One of our major goals is to consider the design and administration of Spanish tests for students at U.S. universities in light of the social implications attached to any specific testing (and teaching) framework. A second goal is to substantiate the need for test administrators to engage in the type of reflective practice (Schön 1983) that will lead them to adapt and modify as needed currently available tests to make them more appropriate to accomplish their specific teaching/learning objectives.

Currently, numerous methods are being used for assessing language in Spanish courses, including

- traditional fill-in-the-blank grammar tests;
- nth word or rational-deletion cloze tasks;
- multiple-choice and open-ended reading comprehension questions on a seen or unseen text;
- listening comprehension checklists of various kinds;
- structured and open writing tasks, usually in response to a prompt;
- structured or improvised oral interviews.

All of the above testing activities, as well as others, are regularly used in Spanish courses taught in most universities in the United States. The fact that these methods of assessment are used rather routinely, however, does not necessarily mean that they are reliable (i.e., that their use would produce the same results each time) or valid (i.e., measuring what they purport to measure). In fact, it may be a challenge to obtain an accurate measure of language ability in the classroom. Yet the construction of reliable and valid assessment measures can have crucial relevance in supporting learners in their efforts to develop Spanish language skills. Hence, it behooves language teachers to enhance their knowledge of what assessing Spanish language ability can entail and to update their knowledge of ways to assess this ability.

1.0 Methods to Assess Classroom Learning

In this section we will briefly describe some selected theoretical aspects of language testing in classrooms, concentrating on the qualities of a test and the models of language competence that inform the field of language testing.

1.1 Assessing the Usefulness and Relevance of a Test

Just as assessment may benefit from the use of multiple measures of language proficiency like the ones described in the previous section, so the worth of any assessment
instrument depends on a combination of methodological factors. Bachman and Palmer (1996) have identified six qualities that they would argue will determine the value of a language testing instrument:

1. Reliability: the consistency of measurement
2. Construct validity: an indicator of the ability we want to measure
3. Authenticity: the correspondence between the characteristics of the test task and the features of the real task
4. Interactiveness: the interaction between the test taker—including language ability, topic knowledge, and the affective situation—and the task
5. Impact: on society and the individuals
6. Practicality: the demands of test specifications can be met with existing resources

Bachman and Palmer (1996) warn us that the evaluation of the usefulness of a test is essentially subjective, predicated on value judgments as well as specific goals and conditions for the test. For instance, they point out that for large-scale testing, reliability and validity are likely to be crucial, whereas for most types of classroom testing, authenticity, interactiveness, and impact are the likely factors to be most relevant (19). Therefore, national standardized tests should not necessarily be viewed as better or more appropriate than locally produced tests. Furthermore, we specifically highlight the importance of the impact of a test, a factor that is sometimes regarded as irrelevant to determine a test’s ultimate overall usefulness.

Consistent with Bachman and Palmer’s list of test qualities, Byram (1997) would remind us that “foreign language teaching is a social phenomenon which is in part determined by the nature of the particular context in which it takes place. . . . The context includes the educational institution and the societal and geo-political factors to which educational institutions and the education system as a whole must respond” (87). Needless to say, Spanish instruction in the United States cannot remain oblivious to the realities of Spanish use in the United States (see Gutiérrez and Fairclough, this volume).

1.2 Models of Communicative Ability
In what has become the seminal work on communicative ability, Canale and Swain (1980) and Canale (1983) offered a four-component model of communicative competence in a second language: (1) grammatical competence (morphology, syntax, lexicon, phonology), (2) sociolinguistic competence (appropriate use of language), (3) discourse competence (cohesion and coherence), and (4) strategic competence (verbal and nonverbal coping mechanism used when communication breaks down). The model served to make certain distinctions that until that point had remained somewhat blurred. For example, they grouped those matters of discourse relating to cohesion (i.e., textual elements that link elements of the text) and coherence (i.e., the
comprehensibility of the text) within their own separate category, whereas prior to this they may have been subsumed within, say, grammar or perhaps even sociolinguistics. In addition, they added a category for strategic ability—perhaps the first official recognition in applied linguistics that strategic ability is not a given but appears to a greater or lesser degree, depending on the learner.

