An Operational Framework for Constructing a Computer-Adaptive Test of L2 Reading Ability: Theoretical and Practical Issues

BY MICHELINE CHALHOUB-DEVILLE, CHERYL ALCAYA, AND VASHTI MCCOLLUM LOZIER

CARLA Working Paper #1

CARLA | Center for Advanced Research on Language Acquisition

University of Minnesota

An Operational Framework for Constructing a Computer-Adaptive Test of L2 Reading Ability:

Theoretical and Practical Issues

Prepared by:

Micheline Chalhoub-Deville, Ph.D.

Cheryl Alcaya

Vashti McCollum Lozier

The Center for Advanced Research on Language Acquisition

University of Minnesota

An Operational Framework for Constructing a Computer-Adaptive Test of L2 Reading Ability: Theoretical and Practical Issues

First Edition

Produced by

Center for Advanced Research on Language Acquisition
University of Minnesota
140 University International Center
331 17th Ave SE
Minneapolis, MN 55414
USA
612.626.8600
carla@umn.edu
http://www.carla.umn.edu

University of Minnesota proficiency testing

In 1986, the second language (L2) programs at the University of Minnesota made a difficult but necessary transition from allowing undergraduate students to fulfill their second language requirement on the basis of seat time, i.e., the completion of the required courses, to requiring students to demonstrate their proficiency by passing proficiency-based tests. The tests were developed by L2 teachers based on the ACTFL Proficiency Guidelines. Today, more than ever, the University of Minnesota remains committed to an accountable approach in assessing students' L2 proficiency and is working hard to update these tests and to testing situation by targeting each student's ability level, and by providing diagnostic feedback.

As part of the Title VI National Language Resource Center (NLRC) grant, the Assessment team at the Center for Advanced Research on Language Acquisition (CARLA) has been involved for the last two years in evaluating the quality of the existing proficiency tests in French, German, and Spanish (see Chalhoub-Deville, Alcaya, Klein, Lozier, and Budlong, 1996; Chalhoub-Deville, Mueller, Lozier, and Juengling, 1996; Chalhoub-Deville, Sweet, Schmidt, and Lozier, 1996; and Lozier and Chalhoub-Deville, 1995). Additionally, the Assessment team, in conjunction with the Minnesota Articulation Project members, L2 teachers from secondary and postsecondary institutions around the state, has been developing new tests that reflect current theory and research and technological advances. These new tests are intended to replace the existing University of Minnesota tests and to be used for articulation purposes by the various foreign language programs in the State of Minnesota. For example, in terms of assessing speaking proficiency, we have developed multiple forms of a simulated oral proficiency interview (SOPI) instrument. With regard to writing proficiency, the newly developed tests incorporate a process approach. As for reading, plans have involved the development of a theory-based computer adaptive test (CAT). The rationale for a CAT is the facility to administer the test on a large scale, and at the same time, to optimize the testing situation by targeting each student's ability level, and by providing diagnostic feedback.

The computer adaptive reading test project

Building on the University of Minnesota tradition in proficiency testing, we applied for funding to the U.S. Department of Education's International Studies and Research Program to develop CATs for assessing students' reading proficiency in French, German, and Spanish at three levels:

(1) an entry level, for those students finishing high school and entering the university;
(2) an exit level, for those students finishing the university L2 requirement; and
(3) at a level appropriate for those university students finishing an L2 major.

The Department of Education funded our proposal, and work on the project began in October, 1995. In the initial phase of the project, we are building the assessment framework that will be used in the development of our CATs.

The purpose of the present document is to review the literature in order to develop the assessment framework that will help inform the selection of the appropriate reading proficiency texts and the construction of the test items. The document also addresses issues pertaining to the measurement model and the test method characteristics.

We have recently also received funding from the U.S. Department of Education Title VI National Language Resource Center program to develop computer based reading tests for diagnostic purposes. We foresee using the diagnostic tests in conjunction with the proficiency tests. Work on the diagnostic tests will begin next fall.

An approach to building an assessment framework

In the present paper, the focus is on proficiency testing. Defining an operational framework for assessing reading proficiency is our first task. There are various theoretical models of first language (L1) and L2 reading ability, as well as general models of L2 ability to help in defining our assessment framework. Unfortunately, as researchers (Chalhoub-Deville, in press) and others, (Tarone, in press) have shown, theoretical models approach constructs from a global, all-encompassing perspective, and may not be applicable, in whole or in part, to specific testing

contexts. While we will rely on theoretical models to build our assessment framework, we are keen that their applicability be weighed in light of the particulars of our testing situation. Thus, we need to adapt these theoretical models to accommodate the specifics of our testing context, and to explain what functional L2 reading ability is for our population, how we propose to measure it, how reading texts are selected, what item types are deemed appropriate, what scores on our measure mean, and how the scores will be used.

To arrive at an assessment framework that provides an operational representation of L2 reading ability as per our testing context, we begin with the broader question of what comprises language ability and then we narrow our focus to L1 and L2 reading research. We then discuss models of L1 reading comprehension that have played an important role in L2 reading investigations. Finally, we consider L2 reading research.

Componential models of language ability

The meaning of "language ability" has evolved as a result of the inputs from the various perspectives that have a stake in defining it. The rather loose term "language ability" is used here because, during the course of the debate among linguists, sociolinguists, psycholinguists, psychologists, language pedagogists, and so on, the terms "competence" and "proficiency" have been used in so many ways that their meanings are no longer clear. "Competence" is probably most closely associated with the linguist Noam Chomsky, who defined linguistic competence quite narrowly as the internalized knowledge of

> the ideal speaker-hearer, in a completely homogeneous speech community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of language to actual performance (Chomsky, 1965, p. 3).

By contrasting competence to performance, he placed competence on a theoretical level which he acknowledged was of little use to L2 practitioners for the purposes of teaching and research (Chomsky, 1973). Chomsky's view of competence is also unidimensional, since it accounts only

for grammatical knowledge. While this narrow focus works well for structural linguistics, communication in the real, day-to-day situations in which humans find themselves involves pragmatic elements of language use. Hymes (1971) first used the term "communicative competence" to denote an integrated concept accounting for both underlying knowledge of a linguistic code and language use for communicative purposes within a community. Hymes' multidimensional conception of communicative competence includes elements of linguistic, cultural, and sociolinguistic knowledge, as well as cognitive, physical, and environmental constraints on communication. It should be noted, however, that Hymes did not use the above terms to identify components of language ability; instead he spoke in terms of judgments that language users make in communicative situations involving what is formally possible, what is feasible, what is appropriate, and what is actually performed.

The communicative competence model was further refined over the years, and the reader is advised to consult a source such as Savignon (1983) or Omaggio (1986) for a thorough discussion of the many contributions made by numerous scholars. One of the most-cited definitions of communicative competence is perhaps that of Canale and Swain (1980), later revised by Canale (1983), who identified four subcompetencies: linguistic competence, sociolinguistic competence, discourse competence, and strategic competence. These four competencies, while not the last word in the quest for a language ability model, continue to contribute to the debate to a large extent.

As we can see, an important focus of L2 proficiency research has been the identification of the components that comprise language ability. Canale and Swain's communicative competence model is one of a number of such componential models. As Savignon (1983) points out, however, identification of components is not enough; we must understand that communicative competence lies in the interaction of the components, and that the relative importance of any one component is highly variable depending upon the language user's ability and willingness to perform in a given situation. The highly interactive nature of the relationships among elements of communicative competence may help to explain in part the findings

supporting Oller's (1983) one-dimensional model. Oller's unitary competence hypothesis (UCH) proposes that there is one general factor that accounts for language proficiency. Nevertheless, Oller's findings have been largely refuted as artifacts of the nature of the data and the analytical methods used; but, as we shall see below in our discussion of reading models and reading research, there is some evidence that it is challenging to empirically identify the theoretical components of a skill such as reading.

A more complex, and recent, componential model is that of Bachman and Palmer (forthcoming). The breadth and adaptability of this model, along with its empirical base, render it particularly useful for test development. Bachman and Palmer discard the term "competence" because "of all the semantic baggage that term has acquired in the fields of linguistics and applied linguistics over the years" (forthcoming). They call the model "A Model of Language Use" (MLU). The MLU is a further refinement of Bachman's (1990) Communicative Language Ability model, which is probably more widely known at present. The MLU incorporates and expands previous conceptions of communicative competence while insisting upon the contextual and interactional nature of language use. The five factors of Bachman's model are language knowledge, metacognitive strategies, knowledge schemata, affective schemata, and characteristics of the language use context. These five factors are elaborated upon briefly below, but the reader is advised to refer to Bachman and Bachman and Palmer (1990; forthcoming) for a thorough discussion and graphic representations of the model. Table 1 displays the different elements of the model in outline form to help the reader see more clearly the relationships among the elements.

Language knowledge: Bachman's language knowledge is generally synonymous with "competence" in other models. It is subdivided into two kinds of knowledge: organizational and pragmatic. Organizational knowledge includes grammatical (phonology/graphology, morphology, syntax) and textual knowledge (cohesion, rhetorical/conversational organization). Pragmatic knowledge includes lexical, functional, and sociolinguistic knowledge, which might be summed up as knowing what language to use for what purpose in a given situation.

- LANGUAGE ABILITY
  - Language knowledge
    - Organizational knowledge
      - grammatical knowledge
        - phonology/graphology
        - morphology
        - syntax
      - textual knowledge
        - cohesion
        - rhetorical/conversational organization
    - Pragmatic knowledge
      - lexical knowledge
        - semantic properties
        - denotation
        - connotation
      - functional knowledge
        - ideational
        - manipulative
        - heuristic
        - imaginative
      - sociolinguistic knowledge
        - conventions of language use
        - dialect/variety
        - register
        - naturalness
  - Metacognitive strategies
    - assessment
    - goal-setting
    - planning

- LANGUAGE USE
  - Characteristics of the language use context
    - propositional
    - functional
    - sociolinguistic
  - Schemata
    - Knowledge schemata
      - world knowledge
      - cultural knowledge
      - topical knowledge
    - Affective schemata

Table 1: Bachman and Palmer's Model of Language Use

Metacognitive strategies are fully integrated and interact with the language knowledge areas identified above. The metacognitive strategies include assessment, goal-setting, and planning. Assessment refers to sizing up the communicative situation to determine what you need to meet its requirements, what you have at your disposal to do so, and assessing your performance. Goal-setting means determining your communicative goal. Planning signifies choosing a route by which you accomplish your goal.

Knowledge schemata refer to the world and cultural knowledge that a language user possesses.

Affective schemata are the feelings individuals have toward a language use context or topic.

Language use context: The characteristics of the language use context can be grouped into propositional, functional, and sociolinguistic features. These categories indicate the content of the discourse, its purpose, and its appropriateness.

