Language Testing Reconsidered

Fox, Janna, Wesche, Mari, Bayliss, Doreen, Cheng, Liying

Published by University of Ottawa Press

Fox, Janna, et al.
Language Testing Reconsidered.

For additional information about this book
https://muse.jhu.edu/book/4459

For content related to this chapter
https://muse.jhu.edu/related_content?type=book&id=103390
What Is the Construct?  
The Dialectic of Abilities and Contexts 
in Defining Constructs in Language 
Assessment

Lyle F. Bachman  
University of California, Los Angeles

Abstract

Understanding the roles of abilities and contexts, and the interactions between 
these as they affect performance on language assessment tasks, has remained 
a persistent problem in language assessment. Approaches to this problem 
over the past half century have led to three general ways of defining the con-
struct, or what we want to assess: 1) ability-focused, 2) task-focused, and 3) 
interaction-focused. While the different theoretical perspectives that underlie 
these approaches are not mutually exclusive, they are based on different sets 
of values and assumptions. Because of these differences, the conundrum of 
abilities and contexts and how they interact in language use and language as-
sessments is, in my view, essentially a straw issue theoretically, and may not 
be resolvable at that level.

Nevertheless, the issues raised by these different approaches have im-
portant implications and present challenging questions for both empirical 
research in language assessment and for practical assessment design, develop-
ment, and use. For research, they imply the need for a much more inclusive 
methodological approach, involving both so-called quantitative and qualita-
tive perspectives. For practice, they imply that focus on any one of these 
approaches (ability, task, interaction), to the exclusion of the others, will lead 
to potential weaknesses in the assessment itself, or to limitations on the uses 
for which the assessment is appropriate. This means that we need to address 
all three in the design, development, and use of language assessments.

Introduction

A persistent problem in language assessment has been that of understanding 
the roles of abilities and contexts, and the interactions between these, as they 
affect performance on language assessment tasks. The way we view these 
roles has clear implications for the way we define the constructs we intend to 
assess, and for the way we interpret and use assessment results. Furthermore, 
the way we view abilities and contexts — whether we see these as essentially 
indistinguishable or as distinct — will determine, to a large extent, the research 
questions we ask and how we go about investigating these empirically. In
the recent history of language testing,\(^2\) one can trace a dialectic, if you will, between what has been called *construct-based* and *task-based* approaches to language testing (e.g., Skehan, 1998). Bachman (2004) has argued that these two approaches are quite distinct because they are based on different ways of defining the construct we intend to assess, and different assessment use arguments. They also use different approaches to the way assessments are designed and developed, and lead to different kinds of score-based interpretations. More recently, some researchers, drawing largely on research in social interaction and discourse analysis, have proposed an interactionalist perspective on language assessment, which views the construct we assess not as an attribute of either the individual language users or of the context, but as jointly co-constructed and residing in the interactions that constitute language use.

I will begin with a brief discussion of what I believe is a central and persistent problem in language assessment, that of relating abilities and contexts in the way we define the construct—or what it is that we want to assess. I will then provide a cursory historical—since the 1960s—overview of different approaches to defining what language assessments assess, which I believe illustrate a dialectic between what has been called “trait/ability-focused” and “task/context-focused” perspectives on or approaches to defining constructs in language testing. I will then discuss what I will refer to as an “interactional perspective” to defining what language assessments measure, pointing out some of the important issues that its proponents have raised, and what I see as some potential problems with this approach. I will then discuss some implications that an understanding of these differing perspectives has for language assessment research and practice.

**Persistent Problem: Relating Abilities and Contexts**

A number of researchers have discussed the problem of disentangling ability from context in language assessment. Bachman (1990), for example, described what he called the “fundamental dilemma” of language testing:

> Language is both the object and instrument of our measurement. That is, we must use language itself as a measurement instrument in order to observe its use to measure language ability ... This makes it extremely difficult to distinguish the language abilities we want to measure from the method factors used to elicit language. (pp. 287–288)

Similarly Skehan (1998), argues that the three underlying problems of language testing are inferring abilities, predicting performance, and generalizing across contexts. For him, the dilemma, or what he calls the “abilities/performance/context conflict” (p. 155), is that these problems cannot be solved simultaneously because each problem conceptualizes what is to be sampled differently. If the focus is on sampling abilities, for example, the performance and contexts
are likely to be ignored or underplayed, and similarly if one chooses to sample performances or contexts.

Bachman’s and Skehan’s observations were based on and have stimulated a great deal of research in language testing that has investigated the relative effects of abilities/processes and test method factors/tasks/contexts. Bachman’s perspective grew out of the trait-method studies that were conducted in the 1970s and 1980s, while Skehan’s perspective has drawn more directly on the research in second language acquisition (SLA), particularly that on the effects of tasks on SLA.

Bachman (1990) and Skehan (1998) both point out that there have been two main approaches to solving this dilemma or conflict. One approach has been to “develop a model of underlying abilities” (p. 155), in Skehan’s terms, which is essentially what Bachman called the “interactive ability approach” (p. 302). The other approach, in Skehan’s terms, “is to bundle together the performance and contextual problems” (p. 155), which corresponds in essence to what Bachman called the “real-life approach” (p. 301). In the next section, I will attempt to demonstrate that language testers have indeed historically attempted to solve this dilemma or conflict by focusing almost exclusively on either ability/process or context/task in their approach to defining the construct to be assessed. Recently a number of researchers have proposed what I will call an “interactionalist perspective,” which attempts, at least implicitly I believe, to solve all three problems simultaneously. In my view, one variation of this approach is essentially an extension or operationalization of the interactional part of Bachman’s interactive ability approach, while the other variation is essentially an extension of performance assessment, in which the performance/context bundle is repackaged as interaction.

**Defining Constructs: Historical Overview**

One way to characterize the recent history of language testing research and practice is in terms of the ways in which it has defined the construct to be measured. From the early 1960s to the present, we can see a dialectic between a focus on *language ability* as the construct of interest and a focus on *task* or *context* as the construct. This dialectic is illustrated in Figure 3.1, in which the focus of the construct is boxed. The dashed arrows trace the shift of the construct from ability/trait to task/context. I would hasten to point out that these different approaches were not strictly ordered chronologically. On the contrary, there has been considerable chronological overlap among the different approaches, both in terms of the theoretical statements of their proponents, and as they were played out in practical test development and use.
### Approaches to defining the construct in language testing, 1960 to the present

<table>
<thead>
<tr>
<th>Construct</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Approach / Major Proponents</strong></td>
<td><strong>Ability / Trait</strong></td>
</tr>
<tr>
<td>1. <strong>Skills and elements</strong> (Lado, 1961; Carroll, 1961, 1968; Davies, 1977)</td>
<td>Elements/aspects/levels Integrated language skills</td>
</tr>
<tr>
<td>3. <strong>Pragmatic language testing</strong> (Oiler, 1979)</td>
<td>Pragmatic expectancy grammar</td>
</tr>
</tbody>
</table>
5. Interaction-ability  
   (Bachman, 1990; Bachman & Palmer 1996)

6a. Task-based performance assessment 1  
   (Brindley, 1994; McNamara, 1996; Skehan, 1998)

6b. Task-based performance assessment 2  
   (Norris et al., 1998; Brown et al., 2002)

7a. Minimalist interactionalist  
   (Kramsch, 1986; Chapelle, 1998; McNamara, 2001)

7b. Strong interactionalist  
   (He & Young, 1998; Young, 2000)

7c. Moderate interactionalist  
   (Chalhoub-Deville, 2003)
1. Skills and Elements

One of the first explicitly defined "models" for language testing was the "skills and elements" model that was articulated largely by Lado (1961) and Carroll (1961, 1968), and later by Davies (1977). In their formulations there was a clear distinction between skills and abilities, on the one hand, and approaches/methods/test types on the other. Lado (1961) described the "variables" of language to be tested as comprising pronunciation, grammatical structure, vocabulary, and cultural meanings. He pointed out that although these elements can be tested separately, they never occur separately in language use. Rather, he stated, "they are integrated in the total skills of speaking, listening, reading and writing" (p. 25). Lado clearly distinguished the construct to be tested—integrated skills and separate language elements—from the situations or types of tests that are used. Indeed, he seems to have anticipated the current debate about task-based performance assessment in his statement that "a situation approach that does not specifically test language elements is not effective. It has only the outward appearance of validity" (p. 27).