Some years later, Bachman (1990) provided his revised model of communicative ability in which he combined under organizational competence the grammatical and the discourse or textual aspects that Canale and Swain had separated. He also grouped under the category pragmatic competence both sociolinguistic ability and a new component he referred to as illocutionary competence. The latter was defined as the ability to understand and express a wide range of language functions, including but not limited to speech acts (e.g., promising, apologizing), ideational, and heuristic functions. He also included in his model strategic competence, though he left it outside of the scope of language ability per se (although it is still assumed to have an effect on language performance). In addition, he gave it a more rigorous subgrouping into strategies for assessing, planning, and executing language tasks (see Johnson 2001 for an in-depth analysis of these models).

A coauthor of this chapter further distinguished sociolinguistic ability into both sociocultural ability and sociolinguistic ability (Cohen 1994). Sociocultural ability refers to knowledge about (1) whether the speech act can be performed at all, (2) whether the speech act is relevant in the situation, and (3) whether the correct amount of information has been conveyed. Sociolinguistic ability, in contrast, refers to whether the linguistic forms (words, phrases, and sentences) used to express the intent of the speech act are acceptable in that situation (e.g., intensifying an apology for hurting someone physically with “really” to indicate regret, rather than with “very,” which may be more an indication of etiquette). Thomas (1995) also saw the need for a distinction of sociolinguistic ability into two categories, though she referred to them as sociopragmatic and pragmalinguistic ability, respectively. As is true with many dichotomies, this one has been criticized by Beebe and Waring (2001) for being too simplistic a distinction and one that is difficult to validate empirically. Still there is some conceptual salience in the sociopragmatic versus pragmalinguistic distinction, and so it continues to enjoy relative popularity in the field of L2 pragmatics and assessment of interlanguage pragmatic ability (see Roever 2004).

More recently, the early models have been revisited. He and Young (1998, 3), for example, contend that Bachman's model "is largely a psychological model that neglects the social, dialogic dimension of cognition and emotion—that is to say, cognition and emotion are not located in the mind of a single individual, but are instead embedded in distributed systems and are shaped and accomplished interactionally." They also assert that "we must now add competence in (at least) the five interactional features: Knowledge of rhetorical scripts, contextually-relevant lexicon and syntax, strategies for managing turns, management of conversation topics and the means to signal boundaries in a conversation" (He and Young 1998, 6–7). Furthering the argument about the role of interactional competence, Chalhoub-Deville (2003, 372) proposes
that “ability, language users, and context are inextricably meshed.” She goes on to add, “It is likely that the interplay between the more and less stable ability features is what researchers need to account for to explain situated language use” (Chalhoub-Deville 2003, 377). A brief introduction to models of communicative competence is found in McNamara (2000, chap. 2), and a more substantive coverage is provided in Johnson (2001, chap. 8).

2.0 Types of Tests

In this section we will critically analyze some concrete methodological issues that have relevance to classroom testing, namely, the distinction between discrete-item and integrative testing, the use of tasks as tests, the validity of real-life tests and semi-direct tests, and the use of complementary measures to assess L2 ability.

2.1 Discrete-item and Integrative Tests

Discrete-item tests focus on the testing of specific (discrete) aspects of language. For instance, a passage with blank spaces to be filled in with past tense forms of verbs provided in their infinitive form in parenthesis is an example of a discrete-item test of grammar. The main advantage of this type of test is that it can be easily designed, scored, and graded. In contrast, an integrative test of past tense use may be represented in the form of a personal narrative about some adventurous situation the test taker may have experienced in the past. It is clear that the latter type of test may have some advantages over the discrete-item test. For one thing, it may appear to the test takers to have a higher level of face validity—that is, to seem like a more realistic measure of language use than the discrete-point focus-on-form approach. Second, an integrative test, by definition, brings several aspects of language competence together. Here are a few examples of more integrative testing formats:

Cloze
• A text with every nth (e.g., 7th word) deleted or with words deleted on some rational basis (e.g., key function words or major content words).
• Reverse-cloze, where students decide which words have been added to a text.
• A C-test, where the second half of every second word is deleted.

Dictation
• A traditional written dictation, delivered at a slow pace.
• A dictogloss where the passage is read at the pace of natural speech. Students take note as they listen and then they are given the chance to fill in missing information.
• Oral repetition, where students repeat or reproduce orally what they have heard.

Summary
• Students need to identify the main ideas while reading a text and then organize these into a coherent summary of the text.