One of the strengths that Bachman and Palmer's theoretical model provides is a flexible interpretation of the language ability construct that recognizes the need to accommodate contextual differences. Bachman and Palmer write:

> the way in which we define language ability and specify test tasks depends on the specific purpose for which the test is intended and the intended test-takers. Thus, in designing a language test we begin by constructing a definition of language ability that is appropriate for the intended use and test-takers. We then operationalize this definition in the form of test tasks that are also appropriate for the intended purpose and test-takers. Thus, the construct, language ability, is likely to be operationally defined in different ways for different testing situations (p. 3).

Because of its flexibility for use in specific contexts, we will consider Bachman and Palmer's MLU in the selection of the salient components of the construct that need to be measured in our context. In addition to considerations of which components are appropriate to include in our assessment framework, we must determine the feasibility of operationalizing these components given our testing method. Currently, the various computer adaptive software relies to a large

extent on using recognition types of questions such as multiple choice. With such item types, it is still not feasible to measure the effect of test-takers' use of metacognitive strategies on comprehension. While other research methods, such as interviews, think-alouds or recall protocols, can provide insights into this very important area of L2 text comprehension, our testing method does not allow us to operationalize metacognitive strategy use as a measurable variable of the language ability construct.

Hierarchical models/evaluative scales

In designing an L2 assessment framework, we must also consider the very important role evaluative scales have played in the debate. Most specifically, we are referring to the ACTFL Proficiency Guidelines (1986), which were derived from the Foreign Service Institute (FSI) scales but modified to accommodate the academic context. The Guidelines generally restrict themselves to components that can be identified in language output or other demonstrations of skill. The "functional trisection" identifies a content, function, and accuracy component in the level descriptors of the Guidelines. The trisection might be summarized as a description of the topics and purposes of language the user can handle and with what degree of accuracy.

It should be noted, however, that scale writing appears to be a widespread activity. Spolsky notes that "the FSI scale has been transported to Europe, too, and interpreted there in a number of different forms" (Spolsky, 1995, p. 350). Such scales are also popular in Australia (e.g., Ingram, 1979; Ingram and Wylie, 1982, 1984). In addition to scales devised by organizations having national or international scope in language teaching and testing, there are scales designed for local use (e.g., Guide to ESL Levels for the Community Literacy Collaborative, St. Paul, Minnesota, 1995) and scales developed by teachers for classroom use. These evaluative scales aim to classify learners' performance into the various levels of their speculated or proposed hierarchy. The question, however, is whose scale is the "right" or "best" scale? These scales are largely experientially based and lack empirical validation. To date, results

of attempts to validate the ACTFL Proficiency Guidelines are inconclusive, since there have been findings both for aspects of their validity (Dandonoli and Henning, 1990; Kaya-Carton and Carton, 1986; Kenyon, 1995a; Kenyon, 1995b) and against (Allen, Bernhardt, Berry, and Demel, 1988; Bernhardt, 1991; Lee and Musumeci, 1988).

North (1993) assembled a list of the pros and cons of evaluation scales for language ability. The attractions of such scales lie in their unifying properties. Widely known and used scales like the ACTFL or FSI scales can provide a common yardstick whose levels are meaningful to diverse bodies needing to classify language ability. These evaluative scales can also influence language teaching and testing to be more coherent within and across systems, e.g. articulation projects in Minnesota (Metcalf, 1995), Ohio (Corl, Harlow, Macián, and Saunders, 1996), and the six New England states (College Entrance Examination Board, 1996).

North, who expands the list compiled by Brindley (1991), points out several problems with evaluative scales. Some of these problems include having very little information about how the descriptors were arrived at, their circular logic, their imprecise language (e.g., "some", "a few", "several", etc.) and the seemingly unprincipled allocation of tasks and cognitive operations to levels that does not accommodate second language acquisition findings such as backsliding or variability. In addition, such evaluative scales typically measure learners' ability against that of the generic, "homogeneous" native speaker (NS). The generic representation of the NS has been challenged based on both theoretical arguments and empirical evidence (Brown, 1995; Chalhoub-Deville, 1995a, 1995b; Elder, 1996).

The ACTFL Proficiency Guidelines are difficult to relate to the componential models, primarily because theoretical models remain at a level of generalizability, while the Guidelines address details of performance along a hierarchy of ability levels. Higgs and Clifford (1982) hypothesized a relationship between the levels on the proficiency scale and the relative role played by various language ability components. In their Relative Contribution Model (RCM), they hypothesized that for any given level on the scale there exists a particular mix in the degree

of importance that the components of vocabulary, grammar, pronunciation, fluency, and sociolinguistic knowledge play in a speaker's performance. Clearly, this restricted group of components encompass a subset of the components identified in other models (e.g. Bachman and Palmer).

The ACTFL proficiency guidelines provide a useful perspective for defining L2 reading. For our assessment framework therefore, we would consider, in addition to Bachman and Palmer's Model of Language Use, the Proficiency Guidelines. Before we describe how we have reconciled the Proficiency Guidelines with the MLU in our assessment framework, we review the literature on reading models and research.

Reading models and research

The field of reading research is vast. Much attention is devoted to reading development in children, the psychomotor operations required for reading, the steps in the reading process, the linguistic and cognitive components of reading comprehension, reading disabilities, and so on. In L2, the characteristics of the population are generally different. In our context, most L2 students are adults who have L1 proficiency, so in our review we will not focus on the literature that deals with first language acquisition or literacy skills. Instead, we will review the literature that sheds light on L2 learners' characteristics and the L2 reading process. Therefore, in addition to L2 studies, we consider L1 work that has influenced how the L2 reading construct has been defined.

L1 reading models have evolved over the years from linear, step-by-step, models to interactive models accounting for constant interplay between the textual, physical, cognitive, and social aspects of reading. Through the advancement and refinement of these models, a great deal has been learned about what it means to read. Grabe (1988) writes that it is necessary to distinguish two kinds of "interaction" in discussions of reading research. On the one hand, there is the interpretive process which occurs between the reader and the text. This interaction

describes a process by which the reader applies his or her knowledge of the world to derive meaning from the text, i.e., top-down processing. Reading is a continual process of hypothesizing and confirming or disconfirming hypotheses, and has been called a "psycholinguistic guessing game" (Goodman, 1967; Smith, 1971). On the other hand, there is interaction in the "processing relations among the various component skills in reading" (Grabe, 1988 p. 59). This second representation of interaction describes the process by which the reader performs lower-level functions involved in decoding (bottom-up processing) while simultaneously applying world knowledge and higher-level functions, such as inferencing, to interpret and learn from the input (again, top-down processing). Models that focus upon this bottom-up/top-down interaction, which have grown out of the work of Rumelhart (1977), McClelland and Rumelhart (1981), and Lesgold and Perfetti (1981) are often referred to as Interactive Parallel Processing models.

Upon carefully studying the Interactive Parallel Processing models, Grabe makes several observations that document the similarities and differences between L1 and L2 reading:

- Lower-level processes such as letter and word recognition are largely automatic in L1, and so do not demand much attention, allowing the reader to attend to higher-order features of the text. For the L2 reader, letter and word recognition processes may not be automatic.
- Linguistic features of text present few problems for good L1 readers, but they do inhibit the reading process of L2 readers.
- A large receptive vocabulary is necessary for reading proficiency; L2 readers generally lack extensive vocabularies.
- Good readers are most easily identified by superior lower-level skills, i.e., speed and accuracy of word recognition operations. Readers do not perform these operations with equal efficiency in L1 and L2.
- L2 readers, like poor L1 readers, may compensate for poor lower-level skills by using more guessing and inferencing strategies.
- Higher-level strategies, such as applying one's world knowledge to a context, may impede comprehension of texts with which the L2 reader does not share a cultural background.

Eskey (1988) also prefers the Interactive Parallel Processing models to the top-down models of Goodman and Smith for application to L2 reading contexts. The recognition that L2

readers need to "hold in the bottom", that is, rely heavily on bottom-up strategies, renders the top-down models only partially useful to understanding L2 reading processes.

Bernhardt (1991) has shown that research in L2 reading has been limited both in quantity and scope. In fact, Bernhardt, like Grabe and Eskey, regrets the over reliance the field of L2 reading seems to have placed on the Goodman/Smith psycholinguistic (top-down) models of reading, when these models are seen as less important in L1 reading research. Bernhardt has offered a developmental, multidimensional model which is specific to L2 reading, and goes beyond previous attempts to make L1 reading models fit the L2 reading context by attaching certain provisos. Bernhardt's model is research based, and as such is linked to a particular research methodology. Bernhardt used error rates and error analysis of recall protocols to develop her model, and one of her assumptions was that "errors in understanding can reveal development in literacy (parallel to observed phenomena in oral language development)" (p. 169). Two other assumptions recall important elements of Higgs and Clifford's Relative Contribution Model: first, abilities (here, text processing abilities) develop over time, and, second, at different points along the developmental continuum, the various facets of language processing abilities play differently weighted roles in relation to each other. Bernhardt explains, "this presupposes an interactive, multidimensional dynamic of literacy elements—not a linear one in which each element is gradually replaced by the next" (p. 169). A fourth assumption states that there is no end point in text-processing development—no one is ever 100 percent proficient.

Although Bernhardt insists, and her data vividly illustrate, that the various elements of the model are inseparable, these elements are singled out in order to describe them. (The inseparability of the elements, as we shall see later, is a critical issue in determining the empirical unidimensionality of the construct.) The elements she proposes are described in two categories: text-driven and conceptually driven categories. There are three text-driven factors and three conceptually driven factors.

Text-driven factors

1. Word recognition denotes the "attachment of semantic value to a word by translation or conjecture."
2. Phonemic/graphemic decoding involves the identification of words. Errors in decoding would be mistaking one word for another due to spelling or phonetic similarities.
3. Syntactic feature recognition means interpreting the relationships among words.

Conceptually driven factors

1. Intratextual perception is concerned with "how the various parts of the text are perceived and reconciled with each other."
2. Metacognition involves readers thinking about and evaluating their comprehension of the text.
3. Prior knowledge refers to what the reader already knows about the world and/or the particular subject of a text.

By analyzing the errors readers made in recall protocols, Bernhardt was able to posit a distribution of the role five of the six above features play in causing readers to misunderstand text as their reading proficiency develops. Metacognition was not included because it cannot be measured in terms of error in recalls. In this theoretical distribution, low proficiency readers owe the majority of their errors to lack of skill in recognition of words and phono-graphemic features. Readers of mid-scale proficiency have greatly reduced the number of errors they make due to word and phono-graphemic feature recognition skills, but syntactic features of texts play a very large role in their misunderstanding of texts. Readers at the high-proficiency end of the scale have greatly reduced their number of syntactic errors, but it is still the most important factor. Background knowledge and intratextual perceptions play a reduced role in misunderstandings as proficiency increases, but the downward slope of these curves is quite gentle. The reasons why these conceptually driven features affect error probably differ along the proficiency continuum. Low-proficiency readers might have to use a great deal of conjecture about the subject of a text given the low level of their decoding and word recognition skills. As these bottom-up, text-driven factors contribute less and less to reading errors, however, background knowledge and

the process of establishing intratextual connections may cause readers to create incorrect interpretations based upon syntactic features they cannot process.