Taking a position similar to Lado's, with respect to defining the construct to be tested, Carroll (1961) described this in terms of "aspects of language competence" (phonology/graphology, morphology, syntax, and lexicon) and skills (auditory comprehension, oral production, reading, writing). He pointed out, however, that "it would be foolish to attempt to obtain these sixteen different measures, for this would be carrying the process of analysis too far" (p. 34). With respect to the method of testing, Carroll stated that it is desirable to test specific points of language knowledge with what he called a discrete structure-point approach. However, he recommended that for testing rate and accuracy in the four skills, an integrative approach is needed. Such an approach requires "an integrated, facile performance on the part of the examinee," in which "less attention is paid to specific structure-points or lexicon than to the total communicative effect of the utterance" (p. 37).

In a later article, Carroll (1968) adopted the Chomskian competence-performance distinction (Chomsky, 1957, 1965), retaining essentially the same components of linguistic competence—aspects and skills—as in his 1961 article. However, in the 1968 article, Carroll argued that "linguistic performance variables" also need to be taken into consideration. In the second part of this article, Carroll elaborated a taxonomy of language test tasks, characterizing these in terms of dimensions such as stimulus, response, modality, complexity, and task. In the final part of the article, Carroll directly addressed the issue of the interaction between task, competence, and performance by introducing the notion of critical performance, which characterized, for him, the necessary relationship between the test task, performance, and underlying competence:

If a language test is to measure particular kinds of underlying competence, its items must call upon language skills and knowledges in a critical way.
each task must operate in such a way that performance cannot occur unless there is a particular element of underlying competence that can be specified in advance. The extent to which this can be true is a function of the nature of the task and the specificity of its elements. (p. 67)

The skills and elements approach was perhaps the first approach in the history of language testing to explicitly draw upon both current linguistic theory and views of language learning and teaching. This approach to language testing was also the first to incorporate notions of reliability and validity from psychometrics. This approach thus provided a conceptual framework for defining the constructs to be tested that was based on linguistic theory and language teaching practice, and which was also in step with measurement theory at that time. This approach was extremely influential, and found its way into a whole generation of practical texts on language testing (e.g., Clark, 1972; Cohen, 1980; Davies, 1977; Finocchiaro and Sako, 1983; Harris, 1969; Heaton, 1975, 1988; Madsen, 1983; Valette, 1967). Equally importantly, this approach also informed a generation of large-scale assessments of foreign or second language in the United States. Early versions of these include:

- the Advanced Placement Examination in French (Educational Testing Service),
- the Comprehensive English Language Tests for Speakers of English as a Second Language (McGraw-Hill),
- the Michigan Test of English Language Proficiency (English Language Institute, University of Michigan),
- the Modern Language Association Cooperative Foreign Language Tests (Educational Testing Service), and
- the Test of English as a Foreign Language (Educational Testing Service).

2. Direct Testing/Performance Assessment

Spurred by an intense interest, in both research and practice, in the testing of oral proficiency, the 1970s saw the emergence, largely in North America, of another view of the construct to be tested in the so-called direct testing. The term performance assessment was also used to characterize this approach (e.g., Jones, 1979a, 1985b; Wesche, 1987). This approach, according to Jones, who was one of its proponents, constituted a major sea-change from the approach advocated by Lado. Nearly twenty-five years after Lado’s book, Language Testing, was published, Jones (1985a) wrote, “In Robert Lado’s book ... we were admonished not to measure a person’s speaking ability directly through a face-to-face test ... By comparison, direct testing today is becoming very commonplace” (p. 77, italics in original).
The published discussions of direct testing/performance assessment tended to focus primarily on the nature of the test tasks, which were claimed to mirror, or approximate real-life language use outside of the test itself (e.g., Clark, 1975). It was this replication of real-life language use tasks that was claimed to be the compelling evidence for the validity of such tests.

In defining the construct to be tested, proponents of direct testing/performance assessment referred to that as real-life language performance, which they viewed as the criterion for test tasks (e.g., Clark, 1972, 1979). Addressing the purpose of the Foreign Service Institute oral interview specifically, Clark (1972) states that this test “is intended to measure the adequacy with which the student can be expected to communicate in each of a number of language-use situations” (p. 121). For Clark, then, performance on a direct test was essentially a predictor of the language performance that could be expected of the test taker in real-life settings. Jones (1979a) took this definition of the construct a step further, explicitly identifying it with performance, and also emphasized the importance of prediction (Jones, 1985a). Scores from performance assessments were interpreted as predictions of future performance, and the evidence for validity lay primarily in the degree to which the test stimulus and the desired response, or both, replicated language use in real-life settings outside the test itself.

3. Pragmatic Language Testing

At about the same time as the direct approach was being promoted, but essentially independently of this, Oiller was conducting a program of factor analytic research that led to his unitary competence hypothesis (e.g., Irvine, Atai, and Oiller, 1974; Oiller, 1976, 1979; Oiller and Hinofotis, 1980). This hypothesis stated that language proficiency is essentially a single unitary ability, rather than separate skills and components, as had been proposed by Lado and Carroll. In the most extensive discussion of this research and the theory that underlay it, Oiller (1979) identified the general factor from his empirical research as “pragmatic expectancy grammar,” which he defined as “the psychologically real system that governs the use of a language in an individual who knows that language” (p. 6).

Having defined the ability to be tested, Oiller then discussed the kinds of tasks that are necessary to test this ability. In a way that echoed Carroll’s earlier notion of “critical performance,” Oiller described a “pragmatic test,” which he distinguished from both discrete-point and integrative tests, as “any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and which requires the learner to relate sequences of linguistic elements via pragmatic mappings to extralinguistic context” (p. 38). Examples of pragmatic tests were the cloze (gap-fill), dictation, the oral interview, and composition writing.
Oiler's conceptualization of the ability to be tested as a single, global ability, was, in my view, both simple and sophisticated. The notion of a single unitary ability meant that language testers did not need to concern themselves about testing the bits and pieces of language, while the notion of pragmatic expectancy drew upon current theory in both linguistics and pragmatics. Similarly, the types of tasks Oiler proposed, such as the cloze and the dictation, were appealing to practitioners, since they promised to be both valid and easy to construct. The research upon which Oiler's claims for a unitary competence were based was eventually rejected (see, for example, Bachman and Palmer, 1980, 1981; Carroll, 1983; Farhady, 1983; Upshur and Turner, 1999; Vollmer, 1980; Vollmer and Sang, 1983), and Oiler himself (1983) admitted that the unitary competence hypothesis was wrong. Nevertheless, Oiler's work has had a major and lasting impact on the field. In terms of language testing practice, his work was instrumental in reviving the use of the dictation and cloze as acceptable methods for testing language ability. His conceptualization of language ability as pragmatic expectancy grammar also foreshadowed later notions of strategic competence (e.g., Bachman, 1990; Bachman and Palmer, 1996; Canale and Swain, 1980).

4. Communicative Language Testing

At about the time the debates surrounding the oral proficiency interview and the unitary competence hypothesis were drawing to a close, or at least losing some of their heat, applied linguists in the U.S., Canada, and the U.K. began exploring a much broader view of language ability, drawing on a wide range of research in functional linguistics, sociolinguistics, discourse analysis, psycholinguistics, and language acquisition, as well as developments in communicative syllabus design and communicative language teaching.