In an effort to avoid being simplistic in making this distinction between discrete-point and integrative measures, we need to point out that some rational-deletion cloze
tests are discrete-point in actuality. When we have interviewed learners about how they have answered cloze tests (Cohen 1984), we have found that they may well have treated sections of such tests as local, focus-on-form exercises rather than as exercises in more integrated language processing. Likewise, it is possible to give a dictation where the focus is just on, say, students’ ability to use the appropriate tenses of the verb. Then it would be an integrative task in principle but used in a more discrete-point manner.

2.2 Using Tasks as Tests
The notion of using a series of tasks to serve as a test has been around for many years, though it has looked more like a classroom project than a test per se (e.g., Swain 1984; Brill 1986). More recently, Hughes (2003) describes another example of a series of writing tasks on the same topic that can be assigned to elicit many representative samples of the test taker’s writing ability. In his test, Hughes proposes the use of four writing tasks centered around the theme of work at a summer camp for children:

1. Having learners write a letter to inquire about a position at a summer camp (the period of employment, accommodations, the pay, and the like)
2. Having them fill out an application form
3. Having them send a postcard to a friend telling him or her where they are, why they are there, and two things they like about the summer camp
4. Writing a note to their friends to apologize for not being able to meet them and to suggest a different day to go out

As the reader can see, all of the examples described above are written tests. It is obvious that a written test represents a more efficient way of collecting performance data from students. The same tests, however, can be easily transformed into oral tasks should teachers desire to do so.

In recent years, task-based instruction has gained considerable prominence. Consistent with this trend, Norris (2002, 343) would argue that task-based assessment has a key role in the classroom as a type of performance test in that it

• serves to determine whether a test taker “can use the target language to engage in and accomplish a given task” (e.g., an exchange student convinces his or her host family to let him or her travel to a nearby city and stay overnight on his or her own);
• focuses on “complex tasks and the criteria by which they are judged beyond the instructional setting” (expressing embarrassment after spilling coffee over another customer at a local coffee shop—this task requires much more than simply linguistic information to be accomplished successfully);
• can be “based on criteria specific to a given genre, setting or audience” (e.g., asking a friend’s mom for more food while having dinner at his or her house as opposed to making the same request from our friend while having dinner alone with him or her.)
A performance assessment instrument is defined by the following three characteristics: (1) examinees must perform a task, (2) the task should be as authentic as possible, and (3) success or failure is based on the outcome of the task (Norris et al. 2002). We note that these features of task-based tests are also features of task-based instruction in general. Willis (1996), for instance, outlines the various stages of task-based instruction. Lest we embrace task-based assessment too quickly, Bachman (2002) would offer several caveats. He warns that there are two serious challenges in the design of tasks: (1) precisely how “real-life” task types are identified, selected, and categorized, and (2) how we actually go about linking pedagogic or assessment tasks to these task types. His concern is that vagueness in task specification inevitably leads to vagueness in measurement.

### 2.3 Real-life Tests

The Oral Proficiency Interview (OPI), a performance test originally developed by the Foreign Service Institute (FSI) and later adapted by the American Council for the Teaching of Foreign Languages (ACTFL), uses the criteria in the ACTFL guidelines (1986, 1999) to assess mostly speaking proficiency. Most important, second language performance tests, such as the ACTFL-OPI test, have been portrayed as authentic real-life direct tests of second language ability. Barnwell (1996, 151), however, warns us about the “test makers’ traditional hubris: the fallacy that a test always measures what its designers say it measures” (see also Lantolf and Frawley 1988, for the same critique). With regard to its authenticity in particular, there are important objections to the claim that the ACTFL-OPI measures real-life oral conversation (e.g., Johnson 2001; van Lier 1989). For instance, Johnson argues that “the OPI lacks both the empirical evidence and theoretical rationales to justify the claim about the conversational nature of its interaction” (143). Furthermore, there are concerns about the validity of the ACTFL-OPI test. In fact, even proponents of the OPI, such as Dandonoli and Henning (1990, 11), acknowledge that “the most significant” criticism against the use of the ACTFL-OPI is that there is no study that supports the validity of such a testing procedure. Finally, it appears that the ACTFL-OPI makes use of an outdated concept of validity to justify its claims. For instance, Messick’s (1994) reconceptualization of validity, specifically incorporated into Bachman’s (1990) model of second language assessment, has not become part of the theoretical framework of the revised version of the ACTFL tester training manual, published in 1999 (cf. Johnson 2001; Salaberry 2000a).