Defining the construct: An example and discussion

Kaya-Carton and Carton (1986) attempted to simplify the reading proficiency construct for purposes of test development by using a "partial model of reading proficiency". The authors intended to ignore any reader-based description in the ACTFL Guidelines and focus to a large extent on the ACTFL text characteristics when selecting texts and developing items. The salient text characteristics in this partial model were: structure/syntax, semantic and pragmatic content, topical reference, level of formality, literal vs. interpretive presentation, and rhetorical organization (Barnett, 1989). Kaya-Carton and Carton reasoned that reader characteristics come into consideration during the item calibration process. This exclusive emphasis on the text brings to mind Bernhardt's legitimate concern regarding whether such an approach identifies proficient texts rather than proficient readers (Bernhardt, 1986). Also, Barnett (1989) writes that Lee (1987, 1988) questioned the decision by Kaya-Carton and Carton to use only a partial model when they had also posited a full model. The full model included, in addition to the text characteristics listed above, the following reader characteristics: cognitive ability, linguistic knowledge, personal and cultural experience, and general knowledge (Barnett, p. 57).

Indeed, it appears, based on a subsequent article, (Kaya-Carton, Carton, and Dandonoli, 1991) that the authors revised their original plans when it came time to select texts for their test. They report a different representation of the reading ability construct from the partial model originally posited: "In this project, reading proficiency was defined as the degree of meaningful interaction between the reader and a sample of text, with the reader and the text contributing a multitude of complex factors that influence the act of reading and comprehending" (pp. 261-262). We feel that we, too, need to conceive of text along the lines of the full model in order to account for the assumptions we make about L2 readers in interaction with text. While our

criteria may not differ markedly from those in Kaya-Carton and Carton's full model, our depiction of the relationship between elements may be unique.

A synthesis of the review

We have briefly discussed several perspectives on language and reading ability. The next step is to reconcile these perspectives in order to identify the components for our assessment framework. One challenging issue in trying to fit together the various models reviewed here is that each model has developed its own terminology, and the scope and meanings of terms vary. For example, Bernhardt's "prior knowledge" probably covers a wide area corresponding to Bachman's "knowledge schemata" as well as his "functional knowledge" of the different roles performed by text and his "sociolinguistic knowledge" governing conventions of language use. A very telling example is demonstrated in Bernhardt's analysis of students' recall protocols of a German business letter. Without the background knowledge to understand the layout of business letters (e.g., where the return address appears), the functions of such letters, and their formal conventions and register, Bernhardt shows that students had great difficulty understanding the text.

Another challenging issue is capturing the depiction of the construct at various levels. The ACTFL Proficiency Guidelines, the Higgs and Clifford Relative Contribution Model, and Bernhardt's theory of second language reading all speak to the developmental nature of L2 ability. In these developmental perspectives, it is assumed that, in general, while the language components are similar at the various ability levels, the relative importance of these components differs at different points along the ability continuum. In a sense, then, development in reading ability is defined not in terms of diverse components but in the different emphasis placed on these components at various ability levels.

Finally, in identifying the components of language ability to be measured, it is equally important, as Bachman and Palmer (forthcoming) write, to identify the elements that will be

excluded from measurement. To illustrate their point, Bachman and Palmer offer the example of a reading comprehension assessment. If the assessment's aim is to measure language ability, and not the test-takers' knowledge schemata with regard to given topics, the content of reading passages must be familiar and accessible to a general population. This example fits very well with our testing situation where we need to identify and include those components pertinent to our testing context and to exclude those, such as topic knowledge, that are not.

Based on the review of the literature, we might sum up our definition of second language reading ability for our reading proficiency assessment context by way of the following assumptions:

- The components of reading ability contribute to the skill in different proportions along the proficiency continuum, but we hypothesize that for any group of individuals at the same point along the continuum, the relative importance of the components is the same. The interaction among these components is indivisible, i.e., reading ability at any given level is best characterized as a holistic construct rather than as a skill made up of discrete, measurable types of knowledge and strategies.
- Our adult readers are familiar with a great variety of text types, genres, and functional characteristics of text in their L1. They also possess a wide range of reading strategies in L1. While we cannot predict the transfer of knowledge about text and reading strategies to L2 for individuals, we cannot assume that low-proficiency readers can work with only certain kinds of text (e.g., "menus, schedules, timetables, maps, and signs") or perform certain kinds of tasks (e.g., scan for individual words). The ACTFL Guidelines are helpful in selecting texts and tasks to use for various proficiency levels, but final determination of the appropriateness of the texts and the corresponding items will be based on empirical evidence.
- We wish to assess reading ability rather than background knowledge. It is not possible to exclude knowledge schemata from the reading process, but their effect on scores may be minimized. For this reason, texts selected will deal with non-specialized, general interest topics, or will have all information required to understand the text contained within it. Also, tasks must focus on language processing and avoid questions that general knowledge and/or common sense could answer.
- Readers at all levels make inferences to connect portions of text and to connect text to their knowledge schemata. Taking into account readers' language knowledge, inference questions can be appropriate at all levels.

Defining the assessment framework

The description of our assessment framework is grouped under two major headings: ***a text selection model*** and ***task criteria***.

A text selection model

Based on our review of language ability and reading models and research, we have developed an operational model for text selection from a combination of elements of Bachman and Palmer's model of language use, Bernhardt's L2 reading model, and the ACTFL Proficiency Guidelines. The model is illustrated in Figure 1.

Our assessment framework consists of four continua that are likely to contribute to text difficulty: text types (wide availability - limited availability), the content (topics, cultural distance), the organizational characteristics (structural and rhetorical complexity), and the pragmatic features (lexicon, function, sociolinguistic factors). Within the space delineated by these four categories (depicted in Figure 1 as the interior of the cube), are the approximate areas where the continua would intersect for *targeting* texts to proficiency levels for assessment purposes in accord with the ACTFL Guidelines. However, by placing these characteristics along continua, we are attempting to show that, theoretically, any one of them can vary independently of the others. In addition, we posit that any one continuum is not inherently hierarchical, and thus the arrows go in both directions. We contend that hierarchies may result from a variety of combinations of characteristics along the continua, which we intend to observe closely during the test development and pilot testing process. A more detailed discussion of how we derived the elements of our model follows.

Figure 1

Text content

The text content corresponds to the "Language Use" portion of Bachman and Palmer's model, i.e., the propositional characteristics of the text (topic) and the schemata (world, topical, and cultural knowledge) imputed to the reader in order to understand the text. We represent this topic-schemata combination along the continuum of general interest topics-specialized topics, as suggested by the ACTFL Guidelines. Unlike the Guidelines which specify a hierarchy that links topics and proficiency based on experiential evidence, we do not make specific links between topics and proficiency levels. Such links will be determined based on students' performance.

How to represent the cultural content of text was debated at length: a topic could be an "easy", general interest topic, but lack of cultural knowledge would render it incomprehensible for both high and low level readers. The texts in the Steffensen, Joag-Dev, and Anderson (1979) study describing an Indian and an American wedding are examples of this issue. We considered including a second component to the continuum consisting of cultural distance between the native and target cultures, but decided to consider cultural content within the concepts of general interest topics and specialized topics.

The ACTFL Guidelines do not address the ability to comprehend target culture-specific information until the Advanced-Plus level, implying the inability of lower-level readers to comprehend cultural content, an assumption not made here. While we recognize that the higher the reading ability of learners, the more likely they are to have culture specific information, we hypothesize that it is also feasible for the lower level reader to be familiar with cultural content.

At the Advanced-Plus level, the ACTFL Guidelines also introduce the ability to deal with conceptually abstract topics. We have omitted a concrete-abstract component on our continuum because we believe that these terms represent a constellation of factors (e.g., lexicon, function, text type, etc.) that cannot be captured under the heading of content and are therefore subsumed by the other headings in the model.

Text types

Text types, on the top of the cube, are interpreted from specific and general references to text in the level descriptors of the ACTFL Guidelines and are displayed in the approximate order in which they appear in the ACTFL proficiency hierarchy. Although the specific claims made with regard to the hierarchical nature of text type have been challenged (see Swaffar, Arens, and Byrnes, 1991 for a review), and while we reject any explicit *a priori* links of text types to levels of reading ability, we will investigate relationships among categories of text type and reading passage difficulty during the item development phase of our project. For example, while it is probably the case that most articles in academic journals are inappropriate for testing at the Intermediate-mid level, analysis of a text of this type by the characteristics described by the other three continua might predict that it would be accessible for Intermediate-mid readers. Conversely, a text that treats a generic topic, is structurally non-complex, and contains no difficult vocabulary might prove difficult because the reader needs knowledge of the conventions of a particular text type in the target culture to understand it. Bernhardt (1991) illustrates this case in the difficulties learners had with a German business letter, and Lee and Musumeci (1988) found similar results with a restaurant receipt.

Pragmatic features

The pragmatic features correspond to Bachman and Palmer's pragmatic knowledge categories of language ability. These features address the lexical, functional, and sociolinguistic knowledge readers possess. Texts will be more or less comprehensible to readers because of the vocabulary present in the text, the transparency of the text's purpose, and sociolinguistic factors such as text conventions, register, language variety, etc. We also maintain that Bachman and Palmer's model provides, with its representation of a separate pragmatic component, a more appropriate way to group these features.

The ACTFL Guidelines make specific reference to vocabulary at the Novice level, but thereafter the lexical aspects of text are largely subsumed by text type descriptions. The

Guidelines address the sociolinguistic and functional features of text within text type, for the most part. For example, it is implied that the low ability reader will understand the purpose and conventions of the category of text called "messages". A message may deal with a technical issue that necessitates certain lexical, functional, or sociolinguistic features not recognizable to a particular reader, rendering the message text type with unchanging pragmatic features not plausible. The separation of the text type and pragmatic features is necessary and sound.

Organizational characteristics of text

The organizational characteristics of text are represented along a non-complex-complex continuum. We include in this category both grammatical knowledge and textual knowledge from Bachman and Palmer and references from the ACTFL Guidelines to grammatical structure, text structure, length, subordination, etc.

Task criteria

A reading test has two principal sources of difficulty which interact with reader characteristics: text difficulty and task difficulty. As discussed above, some texts, for a combination of reasons, are more difficult to understand than others. Tasks, too, vary in difficulty. Davey (1989) writes that

> it has long been recognized that students' responses to questions are influenced not only by text variables and content understandings, but also by ancillary task features of the test itself, and that these features may interact with certain individual differences among readers (p. 694).

Therefore, in addition to investigating the role that texts play in assessing reading, we also need to consider the task.

