the complexities involved in task-based testing.

A more reasonable possible solution to the conundrum of complexity in performance testing is the "evidence-centred assessment design" model proposed by Mislevy, Steinberg, and Almond (2002) for task-based test design. It involves the use of different perspectives (an assembly model, student models, evidence models, & task models) to "design complex tasks, evaluate students' performances and draw valid conclusions therefrom" (p. 477). Given that theirs is only a proposal, it should be studied to determine if an evidence-centred assessment design approach is valid and if it does indeed solve the problems of complex interactions between task characteristics, task conditions, student characteristics, and so forth.

How Should We Best Go about Scoring Performance Tests

Scoring criteria. In addition to the work that has been done over the years to develop scales for rating writing performances (e.g., Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981; Brown & Bailey, 1984; ETS, 1996) and speaking performances (e.g., ETS, 2001; ACTFL, 1984, 2004), some recent, solid, empirically based work has been done on the development of scoring criteria for the Common European Framework (see North & Schneider, 1998; Lenz, 2000; Council of Europe, 2001) and the use of *can-do statements* (see N. Jones, 2000, 2001; North & Schneider, 1998, 2000; Council of Europe, 2001). In addition, Norris and Bonk (2000) and Norris (2001) reported in-depth on the process of developing both *task-dependent* and *task-independent* criteria for scoring performance tests that also appear in the various permutations of Brown, Hudson, Norris, and Bonk (2002) (i.e., Brown, Norris, Hudson, & Bonk, 1999; Norris, Brown, Hudson, & Bonk, 2002). Tierney and Simon (2004) pointed out that the scoring rubrics used in many schools are not as useful as they could be because of inconsistencies in the descriptions of performance criteria. They suggested asking three questions of the descriptions in all scoring rubrics: Are all the performance criteria explicitly stated? Are the attributes explicitly stated for each performance criterion? Are the attributes consistently addressed from one level to the next on the progression scale?

Rater effects. Other recent work has focused on the effects of raters on the whole process of performance assessment. Ross and Berwick (1992) examined the effects of rater discourse on student performance in oral interview tests. A. Brown (1995) used multi-faceted Rasch analysis to examine the effects of different types of raters on the scoring of an occupation-specific language performance test. Chalhoub-Deville (1995) used multidimensional scaling to show how oral assessment scales can be derived from different scaling criteria and tests, and how application of such scales differed for several rater groups. Lumley and McNamara (1995) used multi-faceted Rasch analysis to examine rater characteristics and rater bias in an occupational English performance test with a particular emphasis on the importance of using multi-faceted Rasch analysis of performance test data to improve rater training. Bachman, Lynch, and Mason (1995) studied the variability in tasks and rater judgments in a performance test of foreign language speaking. Henning (1996) investigated the degree to which rater agreement and score reliability are necessarily identical notions for the scores on a simulated performance test. Sakyi (2000) examined how raters evaluate compositions when doing holistic scoring of ESL writing in terms of content-related and language-related factors, as well as in terms of the influences of previous stimuli (i.e., previously read compositions) on ratings. Meiron and Schick (2000) investigated whether students who had not made any quantitatively measurable gains over 11 weeks of training on an oral proficiency tests varied qualitatively in their performances and found that they did indeed vary, especially in terms of lexical control (though different raters appeared to react differently). Kondo-Brown, K. (2002) used multi-faceted Rasch analysis to examine the effectiveness of a five category (content, organization, vocabulary, language use, and mechanics) analytic writing scale for testing Japanese second language writing performance when applied by three raters; she examined especially carefully the effects of rater bias and category bias on the students' scores. Caban (2003) also used multi-faceted Rasch analysis, but she examined the bias of scoring categories (grammar, fluency, content, pronunciation, pragmatics, compensation techniques, and overall intelligibility) and four rater groups (trained L1 and L2 ESL teachers, peers, and English native speakers with no ESL training) in their ratings of speaking test results.

Are Performance Tests Reliable?