Canale and Swain

Perhaps the first, and certainly one of the most influential papers to discuss the implications of this broadened view of language ability for language testing was the seminal article by Canale and Swain (1980). In this article, Canale and Swain adopted the term communicative competence to describe the ability that is of interest in both language teaching and testing. They defined communicative competence as "the relationship and interaction between grammatical competence, or knowledge of the rules of grammar, and sociolinguistic competence, or knowledge of the rules of language use" (p. 6). Canale and Swain explicitly distinguished communicative competence from communicative performance, which they defined as "the realization of these competencies and their interaction in the actual production and comprehension of utterances" (p. 6). In addition to grammatical and sociolinguistic competence, Canale and Swain posited a third component of communicative competence.
which they called \textit{strategic competence}, and defined as “verbal and non-verbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or insufficient competence” (p. 30).

With respect to the context, or kinds of tasks that should be used to elicit evidence of communicative competence, Canale and Swain indicated that “communicative testing must be devoted not only to what the learner knows about the second language and about how to use it (competence) but also to what extent the learner is able to actually demonstrate this knowledge in a \textit{meaningful communicative situation} (performance)” (p. 34; italics added).

Canale (1983) subsequently expanded the Canale-Swain framework in two ways. First, he added \textit{discourse competence}, which he defined as “mastery of how to combine and interpret meanings and forms to achieve unified text in different modes (e.g., casual conversation, argumentative essay, or recipe)” (p. 339). Second, he extended the function of strategic competence to include not only that of compensating for breakdowns in communication, but also “to enhance the rhetorical effect of utterances” (p. 339). Swain (1985) later refined the application of the framework to language testing by elaborating four “general principles of communicative language testing” (p. 36):

1. start from somewhere,
2. concentrate on content,
3. bias for the best, and
4. work for washback.

For Canale and Swain, then, the construct to be measured was clearly an ability or capacity that learners have. Their conceptualization of this capacity was much richer than those that preceded, and it initiated a major shift in the way language testers viewed the construct. They said very little about the specific contexts of tasks that should be used; rather, they suggested general principles for developing and using communicative language tests.

\textit{U.K. Symposia}

At about this same time, a group of applied linguists with interests in language testing met in two symposia, one in Lancaster in 1980 (Alderson and Hughes, 1981, p. 7), and one in Reading, in 1981 (Hughes and Porter, 1983, p. vii). Both of these symposia addressed three broad themes:

1. the nature of language proficiency,
2. communicative language testing, and
3. the testing of English/language for specific purposes.
General language proficiency. At the first symposium there was very little consensus about whether there is such an ability as general language proficiency, with one participant denying that it existed, another wondering why it was even important to attempt to research it, let alone test it, and yet another declaring it a non-issue. If there was any consensus, it was that Oiler's unitary competence hypothesis had generated both controversy and research, which was seen as positive, but that little was really known about language proficiency at the time, and that more research was needed. The discussion of general language proficiency in the first symposium was characterized largely by logical argumentation and not a little speculation, drawing on the experience and professional expertise of the participants. The papers at the second symposium, in contrast, focused largely on methodological issues about the way the empirical evidence supporting or rejecting Oiler's general language proficiency factor had been collected, or presenting fresh evidence that purported to reject it.

Communicative tests. The discussion of communicative language testing at the first symposium consisted of responses to and discussion of Morrow's (1981) paper, "Communicative Language Testing: Revolution or Evolution." The responses and discussion in this section focused on two issues:

1. what constitutes a communicative language test, in terms of the features of the test itself, and
2. whether communicative language tests measure anything different from previous, traditional types of tests.

On the first point, Morrow's position was that communicative tests would necessarily involve performance that is "criterion-referenced against the operational performance of a set of authentic language tasks" (p. 17). Morrow did recognize the problem of extrapolation, an issue that would also be a problem for performance-based and task-based approaches. Nevertheless, Morrow echoed the claim of the direct-testing proponents, stating that "a test of communication must take as its starting point the measurement of what a candidate can actually achieve through language" (p. 17; italics added).

The response papers and the discussion of Morrow's paper focused primarily on the issues of authenticity and extrapolation. In their responses, Weir (1981) and Alderson (1981) essentially rejected Morrow's notion of real-life authenticity as a criterion for language testing. Weir argued that this is unrealistic for language tests, while Alderson argued that language testing constitutes a domain of language use in its own right.

Weir and Alderson, in their responses, also clearly articulated the problems of extrapolation and sampling. Weir stated the extrapolation problem as follows: "A performance test is a test which samples behaviors in a single
Alderson points out the sampling problem as follows: "If one is interested in students' abilities to perform in cocktail parties, and one somehow measures that ability in one cocktail party, how does one know that in another cocktail party the student will perform similarly? The cocktail party chosen may not have been an adequate sample" (1981b, p. 57).

The theoretical discussion of communicative language tests at the second symposium consisted of a keynote paper by Harrison (1983) and a response by Alderson (1983). Harrison built his argument on an analogy between communicative tests and jam, beginning with the observation that, just as we should carefully consider the quality and contents of the jam we buy, so should we be wary of the tests we use. Drawing on the literature in communicative teaching, Harrison discussed several characteristics that should distinguish communicative tests, and a number of issues that he considered crucial in considering communicative testing.

In his response, Alderson (1983) begins by extending Harrison's jam analogy, suggesting that jam itself may not be a desirable product, and that one could well ask who needs jam, and by analogy, who needs communicative language tests. While he agrees with Harrison's warning not to accept as communicative any and every test that claims to be, Alderson goes on to systematically rebut most of Harrison's points.

The consensus view of these two symposia, in terms of the construct of interest in language testing, would seem to be more in terms of what it is not, rather than in terms of what it is. There was general agreement that the ability was not general language proficiency; what was to be tested was a set of areas of language knowledge and skills that interacted in complex ways in communication. What these areas of knowledge and skill were was not clear. On the issue of the context of language testing, there was general agreement that it was not real life, but something short of or different from that. Perhaps it could only be a representation of real life or perhaps language testing was its own context.

U.S. Symposia

At approximately the same time as the U.K. symposia were held, two symposia that also focused on testing communicative competence were held in the U.S. In 1979, what eventually became known as the first Language Testing Research Colloquium (LTRC) was held in Boston as part of the Annual TESOL Convention. The papers and discussions at this colloquium focused on methods, issues, and research in assessing oral proficiency/communication (Palmer, Groot, and Trosper, 1981). The empirical papers discussed results of studies into a variety of approaches to assessing oral proficiency. In my view, two things were particularly significant in this meeting. First, it brought language testers together...
with Michael Canale and Merrill Swain, who presented a short version of their 1980 paper, and thus set a research agenda for the next two years that would focus on efforts to empirically investigate the traits that were being measured by communicative language tests. Second, it introduced to the field of language testing, the research methodology of the multi-trait multi-method matrix (Campbell and Fiske, 1959), which would become, for the next few years, perhaps the dominant methodology in validation research in language testing. Bringing together a substantive theoretical framework of the construct to be tested, communicative competence, with a research methodology that was more sophisticated than the factor analysis of scores from miscellaneous language tests, provided the essential stimulus, in my view, for moving the field forward, beyond the unitary competence hypothesis and into the era of communicative language testing.

Another symposium on language proficiency and its assessment was held in Warrenton, Virginia, in 1981. Selected papers from this symposium that addressed challenging issues to language proficiency testing were published in Rivera (1984). The theoretical papers in the first part of this collection deal with the nature of communicative competence, how this relates to measurement models, and how integrative language proficiency tests could be improved by considering findings in communicative competence research. The papers in the application section of the volume include the results of empirical studies, or plans for such studies into a wide range of issues: the interdependency hypothesis of bilingualism that L1 and L2 proficiency are interdependent, the relationship between linguistic competence and communicative competence in a second language, the relationship of these to achievement in academic subjects, the effects of learners’ background characteristics on the acquisition of linguistic and communicative competence, and issues in identifying an appropriate educational program for L2 learners with primary learning disabilities.