The literature is rich with studies that document the effect of different tasks on students' test scores. Studies such as those by Bachman and Palmer (1981), Clifford (1981), Henning (1983); Shohamy, (1983, 1984), Shohamy, Reeves and Bejarano (1986), Wolf (1993), and Chalhoub-Deville (1995a) show that different tasks influence students' performance, and so,

scores, differentially. Variability can be attributed largely to the different demands that the task places on the linguistic and cognitive processes of the learners, thus, influencing their performance. In measuring students' reading ability, it is our responsibility as test developers to understand the characteristics of the various tasks. Such understanding can help us minimize or explain the pitfalls and biases of the chosen task while interpreting and using test scores. It is important to note that because CAT requires on-line scoring, only selected response types of items are appropriate for our purposes and will be the type of items we have in mind when discussing task features.

In the literature, we do find some degree of consensus with regard to the comprehension abilities that tasks can elicit. To explain, we focus on the work of Swaffar et al. (1991), Langer (1985), and Barrett (1972). Swaffar et al. (1991) cite Schallert, Ulerick, & Tierney (1984): "'the meaning and structure of a text are not inherent in the print but are invited by the author and imputed to the text by the reader'" (p. 22). Swaffar et al. go on to identify the kinds of imputations the reader can make:

> (1) conceptualizing of explicitly stated information, (2) conceptualizing of intentionality created by the author's structuring of that information, and (3) conceptualizing of the significance that the author's message system has for the reader. The first factor, textual assertion, is verifiable in the text. The second factor, inferences to be drawn from explicit language, links text- and reader-based information. The third factor, significance, is verifiable only as a reader-based component (pp. 22-23).

In Swaffar's terms, textual assertion and inference have some objective reality that it would be appropriate to assess in order to verify comprehension.

In the literature, we find a similar three-level model in Langer (1985), but levels are attached to the notion of question difficulty, which is not a feature of Swaffar's description of reader imputations. Langer writes: "At the present time, questions for the assessment of reading comprehension are usually generated to reflect three levels of difficulty: (1) literal (factual), (2) interpretive (inferential), and (3) evaluative (applicative)" (p. 587) Langer suggests that these three categories are derived from Bloom's (Bloom, Engelhart, Furst, Hill, & Kratwohl, 1956)

taxonomy of cognitive functioning. Moreover, Langer states that, according to this taxonomy, these categories "are seen to progress in complexity from knowledge (facts and definitions) to comprehension (paraphrase, infer, imply), to application, to analysis, to synthesis, and finally to generalization or evaluation" (p. 587). Comparing these categories to Swaffar's factors, it can be claimed that the first two categories, i.e., literal and interpretative, correspond to Swaffar's text-based factors and Langer's third category of evaluation corresponds to Swaffer's reader-based category.

Barrett (1972) proposes four categories to represent a model of comprehension for teaching reading skills. They are: (1) literal recognition or recall of explicitly stated information, (2) inference, a synthesis of the literal content of the text, personal knowledge, intuition, and imagination as a basis for conjectures or hypotheses, (3) evaluation, making judgments, and (4) appreciation. The tasks Barrett identifies under the headings of evaluation and appreciation would largely resemble Swaffar's third factor, in that they would be reader-based only. In addition, Smith and Barrett (1979) note that "Appreciation may very well involve inference and evaluation, while evaluation may require inference. In fact, there are those who would argue that these three classifications may be thought of as three types of inference" (p. 67).

DLPT IV HANDOUT	BARRETT'S TAXONOMY OF QUESTION TYPES
Literal comprehension of explicitly stated information	**Literal recognition or recall of explicitly stated information. Recognition of:**
Detail: who, what, where, how, how much, how many, why, which, etc.	Detail: locate/identify/recall names of characters, the time a story took place, the setting of a story, or an incident described in a story, when such facts are explicitly stated in the selection.
Main idea: identify the main idea and/or differentiate between main and subordinate ideas.	Main ideas: locate/identify/ recall an explicit statement which is the main idea of a paragraph or a larger portion of the text.
Sequence: what comes first, last, immediately before, immediately after an event.	Sequence: locate/identify/recall the order of incidents or actions explicitly stated in the selection.
Perceiving relationships in the text: comparison/contrast	Comparisons: locate/identify/recall likenesses and differences among characters, times in history, or places that are explicitly compared by an author.
Perceiving relationships in the text: cause/effect	Cause and effect relationships: locate/identify/recall reasons for certain incidents, events, or characters' actions explicitly stated in the selection.
	Character traits: locate/identify/ recall statements about a character which help to point up the type of person s/he was when such statements were made by the author.
Inferential comprehension questions require some thinking and imagining on the part of the examinees. Inferencing tasks include:	**Inference,** a synthesis of the literal content of the text, personal knowledge, intuition, and imagination as a basis for conjectures or hypotheses. Infer:
Ask for further information about details.	Supporting details: conjecture about additional facts the author might have included in the selection which would have made it more informative, interesting, or appealing.
Arrive at a generalization from a series of details (e.g., what would be the best title?)	The main idea: provide the main idea, general significance, theme, or moral which is not explicitly stated in the text.
	Sequence: conjecture as to what might have taken place between two explicitly stated actions or incidents; hypo-the size what would happen next; hypothesize beginning of a story if the author had not started where he or she did.
Make comparisons	Comparisons: infer likenesses and differences in characters, times, or places.
Determine cause or effect (e.g., what is the cause of this happening?, how did this happen?, what is the result of that?)	Cause and effect relationship: hypothesize about the motives of characters and their interactions with others and with time and place; conjecture as to what caused the author to include certain ideas, words, characterizations, and actions in this writing.
What are people like?	Character traits: hypothesize about the nature of characters on the basis of explicit clues presented in the selection.
Predict outcomes (e.g., what is the result of that?)	Predictable outcomes: on the basis of the reading of an initial portion of the passage, conjecture about the outcome of the selection.
Draw logical conclusions (e.g., what is the result of that?, what would be a logical conclusion?)	
	Signification of figurative language: infer literal meanings from the author's figurative use of language.
Inferential comprehension questions: evaluate or make judgments	**Evaluation, making judgments**

Table 2: Question types in DLPT IV Handout and Barrett's Taxonomy of Reading Comprehension

In a guide for test development prepared by the Defense Language Institute (DLI) (1991), a classification system of the features of reading comprehension similar to the taxonomies discussed above (Swaffar, Langer, and Barrett) is presented. The DLI document, labeled DLPT IV, identifies two major question types: (1) literal comprehension of explicitly stated information, and (2) inferential comprehension questions. The DLI document includes evaluative questions within the second category. In Table 2, we display in detail the first two categories of Barrett's Taxonomy and the DLI classifications and how we equate them. The examples Barrett gives as evaluation questions are reader-based, and so we omit the specific details here. While such questions are certainly appropriate for teaching reading skills, we feel that for testing comprehension our questions must be, in Swaffar's words, "verifiable in the text" (p. 23). The two taxonomies are very similar, although they vary somewhat in the amount of detail they offer for classifying questions. It might also be noted that Barrett's examples appear to deal more with fictional genres while the DLPT IV examples are more generic. Because both taxonomies are useful for the development of test items, they appear together along the vertical axis of Table 3, which represents our task grid. In order to save space in this document, in Table 3 we omit repeated categories and the examples detailed in Table 2. The horizontal axis represents possible type of items to include in our CAT.

The horizontal axis of our task grid in Table 3 refers to the different question formats we might use. An interesting issue arises here with regard to the interaction of the difficulty of the reading passage and the difficulty of the test item. For example, an easy item might be written for a rather difficult text, and vice versa. Selecting texts and targeting items to ability levels *a priori* may be as much art as science. Indeed, as Kenyon (1995b) points out, it happens that items that were intended to be difficult turn out to be empirically easy, while items that were supposed to be easy show high difficulty values upon analysis. A framework like the ACTFL Proficiency Guidelines for targeting levels *a priori* can be a helpful tool. As mentioned earlier, such *a priori* speculated difficulty levels need to be verified empirically. Analyses of students' data will allow us to verify what scales or our intuitions may tell us about passage and item difficulty.

	SELECTED RESPONSE ITEMS			OTHER ITEM TYPES					
SYNTHESIS OF BARRETT'S TAXONOMY OF QUESTION TYPES & DLPT IV HANDOUT	Choose correct response to a question/ to complete a sentence.	Choose the picture best representing the passage	Multiple Matching	Free response	Deletion/ Insertion	Correction	Transformation	Reorganization	Cloze
Literal recognition or recall of explicitly stated information. Recognition of:									
Detail									
Main ideas									
Sequence									
Relationships in the text: comparisons									
Relationships in the text: cause and effect									
Character traits									
Inference, a synthesis of the literal content of the text, personal knowledge, intuition, and imagination as a basis for conjectures or hypotheses. Inferring:									
Supporting details									
A generalization									
The main idea									
Sequence									
Comparisons									
Cause and effect relationship									
Character traits									
Predictable outcomes									
Logical conclusions									
Signification of figurative language									
Evaluation, making judgments									

Table 3: Task grid

Dimensionality of the construct

Our discussion of language ability and reading models demonstrates that reading is a complex and interactive process. In order to describe the reading process in detail, we have discussed the various components of different models in such a way that it appears that the role of these components will be studied in isolation. Indeed, if our primary purpose were to develop diagnostic subtests to pinpoint areas of strength and weakness, items would be constructed to concentrate on specific dimensions of reading ability. While diagnostic assessments administered by computer will be pursued as a second step in our project, our first goal is to assess reading ability as a global construct.

In the present section, we discuss the dimensionality issue, which is a critical assumption for using an item response theory (IRT) model. IRT will be used as the adaptive algorithm in our CAT. The requirement of test unidimensionality is often referred to without clearly differentiating between it and the requirements of local independence and noninvasiveness that are central to IRT (Henning, 1989b). The dimensionality of a test is defined as the number of latent traits or dimensions that underlie test performance. Most IRT models, including the one-, two-, and three-parameter logistic models (1PLM, 2PLM, and 3PLM, respectively), require that a test be unidimensional. Local independence is the requirement that any two items be uncorrelated for any fixed level of ability (Henning, 1989b; Lord, 1980, 1968). Noninvasiveness is the requirement that the item parameters remain constant for any ordering of items in a test. In case of violating the invasiveness assumption, Henning maintains that if unidimensionality and local independence are upheld, and the items retain the same sequence on different test administrations, then IRT models can be used. (For an example of how to explore these issues, see Deville and Chalhoub-Deville, 1993.) The remainder of this section focuses primarily on unidimensionality. Local independence and noninvasiveness are discussed in the following section.

The concepts of dimensionality and local independence, while frequently confused, are distinct. Lord and Novick (1968), pointed out that the two concepts are related to each other in a specific manner. They define dimensionality as the total number of dimensions necessary to satisfy the assumption of local independence. In other words, local independence between the items will exist once all of the dimensions underlying test performance are identified and held constant. If local independence between the items can be obtained by conditioning on just one underlying dimension, the set of test items is unidimensional. When this is the case, just one underlying latent trait can explain the differences in item response patterns across people.