Another area of importance in performance testing is issue of the reliability of performance tests. Certainly, interrater reliability has been reported along the way as an incidental part of many of the performance test development and validation studies reported above. For example, Wesche (1987) reports achieving K-R20 reliabilities of .91 and .92 on performance tests developed for the *Ontario Test of ESL*, and Stansfield, Wu, and van der Heide (2000) report K-R20 reliabilities of .87 and .92 for the Forms A and B of their performance test. Brown, Hudson, Norris, and Bonk (2002) report a large number of different alpha coefficients and interrater correlations (adjusted using the Spearman-Brown prophecy formula for appropriate numbers of raters) for all possible combinations of scales and raters in their study. Indeed, a number of other studies have addressed the classical theory issue of internal consistency reliability of performance tests, and all indications are that language testers can design reasonably reliable performance tests.

However, in the wider educational testing research, more detailed investigations have been conducted into the sources of measurement error in rating performance assessments with the goal of creating such tests that are more reliable. For instance, Shavelson, Mayberry, Li, and Webb (1990) used Generalizability theory (G theory) to enhance the reliability of job performance measurements for Marine Corp riflemen, while Shavelson, Baxter, & Gao (1993) used G theory to examine the effects of numbers of persons, raters, tasks, and occasions on the reliability of a series of science and math performance assessments. Gao and Brennan (2001) examined the variability of the estimated variance components for persons, raters, and tasks over different occasions in performance assessment.

The only such studies I know of in Applied Linguistics are:

1. Brown and Bailey (1984), which looks at the effects of raters and categories on writing performance assessments.
2. Brown (1988, 1989, 1990, 1991), which are a series of validation studies of the Manoa Writing Test (for native speakers of English) in four sequential years, each

of which uses G theory to examine the effects of numbers of raters and prompt topic types on reliability.

3. Bachman, L.F., Lynch, B.K., and Mason, M. (1995) Investigating variability in tasks and rater judgments in a performance test of foreign language speaking.

More complex studies of the effects of numbers of raters, task types, rating categories, occasions, etc. would be very useful and timely. Since I have the detailed performance testing data and a long track record of using G-theory, I'm thinking that this would be a good direction for my own research to go.

Are Performance Tests Valid?

Numerous papers have explored and discussed the many issues involved in the validity of performance tests. This has been true in education circles (Linn, Baker, & Dunbar, 1991; Moss, 1992; Messick, 1994; Arter, 1999; Kane, Crooks, & Cohen, 1999; Nichols & Sugrue, 1999) as well as in language testing circles (Messick, 1996; Chalhoub-Deville, 1996; Chapelle, 1998, 1999; Bachman, 2002; Norris, 2002). While the validity of performance assessments is clearly an important issue, we have had considerable time to reflect and discuss. I would say it is time to base our reflections more on experience. In other words, let's get down to actually developing and validating more performance assessments and learning from the experiences. Discussion tempered by experience and empirical research will be far more valuable than the endless spinning of theoretical distinctions.

Are There Other Issues About Performance Tests That Need to be Researched?

Finally, a number of issues related to performance testing have recently been addressed in the educational measurement literature that might point to future directions for research in second language testing. For instance, Yen (1993) did a study of scaling performance assessments while applying various strategies for managing the issue of *local item dependence*. Zwick, Donoghue, and Grima (1993) addressed the issue of *differential item functioning* for performance tests. Raymond and Chockalingam (1993) used least squares models to correct for *rater effects* in performance assessment. Hardy (1996), and Monk (1996) explored ways to analyze the costs of performance assessment. Baker (1996) addressed the issue of equity in performance assessment. Linn (1996) addressed the important issues of calibrating, equating, and linking performance assessments. Ayala, Shavelson, Yin, and Schultz (2002) used multitrait-multimethod analysis to study the *method effects* in using performance assessment to measure reasoning dimensions in science achievement.

CONCLUSIONS

I started this paper by defining task-based testing as a subset of the broader category of performance testing. If there is one thing I have learned from writing this paper, it is that in the last decade or so the notions of task-based testing and performance testing have merged. That is, today, any project that sets out to design, analyze, and validate a language performance test will inevitably be doing so on the basis of the literatures of task-based language teaching, learning, and testing. In short, task-based testing, which used to be a subset of performance testing, has become inextricably intertwined with the whole idea of language performance testing.

Suggestions for Future Research on Performance Testing

Given that performance and task-based testing seem to have merged in our thinking about language performance testing, all of the following questions and sub-questions should probably be addressed in terms of task-based performance testing.