In summary, the formulations of communicative competence and communicative language testing/assessment that can be found in these discussions demonstrate a clear movement, in the field of language testing. This shift was from the apparent certainty of what was to be tested and how to do this that had been claimed by proponents of the skills and elements, direct testing/performance assessment and pragmatic approaches, to a new level of awareness of the complexity of both the ability to be measured and the contexts or tasks in which it might be measured. With this heightened awareness came a good deal of uncertainty. Communicative competence as an ability or capacity that individuals have, was seen as both a break with and an extension of prior notions of language proficiency, as defined in the direct testing approach, or pragmatic expectancy, as defined by Oller. Although the Canale and Swain framework was reasonably well defined, the richness and variety of the
discussions summarized above illustrate the complexity of the notion of communicative competence, the lack of consensus on what it was, and the paucity, at that time, of solid empirical research into its nature. The contexts in which communicative competence could be assessed were vaguely conceptualized as meaningful communicative situations or authentic tasks. There was very little consensus as to what criteria could be used to identify such situations or tasks, especially if real-life tasks were the criteria for this. There was also a feeling by some that authenticity could not be achieved in a language test which, by its very nature, was artificial and inauthentic. It was definitely a new, if not so brave, world for language testers.

5. Interaction-Ability (Communicative Language Ability)

Bachman (1990, p. 113) conceived of performance on language tests as being a function of both an individual's language ability and of the characteristics of the test method. Nevertheless, he argued that in the design and development of language assessments, as well as in the interpretation of assessment results, it was essential to distinguish, analytically, the ability to be assessed from the assessment contexts or tasks in which language performance was observed. Further, he argued that it was essential to distinguish observable assessment performance from the unobservable abilities about which we want to make inferences.

In an attempt to address these issues in both the design and development of language tests, and in the interpretation and use of assessment results, Bachman proposed an approach that included two frameworks:

1. communicative language ability (CLA), and
2. test method facets.

Communicative language ability was essentially an extension of the Canale and Swain model, in which their notion of strategic competence was expanded from one that functioned essentially in accommodation and compensation to one that he hypothesized underlay all language use. He also reorganized their components of grammatical and sociolinguistic competence into organizational competence and pragmatic competence and elaborated these at lower levels of detail. Bachman (1990) saw test method facets as “analogous to the features that characterize the context of situation, or the speech event, as this has been described by linguists” (p. 111), and argued that these could “be seen as restricted or controlled versions of these contextual features that determine the nature of language performance that is expected for a given test or test task” (p. 112).

Bachman’s two frameworks were subsequently incorporated into an approach to practical test development by Bachman and Palmer (1996), who
renamed communicative language ability as simply *language ability* and the test method facets, *task characteristics*. Bachman and Palmer recognized, as did the proponents of performance assessment, that in order for score-based interpretations to generalize beyond the test itself, the characteristics of the assessment tasks needed to correspond to the characteristics of tasks in test takers' target language use (TLU) domains. Bachman and Palmer argued that by analyzing the characteristics of tasks in the TLU domain, test developers could use these sets of characteristics as templates for generating assessment tasks that would be representative of tasks in the TLU domain. The framework of task characteristics was thus seen as a way to solve the sampling problem of performance assessment, and to thus provide a stronger basis for making inferences to domains beyond the test itself. However, unlike performance assessment, the primary interpretation of test performance was about test takers' capacity for language use, rather than the prediction of future performance.

Bachman's and Bachman and Palmer's approach and frameworks provided richer descriptions of both the construct and the context than had previous approaches to language testing. However, even though both Bachman and Bachman and Palmer recognize and discuss language use in terms of interactions between ability, context, and the discourse that is co-constructed, their two frameworks are essentially descriptive, and provide little guidance for how they interact with each other in language use. Thus, while this approach may provide practical guidance for the design, development, and use of language tests, it does not solve the issue of how abilities and contexts interact, and the degree to which these may mutually affect each other.

6. Task-Based Performance Assessment

In the past decade, another approach to defining the construct to be assessed, *task-based performance assessment*, has been articulated. As Bachman (2002) has suggested, there are two very different conceptualizations of this approach in the language testing literature. The proponents of one version of this approach (Figure 3.1: 6a. in left column) draw on the literature in communicative competence/language ability, communicative language testing, language for specific purpose assessment, and educational assessment, and explicitly build upon previous approaches, specifically communicative language testing and communicative language ability. Proponents of the other version of this approach (Figure 3.1: 6b. in left column) draw more heavily on the research that has focused on the role of tasks in second language acquisition, and the efficacy of using tasks in language teaching (e.g., Candlin, 1987; Crookes and Gass, 1993a, 1993b; Long, 1985; Long and Crookes, 1992), and link themselves explicitly with performance assessment, with little or no link with either communicative language testing or communicative language ability.
a. Task-based performance assessment 1

One approach to language assessment has focused on the kinds of tasks that are presented to test takers, the kinds of processes these tasks engage, and the abilities that are assessed (e.g., Brindley, 1994; McNamara, 1996; Skehan, 1998). The basis for these discussions is the premise that the inferences we want to make are about underlying ability for use, or ability for language use. Thus, Brindley (1994) identifies both language knowledge and ability for use as the construct of interest (p. 75). Similarly, McNamara (1996) discusses construct validity in performance assessments, while Skehan (1998) is explicit that the inferences to be made are about an underlying ability, or what he calls an “ability for use” (Norris, Brown, Hudson and Yoshioka, 1998, p. 1; Brown, Hudson, Norris and Bonk, 2002, p. 1). With respect to the construct that is to be assessed, this version of task-based language assessment differs very little, in my view, from that proposed by Bachman (1990) and Bachman and Palmer (1996). That is, in both approaches, the construct to be assessed is an ability or capacity that resides in the individual, even though the specific details of this construct vary from one researcher to another.

b. Task-based performance assessment 2

Another approach to language assessment has been articulated and studied most extensively by researchers at the University of Hawai‘i at Manoa (Brown et al., 2002; Norris et al., 1998). The Hawai‘i group, who describe their approach as task-based performance assessment, see this as a special case of performance assessment. This approach draws on research that has focused on the role of tasks in second language acquisition, and the efficacy of using tasks in language teaching (e.g., Candlin, 1987; Crookes and Gass, 1993a, 1993b; Long, 1985; Long and Crookes, 1992). In this approach, the construct to be assessed consists of “students’ abilities to accomplish particular tasks or task types” (Brown et al., 2002, p. 9). The context or tasks which they propose as a basis for their approach consist of “the simulation of real-world tasks, associated with situational and interactional characteristics, wherein communication plays a central role” (p. 10). Test takers’ performances on these assessment tasks are evaluated “according to real-world criterion elements (e.g., task processes and outcomes) and criterion levels (e.g., authentic standards related to task success)” (p. 10).

The most salient difference between these two task-based performance approaches to language assessment lies not in the kinds of assessment tasks that are used, but rather in the kinds of inferences their proponents claim they can make on the basis of test takers’ performance on assessment tasks. While the task-based performance assessment 1 approach aims at providing inferences about an ability or abilities that test takers have, the task-based performance assessment 2 approach aims primarily at making predictions about
future performance on real-world tasks. The claims of task-based performance assessment approach are essentially the same as those of the direct testing/performance assessment approach of the 1970s and '80s, discussed above. Specifically, the ability is performance on "real-life-like" tasks, and the assessment tasks are selected to "be as authentic as possible with the goal of measuring real-world activities" (Norris et al., 1998, p. 9).