In a compelling article, Reckase, Ackerman, and Carlson (1988) illustrated a situation in which a test composed of more than one trait can meet the assumptions of a unidimensional IRT model. The authors demonstrated both theoretically and empirically, using real and simulated data, that a test composed of items that measure the same weighted composite of multiple abilities will fulfill the principle of local independence. For example, given a fixed level of the weighted composite of abilities, i.e. θ, item inter-correlations of zero (or near zero) indicate that the items are locally independent. If this is the case, the items, while not unidimensional, nonetheless function as if they were unidimensional and an unidimensional IRT model can be applied.

Relaxing the possibly too rigid assumptions of IRT, Stout (1987, 1990) proposed that the requirements of local independence and unidimensionality be replaced by the two concepts essential independence and essential unidimensionality. Instead of requiring a purely unidimensional test for the use of unidimensional IRT models, Stout believed that it was sufficient if there is one strong dominant dimension that runs through the test data even if it exists in the presence of a number of small, minor, dimensions (Nandakumar, 1991). DIMTEST, a statistical test developed by Stout (1987), and later refined (Nandakumar and Stout, 1993), is designed to determine if a particular test meets the requirement of essential unidimensionality before applying a unidimensional IRT model.

The assumption of unidimensionality states that "the test items and their responses operationally define a single dominant trait in latent space along a unitary continuum of performance" (Henning, 1988). There are divergent claims as to the unidimensionality of reading ability (e.g., Kaya-Carton and Carton, 1886; Kaya-Carton, Carton, and Dandoli 1991; and Laurier, 1993), but it appears that the robustness of measurement models such as IRT tolerate a somewhat flexible understanding of the unidimensionality assumption (Reckase, Ackerman, and Carlson, 1988; Roznowski, Tucker, and Humphreys, 1991). One situation in which the IRT models appear to tolerate seemingly multidimensional data is when the multiple dimensions are highly intercorrelated. Roznowski et al. found that highly intercorrelated dimensions can fit a unidimensional model. In a study of the dimensionality of dichotomous items, Roznowski et al. (1991) determined that "dependable IRT item and person parameters can be obtained from item pools that are not unidimensional in the strict sense of the term—a dominant dimension defined by correlated group factors is sufficient" (p. 109). Henning urges caution, however:

> the nature of the effects of violation of the unidimensionality assumption are not yet clearly understood, nor is there uniform agreement over the best methodology for determining the extent of the violation, nor has the robustness of the various IRT models to such a violation been fully quantified. (1988, p. 84)

It may be necessary to distinguish between theoretical and measurement approaches to reading ability to understand how a multicomponential construct like reading ability can meet the unidimensionality assumption of IRT. L2 reading theorists are concerned with understanding and explaining the reading comprehension process, and so try to identify the components involved in the process and build models to show how they fit together. Measurement researchers may be more concerned with how a test's scores reflect dimensionality; in other words, "what kinds of content fit (the unidimensionality assumption) is an empirical question" (Rentz & Rentz, 1979, p. 5). While the conceptual picture of reading ability portrays many dimensions, in practice, the interactions between these dimensions are so intertwined that it is the interaction itself that is the dominant factor in reading ability. In fact, establishing the dimensionality of a construct is not easily done *a priori*. It is only after the test for the construct

has been written and analyzed that information as to the trait's dimensionality becomes apparent. There are many possible ways to explore a test's dimensionality, including factor analysis, non-linear factor analysis, Stout's DIMTEST, Béjar's method, and others. (For a review, see Hattie, 1985). In our project, once data has been collected, we will use such methods to examine the test's dimensionality.

It is interesting to consider how the issue of dimensionality has been addressed in two recent studies of test development or analysis. Laurier (1993), who developed a computer-adaptive placement test for intensive French language programs in Quebec, distinguished between diagnostic functions of tests and global assessments. He writes that if one administers a battery of tests measuring different traits, the sum of these scores may not be a good indicator of level. One would have to prepare a profile for students, showing strengths and weaknesses. Such a profile would be of little use in forming homogeneous groups, which is required for placement purposes. Laurier suggests, therefore, that the parts of a placement test be all representative of proficiency and all interrelated. The placement test should be unidimensional, because a placement decision is unidimensional (level of course in which to place the student).

Laurier supports the theoretical arguments with empirical evidence. Laurier correlates the three subtests of his test. Correlations among the three subtests were very high; when corrected for attenuation correlations for true scores approach the theoretical limit of 1, or surpass it (p. 104):

	Subtest 1	Subtest 2	Subtest 3
Subtest 1	1.00	0.97	1.04
Subtest 2	0.97	1.00	1.06
Subtest 3	1.04	1.06	1.00

Due to their high intercorrelations, Laurier concludes that the three subtests measure a common trait. When doing split-half analyses to determine the internal coherence of each subtest, Laurier found "excellent internal coherence" (p. 109), and notes that, while by itself internal coherence

does not guarantee unidimensionality in IRT, tests with good internal coherence permit reliable parameterization (Cook, Dorans, and Eignor, 1988; Davidson, 1988; Henning, 1984; Henning, Hudson, and Turner, 1985). In conclusion, Laurier found that reading proficiency, while theoretically multidimensional, is composed of highly intercorrelated factors. The intercorrelations are often so strong that it may be possible to employ an unidimensional IRT model.

Kaya-Carton et al. (1991), who developed a French reading proficiency test for ACTFL, also rejected the notion of unidimensionality for reading proficiency "…and adopted the theoretical stance that reading proficiency is a complex concept involving multiple linguistic and psychological factors..." (p. 261). Based on this view, the authors selected reading texts and developed items. Next, responses to a paper-and-pencil version of the test were "used as raw data in applying multidimensional item response theory [MIRT] to calibrate the difficulty and differentiation indexes of the test items" (p. 264). Once the authors had MIRT data, which purportedly identified four dimensions, it was found that available CAT software could not handle multidimensional data. Therefore, "the test developers decided to include only one of the four dimensions, the item discrimination indexes in the computation of *theta* values (ability scores)" (pp. 272-273) and applied a one-dimensional IRT model. To avoid the disappointment that Kaya-Carton et al. encountered, we decided to survey the market for CAT software with MIRT algorithm capabilities. Our survey shows that such capabilities are still not available.

In summary, in our context, once the test items are written, they will be pilot tested on large samples of French, German, and Spanish students. From this data, we will be able to take an in-depth look at the underlying dimensionality of the test items using procedures such as factor analysis and Stout's DIMTEST (Stout, 1987, 1990). If it is found that the new test items are not unidimensional, the researchers will explore the variables/issues contributing to such multidimensionality.

Item dependencies, noninvasiveness, and testlets

The reading proficiency CAT test being developed by CARLA will consist of a series of reading passages, each followed by items relating to a particular passage, referred to as testlets. According to Wainer and Kiely (1987), a testlet is a group of items related to a single content area or passage that is developed as a unit.

When a test is made up of a series of testlets, the assumption of local independence, conditional independence of the item responses on ability scores, may be violated. Items relating to one passage may be more highly correlated with each other than they will be to items associated with different passages. To illustrate how this might occur, when a student has difficulty understanding a particular reading passage, i.e., passage A, his or her performance on all items relating to passage A will be poor. In contrast, if the same student is quite familiar with the topic of a second passage, i.e., passage B, and, consequently, finds passage B very understandable, he or she will tend to do better on all the items associated with passage B relative to the questions related to the earlier reading, passage A. Hence, when holding reading ability constant, the items within passage A may still be intercorrelated even when the items between passages A and B are uncorrelated. This is because the items in passage A have something in common besides the primary latent trait--the passage. Local independence, however, requires that item responses be independent of one another at a fixed ability level and the violation of this principle could pose a serious problem in implementing an IRT model.

One option for dealing with the violation of local independence is to limit each passage to just one item. While eliminating the problem of item dependencies, this solution has several negative consequences. For example, using just one item per passage would increase the amount of reading necessary by students. This is extremely inefficient (Thissen, Steinberg, and Mooney, 1989) and would counteract the reduced testing time that is one of the advantages of CAT administration. Shorter passages could be used, but this would probably limit the breadth and depth of the reading texts that could be included in the exam. Most importantly, plodding

through a large number of shorter passages is likely to be slow and tiresome for students leaving them dissatisfied with the exam.

Another option is to use testlets consisting of more than one item associated with a passage and arrange the items in a pseudo-adaptive branching hierarchy with a fixed number of possible patterns (Wainer and Kiely, 1987) or as a mini-CAT in which items are selected within each testlet using a true adaptive algorithm (Thomas, 1990). In order to eliminate the problem of item dependencies associated with the testlet format, each set of items associated with a separate passage would be scored as a separate mini-test (or testlet). "Items that naturally fall in testlets are frequently not locally independent when the test is taken as a collection of individual items; but the testlets, taken as units, may be locally independent, permitting 'unidimensional scoring' of each testlet separately" (Thissen et al., 1989, p. 248). In other words, each examinee would have a score on each of the testlets. An overall test score would be obtained by calculating some weighted average of the testlet scores (i.e., a weighted average of the 1,...,n ability estimates associated with testlets 1,...,n). In this way, each testlet is unidimensional. While providing the student with a score on each testlet may help with resolving the issue of item dependencies, the differential ordering of items that occurs with the pseudo-adaptive and mini CAT approaches suggested above is problematic.

There is a problem associated with either the branching or mini-CAT solution, where for the same passage, items are presented to different students in different orders. Although this is the nature of an adaptive approach, it poses a particular challenge in tests of reading comprehension in which the ordering of items pertaining to the same passage is believed to be very important (Deville and Chalhoub-Deville, 1993; Gordon & Hanauer, 1995; Thissen et al., 1989) and different orderings are not necessarily interchangeable. Thissen et al. remarked that "it is not obvious that individual items are interchangeable in an item pool in which clusters of items follow reading passages" (p. 248). In other words, Henning's (1989b) noninvasiveness requirement may not hold in a reading comprehension test. For example, as Gordon and

Hanauer show, a preceding question can impact readers' mental representation of the text and can affect a student's answer to a subsequent question on the same passage. If some students received that earlier question, while others did not, the difficulty of the latter item can differ between these two groups of students. Similarly, in investigating the recall protocol of a reading text, Deville and Chalhoub-Deville (1993) show that some of the same pausal units were recalled differently based on their place in the sentence and the text, which is a violation of the noninvasiveness assumption. (Nevertheless, their analyses showed that the local independence and unidimensionality assumptions were upheld.)

In order to control item ordering effects (invasiveness) in a reading comprehension exam, some researchers have suggested that testlets be used with a fixed sequence of items following each passage (Thissen et al., 1989; Wainer and Kiely, 1987). Wainer and Kiely termed such item groupings linear testlets. In addition, the linear testlets making up an exam may contain varying numbers of items, but any given linear testlet always has a specific number of items administered with it. In a test made up of linear testlets, all students given a particular passage would receive for that passage the same set of items in the same order. As such, the order of the items within a passage is held constant across people, thereby eliminating any differential ordering effect.

For our project, a variation of the linear testlet methodology described above will be considered. Associated with each reading passage would be a linear testlet with items focusing on a specific level of difficulty. As the examinee completes a testlet, the adaptive algorithm would then select the next reading passage with the group of items of the most appropriate difficulty for the examinee's ability level.