What happens when actual performance tests are developed and validated? While it may appear that a great many studies of actual performance test development and validation projects have already been conducted, I would argue that we need many more such studies and the experience and knowledge that will accrue from doing and thinking about such test development and validation projects. To that end, the following sub-questions might usefully be addressed:

1. To what extent would the results be similar when other language performance testing projects are conducted in other similar ESL programs (survival, immigrant, academic, etc.)
2. To what extent would the results be similar when other language performance testing projects are conducted in other sorts of ESL/EFL programs?
3. To what extent would the results be similar when other language performance testing projects are conducted in other second and foreign languages?
4. To what extent would the results of other language performance testing projects of reading, writing, listening, speaking, and pragmatics differ from those that have already been done?
5. To what extent can computer-based, computer-adaptive, and internet-based performance tests be effective?
6. To what extent are computer-based, computer-adaptive, and internet-based performance tests similar to or different from face-to-face performance tests?

What are the characteristics of performance tests? Much space has been devoted in the literature to describing the "characteristics" of performance tests. My analysis of that literature found that basically the authors are describing either (a) task specifications, (b) task content, or (c) criteria for scoring performances, or some mixture of (a) and (b). Some of the sub-questions that might usefully be researched on this topic are the following:

1. What useful and generalizable characteristics can be gleaned from the literature to create adequate/effective task specifications?
2. What useful and generalizable characteristics can be gleaned from the literature to create adequate/effective task content?
3. What useful and generalizable characteristics can be gleaned from the literature to create adequate/effective criteria for scoring performance tests?
4. How do the three categories (task specifications, task content, and criteria for scoring) interact?

What are the benefits of performance testing (why bother?)? The literature claims that a number of benefits will accrue from actually doing performance testing. In my analysis of that literature, I found that the benefits fell into six categories: (a) content, (b) scoring, (c) score interpretations, (d) curriculum development, (e) decision making, and (f) communication. Some of the sub-questions that might usefully be researched are the following:

1. Are there additional benefits from doing performance testing within the (a) content, (b) scoring, (c) score interpretations, (d) curriculum development, (e) decision making, and (f) communication categories that have not been proposed in the literature?
2. Are there other categories of benefits from doing performance testing that have not been proposed, but should be considered?
3. Is there any evidence to support the reality of these claimed benefits?
4. Are these benefits perceived to be real by students, teachers, administrators, parents, etc.?

What factors are involved in performance task difficulty? A number of sets of factors have been posited in the literature as being related to task difficulty. Some of the task-based learning research appears to support some of these factors, though even this research is not conclusive. In the performance testing literature, the empirical studies seem to indicate no effects on students' performances for task difficulty, task complexity, task conditions, task characteristics, examinees' perceptions of task difficulty, etc. In addition, there is some evidence that rater characteristics and test taker characteristics may be important (including their degree of relationship between the raters and test takers). Some of the sub-questions that might usefully be researched are the following:

1. To what degree do task difficulty, task complexity, task conditions, task characteristics, examinees' perceptions of task difficulty, rater characteristics, test taker characteristics, the degree of relationship between the raters and test takers, etc. affect students' performances on tasks?
2. To what degree do task difficulty, task complexity, task conditions, task characteristics, examinees' perceptions of task difficulty, rater characteristics, test taker characteristics, the degree of relationship between the raters and test takers, etc. interact with each other and with students' performances on tasks?
3. Does the "evidence-centred assessment design" approach (proposed by Mislevy, Steinberg, and Almond, 2002) produce a reliable and valid test of tasks for performance testing?
4. Does the "evidence-centred assessment design" approach allow us to adequately deal with any interactions found in #2 above?

How should we best go about scoring performance tests (in terms of scoring criteria and rater effects)? As discussed above, a number of different criteria have been proposed over the years for rating different sorts of language performance assessments. Some of the sub-questions that might usefully be researched are the following:

1. Which of those already existing rating scales function well for different purposes and in different situations?
2. What general types of rating criteria work well for different purposes and different situations?
3. How do analytic and holistic scoring methods compare in terms of effectiveness for different purposes and different situations?
4. In the rubrics currently used in second language teaching, are all the performance criteria explicitly stated, the attributes of each performance criterion explicitly spelled out, and the attributes sequenced consistently from one level to the next?
5. How do task-dependent and task-independent scoring methods compare in terms of effectiveness for different purposes and different situations?

6. What are rater and rating category biases in testing for different purposes and different situations?
7. How can real-time multi-faceted Rasch analysis be used to improve rater training?