To summarize, the different approaches to defining the construct that have been discussed thus far can generally be seen as focusing on either an ability or abilities that test takers have, or on the types of tasks that test takers can perform. The approaches in Figure 3.1 that have boxed entries in the "Ability/Trait" column define the construct in terms of areas of language ability that test takers have, while those with boxes in the "Task/Content" column define it in terms of what test takers can do in contexts beyond the test itself. According to Upshur (1979), defining the construct as what test takers can do limits our interpretations to predictions about future performance. Defining the construct as what test takers have, on the other hand, can potentially tell us something about the nature of the ability itself.

7. Interactional Approach to Language Assessment

The last approach that I will discuss is the social interactional perspective, which has been articulated by several different researchers. Working largely within the area of the assessment of interactive speaking, and drawing on a variety of research literatures outside of language assessment, these researchers have identified a number of problems and lacunae in current conceptualizations of the construct, oral language ability, and how we go about assessing it. They present different, albeit overlapping, perspectives and suggest two general types of implications for language assessment: the need to rethink the way we define the attributes of participants in language assessments, and the way we define and operationalize the contexts in language assessment.

The way we define what we assess

Kramsch (1986), in a discussion and critique of the theoretical beliefs that underlie the notion of proficiency, as operationalized in the ACTFL Proficiency Guidelines, is generally credited with the first use of the term interactional competence. Drawing on the research literature in psycho- and socio-linguistics, Kramsch stated that "the oversimplified view of human interaction taken by the proficiency movement can impair and even prevent the attainment of true interactional competence within a cross-cultural framework and jeopardize our chances of contributing to international understanding" (p. 367). Although Kramsch does not provide an explicit definition of interactional competence in this article, she defines it obliquely by stating what successful interaction presupposes "not only a shared common knowledge of the world, the reference
to a common external context of communication, but also the construction of a shared internal context or "sphere of inter-subjectivity" (p. 367). She elaborates this further, arguing that "learning a foreign language ... entails not only language but also metalanguage skills in the foreign language, such as the ability to reflect on interactional processes, to manipulate and control contexts, to see oneself from an outsider's point of view" (p. 369).

Chapelle (1998), in an analysis of the relevance of construct definition and validation research to SLA research, draws on the research literature in validity theory in educational measurement to discuss three different approaches to construct definition—trait, behaviorist, and interactionist. Chapelle begins by defining a construct as "a meaningful interpretation of observed behavior" (p. 33), arguing that we base this interpretation on performance consistency, and that "the problem of construct definition is to hypothesize the source of performance consistency" (p. 34). The interactionist approach, which is of relevance here, explains performance consistency in terms of "traits, contextual features, and their interactions" (p. 34). Chapelle then goes on to discuss these three perspectives in detail, along with the implications they have for measurement, illustrating these with the example of the construct, interlanguage vocabulary. Chapelle argues that the interactionist approach to construct definition must "specify relevant aspects of both trait and context" (p. 43). However, the interactionist construct is not simply the sum of trait and context. Rather, "when trait and context dimensions are included in one definition, the quality of each changes. Trait components can no longer be defined in context-independent, absolute terms, and contextual features cannot be defined without reference to their impact on underlying characteristics [of language users or test takers]" (p. 43). For Chapelle, what is essential for making an interactionist construct definition work is a component that controls the interaction between trait and context. This component, she further argues, is essentially what Bachman (1990) defined as strategic competence and Bachman and Palmer (1996) called metacognitive strategies. For Chapelle, then, an "interactionist construct definition comprises more than trait plus context; it also includes the metacognitive strategies (i.e., strategic competence) responsible for putting person characteristics to use in context" (p. 44). She cites Bachman's (1990) definition of communicative language ability, which consists of "both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualized communicative language use" (p. 44), as an example of an interactionist approach to construct definition.

Read and Chapelle (2001) describe a framework for vocabulary testing that draws on and extends their earlier work (Chapelle, 1994, 1998; Read, 2000). They present a framework for vocabulary testing that relates test purpose to validity considerations, from test design to validation. They then
illustrate this framework with examples from the three different approaches to construct definition: trait, behaviorist, and interactionist (Chapelle, 1998). Of particular relevance here is their application of Chapelle’s (1998) definition of the interactionist approach, with respect to vocabulary knowledge: “an interactionalist approach to inferences requires that vocabulary knowledge and use should be defined in relation to particular contexts” (p. 22; italics in original). They also discuss the implications for test use: “a new approach to test uses means going beyond tests designs to measure learners’ knowledge of relatively decontextualized word lists and considering what other vocabulary assessment needs have to be met” (p. 22; italics in original).

He and Young (1998) and Young (2000) adopt Kramsch’s (1986) term, interactional competence, and extend or refine it in terms of its components and how it operates in interactive speaking. He and Young (1998) begin with a discussion of assessing “how well someone speaks a second language” (p. 1), reaching the rather unsurprising conclusion that “defining the construct of speaking ability in a second language is in fact a theoretically challenging undertaking” (p. 2). Under the heading “Interactional Competence,” they begin by describing speaking ability as “a subset of the learner’s overall ability — or proficiency — in the language” (p. 3). Thus far, it seems that interactional competence is, indeed, an individual characteristic. However, they then state that “abilities, actions and activities do not belong to the individual but are jointly constructed by all participants” (p. 5; italics in original), which appears to mean that these abilities are not individual attributes. He and Young identify Kramsch’s term, “interactional competence,” with Jacoby and Ochs’ (1995) notion of co-construction, which is essentially an interactive process by which cultural meanings are created. So now interactional competence appears to be a process. He and Young (1998, pp. 5-7) then describe the resources that participants bring to a given interactive practice:

1. knowledge of rhetorical scripts,
2. knowledge of certain lexis and syntactic patterns specific to the practice,
3. knowledge of how turns are managed,
4. knowledge of topical organization, and
5. knowledge of the means for signaling boundaries between practices and transitions within the practice itself.

He and Young state that “participants’ knowledge and interactive skills are local: they apply to a given interactive practice and either do not apply or apply in different configuration to different practices” (p. 7). They then go on to argue that although participants’ knowledge and interactional skills are local and practice-specific, they “make use of the resources they have acquired in previous instances of the same practice” (p. 7; italics in original). Thus, “individuals
do not acquire a general, practice-independent communicative competence; rather they acquire practice-specific interactional competence by participating with more experienced others in specific interactional practices” (p. 7).

Young (2000) pushes the definition of interactional competence further in the direction of its being a characteristic of discursive practice, rather than of individual language users. “Interactional competence … comprises a descriptive framework of the socio-cultural characteristics of discursive practices and the interactional processes by which discursive practices are co-constructed by participants” (p. 4). He contrasts interactional competence with the Canale-Swain (1980) framework. He characterizes the former as being based on a constructivist, practice-oriented view of interaction and competence, while the latter, Young argues, focuses on the individual language user (p. 5). He argues that the theory of interactional competence is characterized by four features:

1. a concern with language used in specific discursive practices,
2. a focus on co-construction of discursive practices by all participants,
3. a set of general interactional resources that participants draw on in specific ways to co-construct a discursive practice, and
4. a methodology for investigating a given discursive practice.

The resources that participants bring to a discursive practice are a recasting of the five (see previous page) that He and Young (1998) describe. Young (2000) again emphasizes the local nature of participants’ knowledge and interactional skills, and that these are “distributed among all participants in a discursive practice” (p. 10).

My reading of He and Young (1998) and Young (2000) is that they seem to vacillate between conceptualizing interactional competence as an ability (i.e., resource) that individual participants bring to an interactional practice, on the one hand, and as an attribute of interactional practice that is locally co-constructed and shared by all participants, on the other. Similarly, the resources that participants bring to a discursive practice are both general resources that appear to be essentially aspects of or expansions of language ability (e.g., Bachman and Palmer) on the one hand, yet localized, in that participants tailor them to a particular interactional practice, on the other.