Finally, once the items are developed and pilot tested, the researchers will examine the items in order to determine if a violation of local independence due to the testlet/passage format exists. An indication of this would be if the average item intercorrelation is higher between items from the same passages than it is for items from different passages.

Testlet scoring

As mentioned earlier, providing a testlet score is an option that might resolve the problem of item dependencies that would arise if each item in every passage were scored separately. With this testlet scoring format, we still need to decide whether to use a polychotomous IRT model, as recommended by Haladyna (1992). (Unlike dichotomous models in which an item is scored either 1 (correct) or 0 (incorrect), polychotomous models were designed to be used with items in which an item score can take on three or more possible values.) If so, should we use for example, the graded-response model or Thissen et al.'s (1989) nominal IRT model?

Thissen et al. (1989) suggested a method for scoring linear testlets using an adaptation of the nominal response IRT model (Bock, 1972) that treats each testlet as a separate item (a "super item"). The nominal model is designed for use with items that can take on three or more categorical scores that have no intrinsic ordering from low to high or better to worse. For example, with a three-item testlet, the response categories would be 0, 1, 2, and 3, depending on the number of items answered correctly. If the items are calibrated using a polychotomous model such as the nominal model, calibrations would be done per testlet rather than per item as is done with the more common IRT models.

Another option is to score each testlet dichotomously. For example, if the examinee gets 3 or more of the 5 testlet items correct, he/she gets a score of 1, otherwise a score of 0 is received. In this case, a standard IRT model could be employed. With such a scoring method, the critical issue pertains to the amount of precision lost if the testlets were scored dichotomously. This issue will need to be researched empirically. Nevertheless, using a dichotomous, instead of a polychotomous model, is preferred mainly because it is the model utilized by readily available CAT software.

CAT entry point

What method should be used in obtaining an individual's initial θ estimate for selecting the appropriate first testlet? One possibility is to have students do a self-assessment of their ability and use that as a starting point. Another possibility would be to present each student with one or more initial passages, each followed by items of varying difficulty. Each student must answer at least one of the items correctly and one of the items incorrectly. Otherwise, it is impossible to calculate an estimate of the person's reading proficiency (θ). For this reason, the items must be sufficiently diverse in difficulty so that all examinees have at least one correct and one incorrect item score on the first part of the test. Based on their performance on these testlet items, the appropriate first testlet of the test is selected. We will also consider combining information from self-assessment and performance on testlets to determine the appropriate entry level.

CAT stopping rules, content sampling, and exposure controls

There are a number of other issues that must be considered. We need to explore whether to have the test be of a fixed length with a fixed number of testlets and items being administered, or whether there be an adaptive termination criterion (stopping rule). Also to be determined is how stringent the termination criterion should be. If we wish to employ some control on the content areas sampled, we will probably need to put a restriction on the test termination so that all areas be sampled before testing stops. We also need to guard against students, teachers, and parents perceiving the test to be unfair if different numbers of items are given to different students. We must also consider setting exposure controls such as the Sympson/Hetter approach (Stocking, 1992). to protect against over-administration of certain items and to enhance test security.

CAT item pool

Related to the above questions, the researchers must estimate, *a priori*, how many items will be needed in the CAT item pool. Stocking (1994) stated that factors in the design of a CAT that influence the required item pool size include "the item selection algorithm, constraints on item content, psychometrics, exposure, stopping rules, overlap restrictions, test scoring, requirements of parallelism with existing paper-and-pencil forms, and so forth" (p. 7). Weiss (1985) suggested that a "CAT operates most effectively from an item pool with a large number of items that are highly discriminating and are equally represented across the difficulty-trait level continuum" (p. 786). In general, the more attributes and properties we add to the CAT, the larger the item pool required. Developing large item pools are quite costly and may prove to be prohibitive. Also, the simpler the CAT design, the easier it is to predict the necessary item pool size using tables provided by Stocking (1994).

Dichotomous IRT models

In the development of the reading proficiency CAT instrument, there are a variety of dichotomous IRT models from which to choose for the calibration of items and for the scoring of examinee responses. The most popular models are the 1PLM, 2PLM, and 3PLM. These models can be used for calibrating each item separately where the empirical evidence suggests that local independence is not violated by the testlet format, for calibrating each item separately when the one item per passage format is used, or when each testlet is treated as a dichotomous "super item", as described in the previous section.

In the 1PLM (also called the Rasch model), the probability that an examinee with trait level θ answers an item i correctly (i =1,...,n) is a function of the item's difficulty (b). In a 2PLM, each item is described by the 2 parameters, item difficulty (b) and item discrimination (a). In the 3PLM, each item is described by item difficulty (b), discrimination (a), and the probability that the correct answer was chosen merely by guessing (c) (see Hambleton, 1985, pp. 37-49).

There is considerable debate regarding which of the three models is most appropriate for selected response items, e.g., multiple choice. One serious issue with the 3PLM is that the *c* parameter is difficult to estimate (Henning, 1987, pp. 116-117). One solution might be to fix *c* to the quantity 1 divided by the number of options (i.e., for a 5 option test item, *c* is fixed at .20). This eliminates the difficulty in estimating *c*. How accurate, however, is the fixed *c* parameter?

An alternative is to use the 2PLM. While guessing is not modeled in the 2PLM, the item discriminations are allowed to vary. This allows more information to use in selecting the next item in a CAT administration. Nevertheless, some researchers (e.g., de Jong, 1996) argue that large variations in discrimination indexes should be examined for possible bias.

The 1PLM has many conceptual and practical advantages over the 2PLM and 3PLM. Its advantages include increased accuracy of estimation, economy of computing time, the reduced sample size required (Lord, 1983), and ease of interpretation (Henning, 1987, 1989a). Furthermore, it may be reasonable to assume that if an item is well-written, no student will make completely blind guesses that take absolutely none of the information in the passage or item content into account. With well-written items, students will attempt to reason through the alternatives and use a hunch based on the item and its alternatives to decide on an answer, even if they fail to understand the content of the reading passage. In this case, one might argue that including *c* in a model may be simply incorporating error variance into the model due to poorly written items. With regard to discrimination, researchers (e.g., Divgi, 1986; Lord, 1983) argue that if item discriminations do vary, a Rasch model may not provide a good fit to test data. A variation on the 1PLM fixes the item discriminations, but not necessarily at 1.0. The discriminations may be fixed at some other value of *a* or to several distinct values of *a* based on the empirical knowledge of the items (Verhelst & Glas, 1995). This option of the 1PLM with the empirically determined discrimination index is the favored approach for the calibration of items and the scoring of examinee responses.

Test method characteristics

To help clarify the specific aspects of our test, we have used Bachman's (1990) list of test method characteristics as a guide.

Characteristics of the environment

Place, materials and equipment: The test will be administered on PCs in a variety of possible settings, such as classrooms, computer labs, testing stations, etc. Since most language classrooms are not equipped with a computer for each student, taking the test will usually require students to go to a different location. Although students will not be in a familiar location, every effort will be made to provide a comfortable environment.

Personnel: The test will be designed to be administered without the intervention of personnel, but how one administers the test will depend upon the population tested, the situation, and so on. At a minimum, there should be personnel present to assist with computer-related problems and to assure that scores reflect individual effort.

Time of testing: In theory, test-takers can choose times at their convenience to take the test. In practice, some limits will probably be placed upon the test's accessibility, depending upon such factors as hours of operation of computer facilities, the amount of control program administrators wish to exercise over test availability, the number of times students are allowed to take the test, the availability of personnel to supervise test-takers, and so on.

Physical conditions: Test-takers should have a comfortable, quiet, and private area in which to take the test. The computer equipment should function correctly and efficiently. Numerous physical setups are possible, but care should be taken that the physical conditions do not impact performance on the test or compromise the requirement that scores reflect individual effort.

Characteristics of the test rubric

Test organization: The test will be organized in two parts. The first part is a reading proficiency test and the second part is a diagnostic test. As indicated earlier, the proficiency test is the focus of this paper, and the diagnostic test will be touched upon only briefly in this section concerning test method characteristics. The two tests are different in conceptualization, design, implementation, interpretation, and use.

At present, we envision the use of our test in this way: If the test is being used for certification or clearance for entrance to, or exit from, a course of study, students who perform to criterion level will take only the proficiency test. If students do not meet the criterion level, they will advance to the diagnostic section of the test. The diagnostic test will attempt to identify problem areas that contributed to the test-taker's not passing the proficiency portion of the assessment. It may be feasible in the future to use the test for placement. Of course, that would entail considerable additional work.

Salience of parts: The two principal parts of the assessment, proficiency and diagnosis, will be clearly indicated and explained to the test-takers. With regard to the subparts, we do not foresee any differential weighting of texts or items.

Sequence of parts: The reading proficiency test is administered first, followed by the diagnostic test. The proficiency test is adaptive, which means that the sequence of items will vary by individual and will be determined by correctness of response. The adaptive feature usually works in this way: If an item is answered correctly, the next item presented will be more difficult, while an incorrect response will cause a less difficult item to be selected. To select a beginning point in the test, students are required to provide information about their language ability that would determine the level at which they are to be tested. Next, all students taking a determined level of the test are given a fixed set of questions. Based on their performance on these items, the computer algorithm branches out and the adaptive aspect of the test begins. In the case of our adaptive tests, we are using fixed linear testlets. Here a testlet is defined as a

reading passage accompanied by *several* questions pertaining to that passage. (A more complete description of testlets appears in the section of this paper entitled **Item dependencies, noninvasiveness, and testlets,** p. 34). A score for the testlet will be calculated and will be the basis for selecting an easier or more difficult next testlet. Students will receive a range of testlets to ensure an appropriate coverage of the foreign language content domain for the given reading ability level.

Relative importance of parts: Different users of the test may view the relative importance of the parts in different manners, depending upon the decisions to be made.

Time allocation: The advantage of computer-administered tests is that, in principle, students may take as much time as they need to respond to the items. For practical reasons, some upper time limit may need to be set. Also, with CAT, it is claimed that testing time is considerably shorter. It remains to be seen how short the test will be, considering that the test has to depict an appropriate coverage of the content domain.

Scoring method: The items will be dichotomously scored, correct/incorrect, using IRT. IRT applications calculate scores as a function of item difficulty, thus a correct response to a more difficult item will contribute more to the overall proficiency score than a correct response to an easy item. It is possible, then, for two individuals to correctly answer the same number of questions, and for one to have a passing score while the other does not.

Criteria for correctness: As we explore the capabilities of CAT software, we will consider different item formats and criteria for correctness. In selected response formats, only one choice will be the correct response. If we are able to use other item formats, e.g., sentence completion, credit may be given for partially correct responses. One issue we must consider in selecting item formats is their familiarity to students—understanding the task to be performed should not contribute to item difficulty. Even when we assume a high degree of familiarity with the item format, the criteria for correctness will be carefully explained.