Are performance tests reliable? We know that we can produce reliable performance tests, but still the following sub-questions need to be pursued:

1. Are performance tests reliable (an ongoing question that must be addressed in all such test development projects) for different purposes and different situations?
2. What are the effects of numbers of raters, categories, tasks, occasions, etc. on the reliability of performance tests?
3. How can we make performance assessments more reliable by adjusting the numbers of raters, categories, tasks, occasions, etc.?

Are performance tests valid? While there has been a great deal of speculation on what makes performance tests valid, a great deal more work needs to be done. Inevitably, a number of issues will be involved: task difficulty, task complexity, task conditions, task characteristics, examinees' perceptions of task difficulty, rater characteristics, test taker characteristics, the degree of relationship between the raters and test takers, etc. and how these factors interact with each other and with students' performances on tasks. Thus, I would say that the sub-questions for this validity issue are the same as those given above for ***What factors are involved in performance task difficulty?*** I would also say that answers to those questions will best be derived from experience and empirical research?

Are there other issues about performance tests that need to be researched? We need to keep a constant eye on the educational measurement literatures on performance testing for ideas on other areas of research that we too may need to address. From my doing so, I have found that the following sub-questions need to be answered for language performance testing:

1. To what degree is local item dependence a problem in language performance testing?
2. What strategies can be used to manage any existing problems with local item dependence in language performance tests?
3. What are the effects of differential item functioning for language performance tests and how does it affect various sub-groups of examinees from different language, nationality, ethnicity, gender, socio-cultural, etc. backgrounds?

4. What issues of equity exist in language performance testing?
5. How can we correct for rater effects in language performance testing?
6. What are the costs of language performance assessment?
7. What are the options in methods of analyzing the costs of language performance tests?
8. How should we go about calibrating, equating, and linking language performance tests across different collections of tasks?
9. How should we go about calibrating, equating, and linking language performance tests across different languages?
10. To what degree are method effects problematic in performance tests?

REFERENCES

ACTFL. (1986). *ACTFL proficiency guidelines (Revised)*. Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

ACTFL. (2004). *ACTFL proficiency guidelines (Revised)*. Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages. Retrieved January 1, 2004 from the World Wide Web: http://www.sil.org/lingualinks/LANGUAGELEARNING/OtherResources/ACTFLProficiencyGuidelines/ACTFLProficiencyGuidelines.htm

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.

Anderson, L. W. (2003). *Classroom assessment: Enhancing the quality of teacher decision making.* Mahwah, NJ: Lawrence Erlbaum Associates.

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers.* San Francisco, CA: Jossey-Bass.

Arter, J. (1999). Teaching about performance assessment. *Educational Measurement: Issues and Practice, 18*(2), 30-44.

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

Ayala, C. C., Shavelson, R. J., Yin, Y., & Schultz, S. E. (2002). Reasoning dimensions underlying science achievement: The case of performance assessment.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453-476.

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

Bachman, L.F., Lynch, B.K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 239-257.

Bailey, K. M. (1985). If I had known then what I know now: Performance testing of foreign teaching assistants. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 153-180). Ottawa: University of Ottawa.

Baker, E. L. (1996). Performance assessment and equity. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises and challenges* (pp. 183-199). Mahwah, NJ: Lawrence Erlbaum Associates.

Banks, J. C. (1997). *Creating and assessing performance-based curriculum projects: A teacher's guide to project-based learning and performance assessment.* Edmonds, WA: CATS Publications.

Berwick, R., & Ross, S. (1996). Cross-cultural pragmatics in oral proficiency interview strategies. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 34-54). Cambridge: Cambridge University.

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).

Brindley, G. (1994). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing, 15*(1), 45-85.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369-394.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15.

Brown, J. D. (1988). *1987 Manoa Writing Placement Examination: Technical Report #1*. Honolulu, HI: Manoa Writing Program, University of Hawai'i at Manoa.

Brown, J. D. (1989). *1988 Manoa Writing Placement Examination: Technical Report #2*. Honolulu, HI: Manoa Writing Program, University of Hawai'i at Manoa.

Brown, J. D. (1990). *1989 Manoa Writing Placement Examination: Technical Report #5*. Honolulu, HI: Manoa Writing Program, University of Hawai'i at Manoa.