I would argue that this conceptualization of interactional competence as resources that participants bring to discursive practice is essentially the same as that of language ability as an attribute of individual participants. I would argue further that the notion that the competence is itself co-constructed and shared by participants, and context-bound, or local to a specific context, is highly problematic and not adequately supported by the research that He and Young (1998) and Young (2000) cite. Nevertheless, I believe that their perspective and the issues they raise provide an important contribution to how our
WHAT IS THE CONSTRUCT?

BACHMAN

Conceptualization of language ability can be enriched, both in terms of what the various components are and in terms of how language ability interacts with specific contexts.

Chalhoub-Deville (2003) and Chalhoub-Deville and Deville (2005) provide probing and insightful discussions of the issues involved in an interaction-based construct definition. Chalhoub-Deville begins her forward-looking overview of the field by echoing Douglas's (2000) somewhat disheartening yet, in my view, accurate assessment of the current state of our field in terms of how we define the construct we want to assess: "while theoretical arguments and empirical evidence have ascertained the multidimensionality of the L2 construct, consensus is absent regarding the nature of these components and the manner in which they interact" (p. 370). She then spends several pages deconstructing Bachman's (1990) conceptualization of communicative language ability (CLA), correctly pointing out that this, by itself, is essentially a "psycholinguistic ability model" of "cognitive, within-user constructs" (pp. 370–371). Referring to the discussion in Chalhoub-Deville and Deville (2005), she contrasts the cognitive-psycholinguistic approach of CLA with a view of the L2 construct as "socially and culturally mediated" (p. 371). She proposes a construct, "ability-in-individual-in-context," which, she argues, represents "the claim that the ability components that a language user brings to the situation or context interact with situational facets to change those facets as well as to be changed by them" (p. 372). Chalhoub-Deville then discusses the social interactional perspective, arguing that this poses two challenges to language assessment: (1) "amending the construct of individual ability to accommodate the notion that language use is . . . co-constructed among participants, and (2) the notion that language ability is local, and the conundrum of reconciling that with the need for assessments to yield scores that generalize across contextual boundaries" (p. 373).

Conceptualizing and operationalizing context

As with several of the approaches discussed above, proponents of an interactional approach define the context of assessment in different ways. Kräsmacher (1986) defines the context holistically as collaborative activity, while McNamara (2001, 1997) conceptualizes the assessment task or context in terms of characteristics of the interaction. Other researchers focus on the criteria for evaluating performance. Thus, Chalhoub-Deville (1995), Upshur and Turner (1999), and Fulcher (1996) discuss task-specific rating scales, while Jacoby and McNamara (1999) argue that we should consider the characteristics of indigenous assessments in developing criteria for rating performance.

McNamara (2001) discusses language assessment in its social context and articulates two of the interactionalist's main points. First, he argues that we need to reconceptualize the construct, language ability, recognizing that this
is a social construction and that the way we define it embodies social values. Second, he argues for a richer conceptualization of context as dynamic rather than static. He correctly notes that "our existing models of performance are inadequately articulated, and the relationship between performance and competence in language testing remains obscure. In particular, the assumption of performance as a direct outcome of competence is problematic, as it ignores the complex social construction of test performance" (p. 337). Thus, while McNamara argues for a richer definition of the construct, he nevertheless clearly sees competence as an attribute of the individuals who interact with assessment tasks in performance assessment. Finally, McNamara challenges language testing researchers, particularly in the context of classroom assessment, to expand the notion of assessment, and he suggests two specific areas for further research and development:

1. "greater research emphasis on the implementation of assessment schemes, including an analysis of the impact of assessment reforms and a critique of their consequences;" and
2. more adequate theorizing and conceptualizing of alternative, more facilitative functions of assessment in classrooms, which would involve "expanding our notion of assessment to include a range of activities that are informed by assessment concepts and that are targeted directly at the learning process." (p. 343)

Some Unresolved Issues Raised by an Interactionalist Approach

Just as the proponents of interactional competence or ability-in-individual-in-context have correctly pointed out the limitations of other approaches to defining the construct, most notably those of Bachman's (1990) and Bachman and Palmer's (1996) conceptualization of language ability, the interactional approach is not without its unresolved issues.

The relationship between interaction and language ability. It seems to me that the differences among proponents of an interactionalist approach to defining the construct can be characterized in terms of the claims they make about the relationship between interaction and language ability. He and Young (1998) and Young (2000), in my view, identify interaction, or discursive practice, with the capacity or ability to engage in such practice. Even though they discuss the resources that participants bring to an interaction, these resources nevertheless are characterized as local and co-constructed by all participants in the discourse. This, I believe, constitutes the strongest interactionalist claim: the interaction is the construct. Cha!houb-Deville (2003) argues that language ability interacts with and is changed by the context and the interaction. This view, in which the ability and context are distinct, with the ability changing
as a result of the interaction, is a moderate interactionalist claim: the ability is affected by the interaction. A third claim is articulated by Chapelle (1998), who sees the capacity for language use (trait plus metacognitive strategies) as distinct from but interacting with the context to produce performance. In other words, for Chapelle, performance, or language use, is a product of both the ability and the context. While Chapelle notes that “the context dimension of an interactionalist definition must provide a theory of how the context of a particular situation . . . constrains the linguistic choices a language user can make during linguistic performance” (p. 45), she stops short of claiming that the ability is changed by the interaction. This, it would seem, is the minimalist interactionalist claim: the ability interacts with the context.

The strong interactionalist claim raises, it seems to me, some thorny issues. As Messick (1989) and Chapelle (1998) have pointed out, whatever our perspective (trait, behaviorist, interactionist), what testers generalize from and attach scores and meaning to are consistencies in performance across a range of assessment tasks. Bachman (2006) makes essentially the same point about empirical research in applied linguistics in general. Thus, we might begin by asking about the source of the performance consistencies that enable researchers, whether in language assessment, discourse analysis, or SLA, to generalize. If the construct is strictly local and co-constructed by all of the participants in the discursive practice, this would imply that each interaction is unique. If so, what performance consistencies, if any, would we expect to observe from one interaction to the next? If we put the same participants in a different context, will their performances share any consistent features? Or, if we put different participants in the same context, what features, if any, will their performances share? If there are no consistencies in performance across contexts or participants, then we have no basis for generalizing about the characteristics of either. This problem has been pointed out by Chalhoub-Deville (2003), with specific reference to language assessment: “If internal attributes of ability are inextricably enmeshed with the specifics of a given situation, then any inferences about ability and performance in other contexts is questionable” (p. 376). She points to this as a challenge for language testers, to reconcile the local nature of language ability with the need for assessments that generalize across contexts (p. 373). If, on the other hand, there are performance consistencies, where do these come from? Since each discursive practice is uniquely co-constructed, performance consistencies cannot arise from the interaction itself. To what, then, are these attributable? Are these due to the attributes of the participants or the features of context? It would thus appear that this confounding of the roles of ability and context in interaction is problematic for generalizing, whether there are performance consistencies or not. If there are no consistencies in performance, any attempt to generalization is suspect. If there are consistencies, we are unable to explain or interpret them.
A second issue raised by the strong interactionalist claim is its identification of language use and language ability. The research literature upon which this claim is based comes largely from the various approaches to the analysis of oral discourse (e.g., sociolinguistics, conversation analysis, speech act theory, ethnography). In this research, the focus is clearly on language use rather than on the language abilities of the participants in the conversation or discursive practice. Young (2000) and He and Young (1998), for example, draw on the research in areas such as conversational analysis, linguistic anthropology, sociolinguistics, and speech act theory, all of which focus clearly on the speech event, the interaction, language use. In this regard, the same criticism could be applied here as Tarone (2000) applied to SLA researchers who take a narrowly “sociolinguistic or co-constructionist orientation”:

while . . . [they] have a good deal of evidence showing that L2 learner’s IL [Inter Language] use is variably affected by identifiable features of the social context, they have usually not tried to show that those social features change the process of L2 ACQUISITION — specifically, the acquisition of an IL system — in any clear way. They have assumed it, and asserted it, but not often accumulated evidence to prove it. (p. 186)