Procedures: Scoring will be performed by the computer as the test progresses, i.e., on-line scoring. Scoring information will be used to determine the next testlet presented.

Explicitness of criteria and procedures: The criteria and procedures for scoring will be explained clearly to test-takers. However, item difficulty level information will be withheld as it may contribute to the test-takers' anxiety.

Score reporting: A final consideration is how the scores should be reported to the students. Given the particular needs of the various institutions, the examinees can either get immediate feedback on their performance via the computer or be sent that information. It is also possible to report the results, either as scaled scores or a simple pass/fail. For those who do not pass, and are branched into the diagnostic part of the test, they are provided with detailed profiles of their performance showing their strengths and weaknesses in specific areas.

Instructions

Language (native, target): The instructions will be presented in English.

Channel (aural, visual): Instructions will be presented visually on the computer screen.

Specification of procedures and tasks: The instructions will explain how to proceed through the test and how to respond on the computer. Most test-takers will be familiar with the item formats; nonetheless, explicit instructions will be provided along with several sample items at the beginning of the test.

Characteristics of the input

Format

Channel: The input will be presented visually on the computer screen.

Mode: Test-takers will read the input (receptive mode).

Form: For the most part, the input will consist of language samples, but reading passages may include graphic displays that require interpretation, and tasks may incorporate pictures. Thus, both language and nonlanguage forms may be used in the input.

Language: Reading passages will be presented in the target language, and tasks will be in English, the L1 for most of the test-takers.

Length: Passage lengths will vary due to many factors. Text type, topic, authenticity, and the hierarchical level the testlet is targeting are some of the determinants in passage length. We will be somewhat constrained by the computer format of the test; for example, we would prefer that the entire passage be presented on one screen so that test-takers are not required to scroll through the text. Having to scroll might inhibit looking back in the text—a common reading behavior. We must also consider the effects of having the test items on a separate screen from the reading passage. A split screen may be necessary so that it is easy to see the passage at the same time one is attempting to respond to items. Different types of presentation could be pilot tested to check for effect on scores and student preferences.

Type: Testlets combine two types of input: a reading passage (input for interpretation and comprehension) followed by several short stand-alone items. We intend to use selected response item formats, but we will investigate the possibility of using other item formats as well.

Degree of speededness: It is intended that test-takers use as much time as they need to process the information in the input (passage and question items). However, some upper limit of time may be needed in order to stop the test. Cases may arise where test-takers spend inordinate amounts of time trying to understand difficult texts.

Identification of language problem: In selected response items, the 'problem' or task to perform will be identified in the ways frequently reported in the literature found, e.g., fill in the blank, rearrange sentences/paragraphs. Depending upon the adaptability of the CAT software, other item formats may be used. Whatever the question format, care will always be taken to make sure that the task to be performed is clear.

Language of input

Organizational characteristics

Grammatical (syntax/morphology, phonology/graphology) and textual (cohesion, rhetorical organization): We will use authentic texts as much as possible. When targeting lower ability levels, it may be necessary to construct some texts. In such cases, the language of the input will be organized according to L2 norms. Grammatical and textual organization characteristics will vary by level targeted. Both the ACTFL Proficiency Guidelines, specifying degree of grammatical and textual complexity, and L2 teaching experts will be used up front to help us select texts appropriate for the different ability levels. Also piloting will help refine *a priori* hypothesized scheme.

Pragmatic characteristics

Propositional content

Topic: the selection of appropriate topics for different levels will be informed by teachers, experts, the ACTFL Guidelines, and other related research. Nevertheless, while it might be reasonable to assume that topics appropriate for low levels of proficiency will be familiar, non-specialized, non-technical, and not too culture specific, such an *a priori* determined hierarchy will need to be verified empirically.

Genre: As discussed above with regard to topics, familiarity will be a primary consideration in our selection of text types. Also, to some degree, different genres imply varying levels of difficulty by their very nature, and we must be concerned with selecting genres appropriate to ability levels. It is difficult to develop hard rules about genre, however, since texts must be considered in light of other factors such as organizational complexity, pragmatic function, vocabulary, topic, the task the test-taker is asked to perform with the text, and so on, and not deemed appropriate or inappropriate solely because they do or do not meet a criterion of text type.

Type of information: Whether the information contained in a text is concrete or abstract, positive or negative, factual or counterfactual will all play a role in the selection of texts for the different target levels in the test. It would seem best to avoid abstract, negative, and counterfactual information for all levels but those at the highest proficiency levels.

Vocabulary: Bachman and Palmer (forthcoming) list frequency, domain of usage (specialized or not), cultural references, and figures of speech as factors to consider in this category. The ACTFL Proficiency Guidelines describe appropriate vocabulary for different levels, and we will consider their recommendations.

Functions: Bachman identifies the functions of text as ideational, manipulative, heuristic, and imaginative. The ACTFL Proficiency Guidelines include language identifying the purposes of texts, such as instructional, directional, social, informational, and for pleasure. They also identify texts as serving basic, personal, social, academic, and professional needs. We will select texts written to perform a variety of functions and purposes, as appropriate to targeted ability levels.

Sociolinguistic characteristics: The dialect or variety of language used in the input text should be 'standard', although it may be difficult to arrive at a precise definition of standard. It may be easiest to meet this aim by eliminating texts that deviate from what is generally accepted as standard. French, German, and Spanish all have varietal and dialectical differences. The Spanish language is particularly challenging in this regard given the large number of varieties found throughout the Spanish-speaking world. Texts containing words or expressions that can be identified, for example, specifically as Argentinean Spanish or Cuban Spanish or some other variety are likely to be avoided. Here again, however, we must keep in mind the impact that 'nonstandard' varieties or dialects have on the actual tasks test-takers are asked to perform. It may be possible to construct items in such a way that nonstandard language in the input text does not interfere with task completion. Register is another sociolinguistic factor we must consider. Sensitivity to register is expected to become more acute as we progress on the proficiency scale.

Characteristics of the expected response

Format

Channel: Responses will be selected from a visual display on a computer screen.

Mode: Test-takers will select the appropriate response and register their selection via the computer.

Form: The form of the expected response will be a letter of the alphabet, a number, or a click of a mouse. Any of these forms will represent the selection of an appropriate response from among a set of alternatives.

Language: While the reading passages are in L2, the selected response test items will be in English. Thus, test-takers will not be required to produce or in any way manipulate the target language.

Length: Length of response is not a consideration, inasmuch as test-takers will not produce language. Length will be an issue for item writers, however, since consistency in length of distractors, for example, is important in writing good items.

Type: The response type is selected response.

Degree of speededness: No time limits will be set for selecting responses. An overall time limit may be required for the test as a whole, but it is envisioned that it would be set well beyond the time estimated for the majority of test-takers to complete the test.

Language of expected response

Organizational characteristics: Because test-takers will not produce language in their responses, organizational characteristics (grammatical, textual) and pragmatic characteristics (propositional content, language functions, sociolinguistic characteristics) are more a concern for item writers in the initial test development than they are for test-takers. Care must be taken to write appropriate items for targeted levels, and to avoid "trick" questions. An example of an inappropriately written item might include one that requires knowledge of a particular

vocabulary word that is unlikely to be part of the repertoire of learners at the item's targeted level, or item distractors written in negative rather than positive language.

Relationship between input and response

Reactivity: There is an adaptive relationship between input and response. Items will be presented to the test-taker as a function of performance on preceding items.

Scope of relationship: According to Bachman, this category "pertains to the amount or range of input that must be processed in order for the test-taker to respond as expected." The scope of our items will range from interpreting detail to interpreting a whole text, such as in an item asking for the main idea.

Directness of relationship: Bachman writes that the directness of relationship concerns "the degree to which the response deals primarily with information in the input, or whether the test-taker must rely on information in the context or in his own knowledge schemata." Items will be written to reflect a direct relationship, i.e., information required to respond correctly is contained in the reading passage. However, test-takers will use their knowledge schemata as part of the reading process, and so some measure of indirectness is inherent in a reading assessment.

Conclusion

The purpose of the present document is to review the language and measurement literature to inform the development of the assessment framework for our computer adaptive reading proficiency test. The posited CAT framework, as well as the input of teachers involved in teaching at the targeted levels will be employed when selecting the reading texts and constructing the related test items. These CATs will be used to certify French, German, and Spanish students at three levels: exit from secondary/entrance into post-secondary instruction, exit from post-secondary instruction, and completion of the foreign language requirement.

We recognize that the *a priori* defined components of the assessment framework and the related features are likely to be modified as per the empirical evidence garnered during pilot

testing and future research work. Given the performance of the students, which represents the interaction between the test and the test-taker characteristics, we are likely to expand the framework, revise aspects of its components, etc. In other words, the up-front work will be continually refined based on pilot testing and empirical research. Pilot testing and future research on our CATs will be appropriately disseminated to continue to contribute to the research base and to inform others involved in CAT development.

References

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

Allen, E. D., Bernhardt, E. B., Berry, M. T., & Demel, M. (1988). Comprehension and text genre: An analysis of secondary school foreign language readers. *Modern Language Journal*, 72(2), 163-172.

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

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

Bachman, L. F., & Palmer, A. S. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking reading. In A. S. Palmer, P. J. M Groot, & G. A. Trosper (Eds.), *The construct validation off tests of communicative competence* (pp. 149-165). Washington, DC: TESOL.

Barnett, M. A. (1989). *More than meets the eye: foreign language reading: theory and practice*. Englewood Cliffs, NJ: Prentice-Hall.

Barrett, T. C. (1972). Taxonomy of reading comprehension. In *Reading 360 Monograph*. Lexington, MA: Ginn, A Xerox Educational Company.

Bernhardt, E. B. (1991). *Reading development in a second language: theoretical, empirical, and classroom perspectives*. Norwood, NJ: Ablex Publishing Corporation.

Bernhardt, E. B. (1986). Proficient texts or proficient readers? *ADFL Bulletin, 18* (1), 25-28.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Kratwohl, D. R. (Eds.). (1956). *Taxonomy of educational objectives: cognitive domain*. New York: David McKay.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Brindley, G. (1991). Defining language ability: the criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing* (pp. 139-164). Singapore: SEAMEO Regional Language Centre.

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

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47.

Chalhoub-Deville, M. (in press). Theoretical models, assessment frameworks, and test construction. *Language Testing*.

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

Chalhoub-Deville, M. (1995b). A contextualized approach to describing oral proficiency. *Language Learning*, 45, 251-281.

Chalhoub-Deville, M., Alcaya, C., Klein, F., Lozier, V. M., Budlong, E. (1996). *Qualitative and Quantitative Review of the University of Minnesota CLA French Entrance and Graduation Proficiency Tests*. (Technical Report No. 3), The Center for Advanced Research on Language Acquisition, University of Minnesota, Minneapolis.

Chalhoub-Deville, Mueller, I., F., Lozier, V. M., Juengling, F. (1996). *Qualitative and Quantitative Review of the University of Minnesota CLA German Entrance and Graduation Proficiency Tests.* (Technical Report No. 1), The Center for Advanced Research on Language Acquisition, University of Minnesota, Minneapolis.