Brown, J. D. (1991). *1990 Manoa Writing Placement Examination: Technical Report #11*. Honolulu, HI: Manoa Writing Program, University of Hawai'i at Manoa.

Brown, J. D. (Ed.). (1998). *New ways of classroom assessment*. Washington, DC: Teachers of English to Speakers of Other Languages.

Brown, J. D. (2000). Observing pragmatics: Testing and data gathering techniques. *Pragmatics Matters, 1*(2), 5-6.

Brown, J. D. (2001). Six types of pragmatics tests in two different contexts. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301-325). Cambridge: Cambridge University.

Brown, J.D. & Bailey, K.M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*, 21-42.

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

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Brown, J. D., Hudson, T., & Kim, Y. (2001). *Developing Korean language performance assessments*. (NFLRC Research Note #27). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

Brown, J. D., Hudson, T., Norris, J. M., & Bonk, W. (2002). *An investigation of second language task-based performance assessments.* (Technical Report #24). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

Brown, J. D., Norris, J. M, Hudson, T., & Bonk, W. (1999). Performance assessments of Japanese ESL and EFL students. *University of Hawaii Working Papers in ESL, 18*(2), 99-139.

Brualdi, A. (1998). Implementing performance assessment in the classroom. *Practical Assessment, Research & Evaluation, 6*(2). Retrieved March 1, 2004 from http://PAREonline.net/getvn.asp?v=6&n=2 .

Bygate, M., Skehan, P., & Swain, M. (2001). Introduction. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 1-20). Harlow, UK: Pearson Education.

Byrnes, H. (2002). The role of task and task-based assessment in a content-oriented collegiate foreign language curriculum. *Language Testing, 19*(4), 419-437.

Caban, H. L. (2003). Rater Group Bias in the Speaking Assessment of Four L1 Japanese ESL Students. *Second Language Studies, 21*(2), Spring 2003, 1-43.

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.

Candlin, C. N. (2001). Afterword: Taking the curriculum to task. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 229-243). Harlow, UK: Pearson Education.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Chalhoub-Deville, M. (1996). Performance assessment and the components of the oral construct across different tests and rater groups. In M. Milanovic & N. Saville (Eds.), *Performance testing , cognition and assessment* (pp. 55-73). Cambridge: Cambridge University.

Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 210-228). Harlow, UK: Pearson Education.

Chapelle, C. (1998) Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interfaces,* (pp. 32-70). Cambridge: Cambridge University.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254-272.

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.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching and assessment.* Cambridge: Cambridge University.

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.

Crookes, G., & Gass, S. M. (Ed.). (1993). *Tasks in a pedagogical context: Integrating theory and practice*. Philadelphia: Multilingual Matters.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D., & Selinker, L. (1993). Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. In D. Douglas, & C. Chapelle (eds.), *A new decade of language testing* (pp. 235-56). Alexandria, VA: TESOL.

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.

Eisner, E. W. (1999). The uses and limits of performance assessment. *Phi Delta Kappan*, *80*(9), 658-660.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*(4), 347-368.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford University.

ETS. (1996). *TOEFL Test of Written English Guide.* Princeton, NJ: Educational Testing Service.

Fitzpatrick, T., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

ETS. (2001). *TSE and SPEAK score user's guide (2001-2002 edition).* Princeton, NJ: Educational Testing Service.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 237-270). Washington, DC: American Council on Education.

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.

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14*(2), 191-203.

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

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performances. In R. M. Gagné (Ed.), *Psychological principles in systems development* (pp. 419-474). New York: Holt, Rinehart, & Winston.

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan*, *80*(9), 662-666.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs*. TOEFL Research Report # 54. Princeton, NJ: Educational Testing Service.

Hardy, R. (1996). Performance assessment: Examining the costs. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises and challenges* (pp. 107-117). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Hill, K. (1998). The effect of test-taker characteristics on reactions to and performance on an oral English proficiency test. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 209-229). Mahwah, NJ: Lawrence Erlbaum Associates.

Horowitz, D. M. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20*(3), 445-462.

Hudson, T., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics*. Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai‘i Press.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics*. Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai‘i Press.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401-436.

Jacobson, W. H. (1986). An assessment of the communication needs of non-native speakers of English in an undergraduate physics lab. *English for Specific Purposes, 5*(2), 189-195.