The moderate interactionalist claim that language ability is changed by interaction raises, in my mind at least, questions about the generalizability and relevance of the research upon which the claim is based. Chalhoub-Deville (2003) draws largely on the literature in learning and cognition in building a case for the construct as ability-in-language user-in-context. One perspective on learning that she discusses is that of situated/reflective learning, or what Sfard (1998) refers to as the participation metaphor, according to which learning is seen as a set of ongoing activities that are “never considered separately from the context within which they take place” (p. 6). Sfard contrasts this with the acquisition metaphor, which views learning as “gaining possession over some commodity” (p. 5), such as knowledge, concepts, or skills. If we want to generalize from the research in learning to language use or discursive practice, the first question that needs to be asked, I believe, is how strong a consensus there is in the field of learning about the nature of learning. If one can judge by debates in the literature (e.g., papers in Resnick, 1993) and recent overviews (e.g., Hofer and Pintrich, 1997; Saloman and Perkins, 1998; Sfard, 1998), it is clear that there is still considerable debate on this issue in the field of learning. Therefore, I would question the extent to which this research supports the moderate interactionalist claim. Although Chalhoub-Deville does not draw on the literature in SLA, she might well have cited research suggesting that acquisition varies as a function of context. Tarone (2000), for example, discusses a number of studies in SLA that suggest that differing social contexts affect what gets acquired and how it gets acquired. However, the research cited by Tarone
was all conducted with L2 learners, so one must therefore ask whether the effect of interaction on the language acquisition of language learners differs from its effect on the language use of language users with native-like language ability. Thus, with both the literature in learning, in general and with SLA, I would argue that we need to question the extent to which metaphors for learning generalize to performance or use.

Chalhoub-Deville also draws on the literature on situated cognition, which "focuses attention on the fact that most real-world thinking occurs in very particular (and often very complex) environments ... and exploits the possibility of interaction with and manipulation of external props" (Anderson, 2003, p. 91). But as with language use and language ability, one might well ask what the relationship is between cognition and knowledge, concepts, and so forth. That is, if cognition is an activity that is highly situated, what, if anything is the product of this activity. Furthermore, as with situated learning, within the various fields that constitute the cognitive sciences there appears to be considerable debate about the nature of cognition (e.g., Anderson, 2003; Roth, 1998). Thus, to draw on this literature to support the moderate interactionalist claim again would appear questionable.

And what about what I've called the minimalist interactionalist claim? As with proponents of the strong and moderate interactionalist claims, Chapelle (1998) also discusses a number of challenges that an interactionalist approach poses, not only for language testers but for SLA researchers as well. For example, she points out that content analysis, as an empirical validation method, "requires that the person and context sources of learner's performance in the operational setting be hypothesized, but the analytic procedures for making such process-oriented hypotheses have not yet been developed" (pp. 64–65). Similarly, empirical item analysis requires that the researcher operationalize both context and person variables, as well as their interaction, but such operational definitions have not been specified (p. 65). Finally, she refers to Messick's measurement conundrum: "strategies can vary across people and tasks even when the same results are achieved" (p. 65). I will return to this last point below as something that differentiates language testers as practitioners from language testers as researchers.

I have argued that the research that has been cited in support of the strong and moderate interactionalist claims is either controversial in the fields from which they are drawn or of questionable relevance to these claims. I have also argued, as have the proponents of all three types of interactionalist claims, that the research evidence in support of any of these claims, in the context of language assessment, is scanty, if not non-existent. What I have not argued is that these claims are wrong. I have pointed out what I believe are some unresolved issues with these claims, as have their proponents. Chapelle (1998), Young (2000), McNamara (2001), and Chalhoub-Deville (2003) all discuss
challenges that an interactionalist approach poses for language assessment, and these challenges imply a research agenda for future language testers. On this point I am in complete agreement with them, and list in the following section, some unanswered questions that might help focus such research. These are not, of course, entirely new questions, and some research has been conducted on virtually each of them. What has not happened, I believe, is for these questions to be investigated carefully in the context of language assessment.

Some Implications for Language Assessment

Each of the different approaches to defining the construct of interest that have been discussed above have drawn on research literatures outside language testing. However, I believe that it is the interactionalist approach, which has drawn on research approaches and perspectives that are in the starkest contrast to those generally associated with language assessment, that poses the most serious and interesting challenges for the field. But in order to appropriately translate these challenges into implications for language testing, I believe that we must first distinguish two roles of language testers that Bachman (1990) pointed out: the language testing practitioner and the language testing researcher. The fact that the majority of language testing researchers are also practitioners does not lessen the importance of the difference between these two roles.

Language Testers as Researchers and as Practitioners

Perhaps the most important distinction between the roles of language testing researcher and practitioner is that of purpose, or goal. The language testing researcher’s goal, I believe, is to better understand, *inter alia*, the psychological and contextual factors that affect performance on language assessments, the types of language use that language assessments elicit, the relationship between language use elicited in assessments and that created in *real-life* settings, and the relationship between the abilities engaged in language assessments and those engaged in *real-life* settings. I would argue that the goal of the language testing practitioner, on the other hand, is to design and develop language assessments that are useful for their intended purposes. In either role, I believe that it is essential that we clearly define what it is we want to measure or what we want to investigate.

Because of these differing purposes, practitioners and researchers may investigate different constructs, may define the constructs of interest with different degrees of specificity, and may investigate differing ranges of constructs. When we language testers wear our researcher hat, we are essentially applied linguists who are seeking to expand our knowledge. What we choose to investigate, or observe, as well as how we define this, how we choose to observe it, and how we interpret it, will be influenced by a number of different
dimensions, including our perspective on knowledge, our purpose, where we place ourselves in defining constructs and contexts and the relationship between these, and our view of the world (Bachman, 2006). As researchers, we are interested in both the how and why questions. For example:

- How do the features of a graphic prompt influence the kinds of oral language test takers produce, and why?
- How do test takers interact with and respond differently to different types of tasks based on a reading passage, and why?
- How do ESL teachers in elementary schools interpret and use standards-based assessments, and why?
- How do test takers with different personality traits interact with each other in a group oral discussion assessment, and why do they interact this way?

More often than not, we can find tentative answers to the how questions, and are left to speculate on the why. Thus, as researchers, we have the luxury of investigating not only constructs defined in very precise ways but also a large range of these in a single study. We can also investigate constructs that we only hypothesize to be relevant to test performance. Finally, the impact of our research is generally minimal, typically being limited to other researchers, journal reviewers, reviewers of grant proposals, and tenure and promotion committees.

As language testing practitioners, on the other hand, we work under a very different set of constraints. First of all, we may be held accountable for the decisions that are made on the basis of our test, and the higher the stakes of the decisions, the greater the burden of accountability. Thus, if the decisions that are made on the basis of the test will have a major impact on the lives of test takers (high stakes), then the language testing practitioner must collect considerable evidence to support these decisions. Because of this burden of accountability, language testing practitioners must deal with known constructs that may be defined very broadly. Thus, while a researcher might be interested in the internal construction of meaning formed by interacting with a written text, the practitioner is more likely to be interested in reading comprehension. Part of the accountability equation involves practicality, so the language testing practitioner must generally focus on constructs that are directly relevant to the decisions that are to be made. Thus, while the researcher might be able to investigate the effects of different personality types on oral test performance, the practitioner may not have the resources, such as time, personnel, expertise, or funding, to include the assessment of personality type in his assessment. While the researcher may have resources to collect verbal protocols from test takers and to analyze the discourse they produce, the practitioner will seldom
have the resources for this, or be able to justify using them in this way. In using the measurement models and statistical tools needed to provide accountability, the practitioner must generally work with interpretations that are group-based. Thus, even though research tells us that each test taker may approach the same reading passage and questions differently, interact with these differently, and draw on different areas of knowledge, processes, or strategies in responding to the questions, the practitioner must assume that the scores of test takers on this reading test can all be interpreted in the same way: level of reading ability. The language testing practitioner is thus interested in the what and how much questions. For example:

- What does this test score tell us about what test takers know or can do?
- How much of this do different test takers have, or how well can different test takers do it?
- What kind and how much impact will our intended decisions have on stakeholders?