Chalhoub-Deville, M., Sweet, G., Schmidt, K., Lozier, V. M. (1996). *Qualitative and Quantitative Review of the University of Minnesota CLA Spanish Entrance and Graduation Proficiency Tests.* (Technical Report No. 4), The Center for Advanced Research on Language Acquisition, University of Minnesota, Minneapolis.

Chalhoub-Deville, M. & Lozier, V. M. (1995). Preliminary Item Response Theory Analysis of the University of Minnesota CLA Language Proficiency Tests in French, German, and Spanish. Center for Advanced Research on Language Acquisition, University of Minnesota, Minneapolis.

Chomsky, N. (1973). Linguistic theory. In O. J. & J. Richards (Eds.), *Focus on the Learner: Pragmatic perspectives for the language teacher*. Rowley, MA: Newbury House.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clifford, R.T. (1981). Convergent and discrimenant validation of integrated and unitary language skills: The need for a research model. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 149-165). Washington, DC: TESOL.

Cohen, A. D. (1984). *Assessing language ability in the classroom*. Boston: Heinle & Heinle.

College Entrance Examination Board (1996). *Articulation and achievement: connecting standards, performance, and assessment in foreign language*. New York: College Entrance Examination Board.

Community Literacy Collaborative (1995). GEL: Guide to ESL Levels for the Community Literacy Collaborative, St. Paul, MN.

Cook, L.L., Dorans, N.J., & Eignor, D.R. (1988). As assessment of the dimensionality of three SAT-verbal test editions. *Journal of Educational Statistics*, *13*, 19-43.

Corl, K.A., Harlow, L.L., Macián, J.L., Saunders, D.M., (1996). Collaborative partnerships for articulation: asking the right questions. *Foreign Language Annals*, *29*(2), 111-124.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals*, *23*(1), 11-22.

Davey, B. (1989). Assessing comprehension: Selected interactions of task and reader. *Reading Teacher*, *42*(9), 694-697.

Davidson, F. G. (1988). *An exploratory modeling survey of the trait structures of some existing language test datasets*. Unpublished doctoral dissertation. University of California, Los Angeles.

Defense Language Institute (1991). *DLPT IV guidelines and checklists for test development.* Monterey, CA.

de Jong, J. & Stoyanova, F. (1996). Discrimination Parameters and Test Dimensionality. Paper presented at the Language Testing Research Colloquium, Tampere, Finland.

Deville, C. & Chalhoub-Deville, M. (1993). Modified scoring, traditional item analysis and Sato's caution index used to investigate the reading recall protocol. *Language Testing, 10,* (2), 117-132.

Divgi, D.R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, *23*(4), 283-298.

Downing, J., & Leong, C. K. (1982). *Psychology of reading*. New York: Macmillan.

Elder, C. (1996). Performance testing for the professions: language proficiency or strategic competence? Paper presented at the Annual Conference of the American Association of Applied Linguistics.

Eskey, D. E. (1988). Holding in the bottom: an interactive approach to the language problems of second language readers. In P. L. Carrell, J. Devine, & D. E. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 93-100). Cambridge: Cambridge University Press.

Goodman, K. S. (1967). Reading: a psycholinguistic guessing game. *Journal of the Reading Specialist*, *6*(1), 126-135.

Gordon, C. & Hanauer, D. (1985). The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly, 29*(2), 229-324.

Grabe, W. (1988). Reassessing the term "interactive". In P. L. Carrell, J. Devine, & D. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 56-70). Cambridge: Cambridge University Press.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice* (Spring), 21-25.

Hambleton, R. K. &. Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164.

Henning, G. (1989a). Does the Rasch model really work for multiple-choice items? Take another look: a response to Divgi. *Journal of Educational Measurement*, *26*(1), 91-97.

Henning, G. (1989b). Meanings and implications of the principle of local independence. *Language Testing*, *6*(1), 95-108.

Henning, G. (1988). The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. *Language Testing*, *5*(1), 83-99.

Henning, G. (1987). *A guide to language testing: development, evaluation, research*. Cambridge, MA: Newbury House Publishers.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing*, *1*(2), 123-133.

Henning, G. (1983). Oral proficiency testing: comparative validities of interview, imitation, and completion methods. *Language Learning*, *33*(3), 315-332.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, *2*, 141-154.

Higgs, T. V., & Clifford, R. (1982). The push toward communication. In T. V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher*. Lincolnwood, IL: National Textbook.

Humphreys, L. G. (1985). General intelligence: an integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 201-224). New York: Wiley.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist*, *17*(475-483).

Hymes, D. (1971). Competence and performance in linguistic theory. In R. Huxley & E. Ingram (Eds.), *Language acquisition: models and methods* London: Academic Press.

Ingram, D. E. (1979). Introduction to the Australian second language proficiency ratings (ASLPR). A paper in the AMEP Teacher's Manual. Canberra: Australian Government Publishing Service.

Ingram, D. E., & Wylie, E. (1984). Australian second language proficiency ratings (ASLPR). Canberra: Australian Government Publishing Service.

Ingram, D. E., & Wylie, E. (1982). Australian second language proficiency ratings (ASLPR) revised. Canberra: Australian Government Publishing Service.

Kaya-Carton, E., Carton, A. S., & Dandonoli, P. (1991). Developing a computer-adaptive test of French reading proficiency. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: research issues and practice* (pp. 259-284). New York: Newbury House.

Kaya-Carton, E., & Carton, A. S. (1986). Multidimensionality of foreign language reading proficiency: Preliminary considerations in assessment. *Foreign Language Annals*, *18*(2), 95-102.

Kenyon, D. M. (1995a). An investigation of the validity of the demands of tasks on performance-based tests of oral proficiency. In paper presented at the 16th annual Language Testing Research Colloquium. Long Beach, CA.

Kenyon, D. M. (1995b). *Linking multiple-choice test scores to verbally-defined proficiency levels: an application to Chinese reading proficiency*. Doctoral dissertation. University of Maryland.

Langer, J. A. (1985). Levels of questioning: an alternative view. *Reading Research Quarterly*, *20*(5), 586-602.

Laurier, M. (1993). *L'informatisation d'un test de classement en langue seconde (Computerized second language placement test)*. Quebec: International Center for Research on Language Planning.

Lee, J. F. (1988). Toward a modification of the "proficiency" construct for reading in a foreign language. *Hispania*, *71*(4).

Lee, J. F. (1987). Comprehending the Spanish subjunctive: an information processing approach. *Modern Language Journal*, *71*(1), 50-57.

Lee, J. F., & Musumeci, D. (1988). On hierarchies of reading skills and text types. *Modern Language Journal*, *72*(2), 173-187.

Lesgold, A., & Perfetti, C. (1981). Interactive processes in reading: where do we stand. In A. Lesgold & C. Perfetti (Eds.), *Interactive processes in reading* (pp. 387-407). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1983). Small N justifies the Rasch model. In D. J. Weiss (Ed.), *New horizons in testing.* New York: Academic Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. &. Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Lozier, V. M. & Chalhoub-Deville, M. (1995). *Preliminary item response theory analysis of the University of Minnesota CLAlanguage proficiency tests in French, German, and Spanish* (Technical report No.2). The Center for Advanced Research on Language Acquisition, University of Minnesota, Minneapolis.

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of the effect of context in perception. *Psychological Review*, *88*, 375-407.

Metcalf, M. (1995). Articulating the teaching of foreign languages: the Minnesota Project. *ADFL Bulletin*, *28* (3), 52-54.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement*, *28*(2), 99-117.

Nandakumar, R. &. Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of educational statistics*, *18*(1), 41-68.

North, B. (1993). The development of descriptors on scales of language proficiency. Occasional papers. National Foreign Language Center, Johns Hopkins University, Washington, DC.

Oller, J. W., Jr. (1983). Evidence for a general language proficiency factor: an expectancy grammar. In J. W. Oller Jr. (Ed.), *Issues in language testing research* (pp. 3-10). Rowley, MA: Newbury House.

Omaggio, A. C. (1986). *Teaching language in context: proficiency-oriented instruction*. Boston, MA: Heinle and Heinle.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*(3), 193-203.

Rentz, R. R., & Rentz, C. C. (1979). Does the Rasch model really work? A discussion for practitioners. *Measurement in Education*, *10*, 1-8.

Rost, D. H. (1993). Assessing different components of reading comprehension: fact or fiction? *Language Testing*, *10*(1), 79-92.

Roznowski, M., Tucker, L. R., & Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary items. *Applied Psychological Measurement*, *15*(2), 109-127.

Rumelhart, D. E. (1977). Toward and interactive model of reading. In S. Dornic (Ed.), *Attention and performance* (pp. 573-603). New York, NY: Academic Press.

Savignon, S. J. (1983). *Communicative competence: theory and classroom practice*. Reading, MA: Addison-Wesley.

Schallert, D. L., Ulerick, S. L., & Tierney, R. J. (1984). Evolving a description of text through mapping. In C. D. Holley & D. F. Dansereau (Eds.), *Spatial learning strategies: techniques, applications, and related issues* (pp. 255-274). New York, NY: Academic Press.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*, 147-170.

Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedure. *Language Learning, 33*. 527-539.

Shohamy, E., Reeves, T., Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal, 40,* 212-240.

Smith, F. (1971). *Understanding reading: a psycholinguistic analysis of reading and learning to read*. New York, NY: Holt, Rinehart and Winston.

Smith, R. J., & Barrett, T. C. (1979). *Teaching reading in the middle grades* (2nd ed.). Reading, MA: Addison-Wesley.

Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.

Steffensen, M. S., Joag-Dev, C., & Anderson, R. C. (1979). A cross-cultural perspective on reading comprehension. *Reading Research Quarterly*, *15*, 10-29.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (Research Report No. 94-5). Educational Testing Service.

Stocking, M. L. (1992). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report No. 93-2). Educational Testing Service.

Stout, W. (1990). A new item response theory approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*(2), 293-325.

Stout, W. (1987). A nonparametric approach for testing latent trait unidimensionality. *Psychometrika*, *52*(4), 589-617.

Swaffar, J. K., Arens, K. M., & Byrnes, H. (1991). *Reading for meaning: an integrated approach to language learning*. Englewood Cliffs, NJ: Prentice-Hall.

Tarone, E. (in press). In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* New York: Cambridge University Press.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace line for testlets: A use of multi-categorical-response models. *Journal of Educational Measurements*, *26*, 247-260.

Thomas, T. J. (1990). Item-presentation controls for multidimensional item pools in computerized adaptive testing. *Behavior Research Methods, Instruments, & Computers*, *22*(2), 247-252.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Educational Research Institute of British Columbia.

Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In G.H. Fischer * I.W. Molenaar (Eds.) *Rasch models: their foundations, recent developments and applications.* New York, NY: Springer.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185-201.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53*, 744-789.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *Modern Language Journal, 77*, 473-489.