Jacobs, H.L., S.A. Zinkgraf, D.R. Wormuth, V.F. Hartfiel, & J.B. Hughey. (1981). *Testing ESL composition: a practical approach*. Rowley, MA: Newbury House.

Jones, N. (2000). Background to the validation of the ALTE 'can-do' project and the revised common European framework. English as a Foreign Language (EFL) Research Notes, Issue 2 retrieved February 20, 2004 from: http://www.cambridge-efl.org/rs_notes/0002/rs_notes 2_6.cfm.

Jones, N. (2001). Using can-do statements to equate computer-based tests across languages. Paper presented at the 23rd Annual Language Testing Research Colloquium, St. Louis, Mo.

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.

Kane, M., & Mitchell, R. (1996). *Implementing performance assessment: Promises, problems, and challenges*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kenyon, D. M. (1998). An investigation of the validity of task demands on performance-based tests of oral proficiency. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 19-40). Mahwah, NJ: Lawrence Erlbaum Associates.

Khattri, N., Reeve, A., & Kane, M. (1998). *Principles and practices of performance assessment*. Manwah, NJ: Lawrence Earlbaum Associates.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language performance. *Language Testing, 19*(1), 3-32.

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

Lenz, P. (2000). *Piloting the Swiss model of the European language portfolio: Evaluator's final report.* Fribourg: Centre d'enseignement et de recherche en langues étrangères.

Linn, R. L. (1996). Linking assessments. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises and challenges* (pp. 91-105). Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15-21.

Long, M. H. (1985). A role for instruction in second language acquisition: Task-based training. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 77-100). Clevedon, Avon: 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. (1997). Authenticity and learning potential in L2 classroom discourse. In G. M. Jacobs (Ed.), *Language classrooms of tomorrow: Issues and responses* (pp. 148-169). Singapore: SEAMEO Regional Language Center.

Long, M. H. (1998). Focus on form in task-based language teaching. *University of Hawaii Working Papers in ESL*, *16*(2), 35-49.

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.

Long, M. H., & Norris, J. M. (2001). Task-based language teaching and assessment. In M. Byram (Ed.), *Encyclopoedia of language teaching*. London: Routledge.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

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.

McNamara, T. (1997). Performance testing. In Clapham, C. Corson, D. (Eds.), *Encyclopedia of language and education*: Volume 7 Language testing and assessment (pp. 131-139). Dordrecht, NL: Kluwer Academic.

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

Meiron, B. E., & Schick, L. S. (2000). Ratings, raters, and test performance: An exploratory study. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 153-176). Cambridge: Cambridge University.

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.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477-496.

Monk, D. H. (1996). Conceptualizing the costs of large-scale pupil performance assessment. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises and challenges* (pp. 119-137). Mahwah, NJ: Lawrence Erlbaum Associates.

Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation, 8*(14). Retrieved March 1, 2004 from http://PAREonline.net/getvn.asp?v=8&n=14

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice, 18*(2), 18-29.

Norris, J. M. (2000). Purposeful language assessment: Matching testing alternatives with intended test use. *English Teaching Forum*, *38*(1), 18-23.

Norris, J. M. (2001). Identifying rating criteria for task-based EAP assessment. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (Technical Report #21) (pp. 163-204). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing, 19*(4), 337-346.

Norris, J., & Bonk, W. (2000). Comparing the success criteria of various groups of stakeholders in L2 performance assessments. Paper presented at AAAL conference, Vancouver (March, 2000).

Norris, J., Brown, J., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments* (Technical Report #18). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395-418.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.

North, B., & Schneider, G. (2000). *«Dans d'autres langues, je suis capable de…» Échelles pour la description, l'évaluation et l'auto-évaluation des competences en langues étrangères: Rapport de valorisation.* Berne, Switzerland: Centre Suisse de coordination pour la recherché en education (CSRE).

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. M. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 55-68). Clevedon, Avon: Multilingual Matters.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277-295.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56.

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.

Raymond, M. R., & Chockalingam, V. (1993). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement, 30*(3), 253-268.

Roeber, E. D. (1996). Guidelines for the development and management of performance assessments. *Practical Assessment, Research & Evaluation, 5*(7). Retrieved March 1, 2004 from http://PAREonline.net/getvn.asp?v=5&n=7 .

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*(1), 99-140.

Robinson, P. (1996a). Introduction: Connecting tasks, cognition and syllabus design. *The University of Queensland Working Papers in Language and Linguistics, 1*(1), 1-14.