When the language testing practitioner investigates these questions, he is attempting to answer questions about the reliability of the scores, the validity of the interpretations, and the fairness and appropriateness of the decisions that are made. He is preparing to be held accountable by stakeholders.

**Implications for language testing researchers**

The approaches to language testing that have been discussed above have been based not only on theoretical perspectives drawn from other disciplines but also on empirical research. If we were to examine this research closely, I believe we would see that both the types of research questions that have been investigated and the specific research approaches that have been used derive largely from differences in several dimensions of research. These dimensions are discussed by Bachman (2006), and although these are framed broadly within empirical research in applied linguistics, I believe that they also apply to research in language testing. The lesson to be learned here, I believe, is that the issues and questions in language testing research are far too complex, and the perspectives that are involved far too diverse to admit to any doctrinaire positions, with respect to either what the true construct is or what the correct methodological approach is.

Consideration of the historical dialectic in language assessment between abilities and contexts, and of the interactionalist perspective in particular, also raises a host of interesting questions for language testing researchers. I list a few that come to mind below.
1. What cognitive/neurobiological/socially constructed abilities/predispositions/resources do language users bring to an interaction and what cognitive/neurological residue do they take away with them?

2. To what extent can we distinguish resources for interactional competence (e.g., knowledge of rhetorical scripts, knowledge of register, knowledge of how to take turns-at-talk) from other resources that have been discussed in the literature (e.g., knowledge of grammar, lexicon, cohesion, rhetorical organization)?

3. To what extent and how does the effect of interaction on the language ability of L2 learners differ from its effect on the language use of language users with native-like language ability?

4. How does interaction for the purpose of acquiring a language differ from interaction for communication, socialization, acquiring, and creating knowledge, etc.?

5. Chalhoub-Deville (2003) points out that “task specificity, i.e., inconsistent performance across tasks, is well-documented in the literature” (p. 378). What is the role of variable contexts in interactional competence or ability-in-individual-in-context? “Where does performance inconsistency across contexts come from?”

Implications for practical test development

In my view, the most important implication for practical test development of this historical perspective on approaches to language testing is that the theoretical frameworks (linguistic, psychological, cognitive, affective, neurobiological, sociological, ethnographic, etc.) that researchers draw upon are typically too broad and complex to apply to all test development contexts. That is, the complexity and breadth of theoretical frameworks generally render these unsuitable for practical test development (e.g., cognitive processing models of reading for developing assessments of reading comprehension). These theories also present the testing practitioner with a specification dilemma, in that they may be underspecified in certain ways, and overspecified in others, for application to a specific language testing mandate. For this reason a given theoretical framework may apply only partially or poorly to any particular test development effort. Finally, many of the theoretical frameworks that researchers draw upon are developed within a philosophical perspective that uses falsifiability as a criterion, while for any particular language assessment, I have argued, the appropriate criteria are: 1) the cogency of the assessment use argument that informs it, and 2) the quality of the evidence that supports this argument (Bachman, 2005).

For all of the reasons above, I would propose that for purposes of practical language test development we need to develop local theories, or what
Chalhoub-Deville and Tarone (1996) refer to as “operational models” (p. 11). I would suggest that an assessment use argument (AUA), such as that described by Bachman (2005), constitutes, in essence, a local or operational theory for any given test development project. Bachman (2005) describes an AUA as “the overall argument linking assessment performance to use (decisions)” (p. 16), and argues that an AUA will guide both the design and development of language assessments and validation, which he characterizes as the process of collecting evidence in support of a particular score interpretation or use. For Bachman, then, an AUA is essentially a local or operational theory that guides both the design and development of a specific assessment, and the collection of evidence in support of a specific intended use.

Consideration of the ways in which the different approaches to defining the construct have influenced practical test design, development, and use over the past half century also raises a host of questions for practitioners. At the top of this list, for me as a language testing practitioner, at least, would be a set of questions that follow from the research questions listed above. For each of these questions, the practical follow-up question would be something like, “If we knew the answer to this, would this make a difference in how we design, develop and use language assessments?” “If so, how?” “If not, why not?”

Conclusion

Issues related to language ability and language use contexts and the interaction between these have been addressed, in a dialectic, in language assessment research, and have led to three general approaches to defining the construct, or what we want to assess: (1) ability-focused, (2) task-focused, and (3) interaction-focused. While the different theoretical perspectives that underlie these approaches are not mutually exclusive, they are based on different sets of values and assumptions. These, in turn, have derived largely from the differing research milieux or Zeitgeisten in which they were formulated, the hopefully cumulative experience of the field over time, and from individual differences among researchers in their ontological stances, perspectives, and purposes, and how they define the phenomena that are the focus of their research. Because of these differences, the conundrum of ability and context and how they interact in language use and language assessment is, in my view, essentially a straw issue, theoretically, and may not be resolvable at that level.

Nevertheless, the theoretical issues raised by these different approaches have important implications and present challenging questions for both empirical research in language testing and for practical test design, development, and use. These theoretical issues also provide valuable insights into how we can enrich the ways in which we conceptualize what we assess and how we go about assessing it. For research, they imply the need for a much broader, more catholic methodological approach, involving both so-called quantitative
and qualitative perspectives and methodologies. For practice, they imply that exclusive focus on any one of these approaches (ability, task, interaction), to the exclusion of the others, will lead to potential weaknesses in the assessment itself, or to limitations on the uses for which the assessment is appropriate. This means that we need to address all three in the design, development, and use of language assessments.

Notes

1 Although Carroll (1973) discussed several “persistent problems” that he believed would “continue to exist and to challenge our best efforts,” he touched on this particular problem only peripherally in his discussions of problems of “validity and realism” and of “scope.”

2 In this paper, I will focus on research and theory in language testing since the early 1960s. For discussions that take both longer and broader historical perspectives on language testing, see Barnwell (1996) and Spolsky (1995).

3 See Develle and Chalhoub-Deville (2005) for an excellent historical overview of these approaches, from a slightly different perspective.

4 Another perspective is that of Spolsky’s (1978) well-known three trends or approaches: pre-scientific, psychometric-structuralist and integrative-sociolinguistic. Spolsky (1995) takes a different perspective, viewing language testing in the broad context of historical developments in large-scale institution testing, particularly from the mid-1940s onward, with focus on the development of the TOEFL and the Cambridge EFL tests.

5 McNamara (1996) and Skehan (1998) characterize this differently, as a distinction between “construct-based” and “task-based” approaches to language assessment. However, it seems to me that the critical issue is how we define the construct to be assessed— as ability or as task.

6 A variety of terms have been used by different authors for this general approach to assessment. Although McNamara (1996) uses the term performance assessment, he does characterize earlier work in performance assessment (e.g., Clark, 1972) as a task-centred approach. Skehan (1998) uses the term performance testing, but discusses ways of testing task-based performance in terms of a processing or task-based approach, which he appears to use more or less synonymously. Brindley (1994) uses the term task-centered. The University of Hawai’i group (e.g., Brown et al., 2002; Norris et al., 1998) appear to consider task-based assessment as an approach to performance assessment, while Brown et al. (2002) use the term task-based performance assessment.

7 These were originally published as the ACTFL Proficiency Guidelines (1983) and revised in 1985. They are currently available on the Web at: www.sfl.org/lingualinks/LANGUAGELEARNING/OtherResources/ACTFLProficiencyGuidelines/contents.htm

The most recent versions of the guidelines for speaking and writing can be downloaded from www.actfl.org/i3a/pages/index.cfm?pageid=3318.