Robinson, P. (1996b). Task-based testing, performance-referencing and program development. *The University of Queensland Working Papers in Language and Linguistics, 1*(1), 95-117.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 21*(1), 27-57.

Robinson, P., & Ross, S. (1996). The development of task-based testing in English for academic purposes contexts. *Applied Linguistics, 17*(4), 455-476.

Robinson, P., Chi-chien, S., & Urwin, J. (1996). Three dimensions of second language task complexity. *The University of Queensland Working Papers in Language and Linguistics, 1*(1), 15-30.

Roever, C. (2001). A web-based test of interlanguage pragmalinguistic knowledge: Speech acts, routines, implicatures. Unpublished Ph.D. dissertation. Honolulu, HI: University of Hawaii at Manoa.

Rose, K. R. (1994). On the validity of discourse completion tests in non-Western contexts. *Applied Linguistics, 15*, 1-14.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *SSLA, 14*, 159-176.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129-152). Cambridge: Cambridge University.

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.

Shameem, N. (1998). Validating self-reported language proficiency by testing performance in an immigrant community: The Wellington Indo-Fijians. *Language Testing, 15*(1), 86-108.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347-363.

Shavelson, R. J., Mayberry, P. W., Li, W., & Webb, N. M. (1990). Generalizability of job performance measurements: Marine Corp riflemen. *Military Psychology, 2*, 129-144.

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.

Short, D. J. (1993). Assessing integrated language and content instruction. *TESOL Quarterly, 27*(4), 627-256.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing, 1*, 202-20.

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

Skehan, P. (1998a). Task-based instruction. *Annual Review of Applied Linguistics, 18*, 268-286.

Skehan, P. (1998b). *A cognitive approach to language learning.* Oxford, Oxford University.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167-185). Harlow, UK: Pearson Education.

Skehan, P., & Foster, P. (1999). The influence of task-structure and processing conditions on narrative retellings. *49*(1), 93-120.

Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *Modern Language Journal, 76*(2), 129-141.

Stansfield, C. W., Wu, W. M., & van der Heide, M. (2000). A job-relevant listening summary translation exam in Minnan. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 177-200). Cambridge: Cambridge University.

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

Swezey, R. W., Hutcheson, T. D., & Swezey, L. L. (2000). Development of second-generation computer-based team performance assessment technology. *International Journal of Cognitive Ergonomics, 4*(29) 163-170.

Thompson, C. K. (1992). Learner-centered tasks in the foreign language classroom. *Foreign Language Annals, 25*(6), 523-531.

Tierney, Robin, & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation, 9*(2). Retrieved March 1, 2004 from http://PAREonline.net/getvn.asp?v=9&n=2 .

Tulou, G., & Pettigrew, F. (1999). Performance assessment for language students. In M. Kassen (Ed.), *Language learners of tomorrow* (pp. 189-231). Lincolnwood, IL: National Textbook Co.

van den Branden, K., Depauw, V., & Gysen, S. (2002). A computerized task-based test of second language Dutch for vocational training purposes. *Language Testing, 19*(4), 438-452.

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.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 186-209). Harlow, UK: Pearson Education.

Wiley, D. E., & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises and challenges* (pp. 61-89). Mahwah, NJ: Lawrence Erlbaum Associates.

Willis, J. (1996). *A framework for task-based learning*. Essex, England: Addison Wesley Longman.

Wu, W. M., & Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language Testing, 18*(2), 187-206.

Yamashita, S. O. (1996a). *Comparing six cross-cultural pragmatics measures.* Unpublished doctoral dissertation, Temple University, Philadelphia, PA.

Yamashita, S. O. (1996b). *Six measures of JSL Pragmatics* (Technical Report #14. Second Language Teaching and Curriculum Center). Honolulu, HI: University of Hawai‘i Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Yoshitake, S. S. (1997). Measuring interlanguage pragmatic competence of Japanese students of English as a foreign language: A multi-test framework evaluation. Unpublished doctoral dissertation, Columbia Pacific University, Novata, CA.

Yoshitake, S., & Enochs, K. (1996). Self assessment and role play for evaluating appropriateness in speech act realizations. ICU Language Research Bulletin, 11, 57-76.

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.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tests. *Journal of Educational Measurement, 30*(3), 233-251.