### 2.4 Semidirect Tests

Semidirect tests simulate an oral interview through prerecorded questions that the respondent is to answer in a recorded session. The Simulated Oral Proficiency Interview (SOPI) is the best known of such measures (see Malone 2000). The SOPI follows the general structure of the OPI, but it relies on audiotaped instructions and a test booklet to elicit language from the examinee. The SOPI makes an effort to contextualize tasks so that they appear as authentic as possible. The prototypical SOPI
follows the same four phases as the OPI: warm-up, level checks, probes, and wind-down. The warm-up phase, designed to ease examinees into the test format, begins with background questions, then level checks and probe phases follow, assessing the examinee’s ability to perform different functions at the ACTFL’s Intermediate, Advanced, and Superior levels. The prototypical SOPI includes picture-based tasks that allow examinees to perform activities such as asking questions, giving directions based on a simple map, describing a place, or narrating a sequence of events based on the illustrations provided.

The Center for Advanced Research on Language Acquisition (CARLA) at the University of Minnesota has developed a tape-mediated instrument for assessing speaking, the Contextualized Speaking Assessment (CoSA), requiring students to listen to a master cassette. The test can also be administered to large groups in a language lab or large room. After listening to the instructions, a sample response, and a description of the overall theme for the test, the test takers are presented with situations and topics for their responses. All instructions are provided in English. (See www.carla.umn.edu/assessment/MLPA/CoSA.html for more on the CoSA.)

2.4.1 Do Direct and Indirect Tests Measure the Same Construct?
Stansfield and Kenyon (1992) report correlations of .89 and .95 between the OPI and the SOPI in various languages. Considering the substantive criticism against the conversational nature of the face-to-face OPI (e.g., van Lier 1989), such high correlations may exist because neither test allows candidates to demonstrate interactive skills. Despite the findings from Stansfield and Kenyon, however, Shohamy (1994) argues that the functions and discourse features elicited in the face-to-face OPI and the tape-mediated SOPI are not necessarily the same. Along the same lines, Chalhoub-Deville’s (2001) analysis of data from the OPI, the University of Minnesota’s CoSA, and the San Diego State’s Video/Oral Communication Instrument (VOCI) revealed that each test seemed to be tapping different language abilities. With respect to the testing of Spanish in particular, Koike’s (1998) empirical study revealed differences between the OPI and the SOPI, especially with regards to the effect of the live interlocutor (e.g., more turns, more use of English).

2.4.2 Advantages of Semidirect Tests
There are various factors that can make a semidirect format attractive and help explain why it has been used widely in China and elsewhere in the world. For one thing, it allows for a uniformity of elicitation procedures, which helps to promote reliability. In addition, it is economical to administer, since there is no need to hire test administrators to interact with each respondent. Furthermore, it eliminates the interview effect which can play a role in oral interviews. Brown (2003, 3), for instance, argues that interviewers may influence the outcome of an interview by means of factors such as their level of rapport with the test takers, their choice of topics and functions, their phrasing of questions and prompts, and the extent to which they accommodate to the test taker’s abilities. The dilemma is that differences in interviewer
reactions to various test takers are actually supportive of the nonscripted natural variation we normally find in most conversations, where a change in interlocutor will often naturally lead to changes in the type of interaction. Brown (2003, 20), nevertheless, claims that such changes in interviewer behavior may turn out to be relevant for the construct of language proficiency.

To help underscore the inconsistency across interviewers, Brown (2004) conducted a study that demonstrated how two different interviewers could rate the same non-native English speaker differently. Through close conversational analysis of the two interviews, she demonstrated how oral assessment instruments can be dramatically different from ordinary conversation and how the individual interviewer conducts the session can sway the subsequent ratings made by outside raters. In our view, for the sake of economy, the semidirect approach is the only feasible alternative for classroom Spanish teachers with limited resources who wish to obtain a measure of speaking from all students. An alternative mixed format would be to have students interacting with each other with prompts from a computer or a tape. As Hughes (2003, 121) points out, “An advantage of having candidates interacting with each other is that it should elicit language that is appropriate to exchanges between equals.” This mixed format, although in principle potentially as valid as any one of the other test formats, has not been sufficiently researched, nor have teachers used it often enough yet in order to have substantive information on its usefulness.

### 2.5 Complementary Measures to Assess Second Language Ability

Regardless of efforts made to ensure that a given measure is a true estimate of a learner’s ability in that area, there are bound to be method effects. These effects result from differences in discourse tasks (e.g., reporting vs. interviewing), elicitation methods (e.g., personal vs. machine, direct vs. indirect), genres (e.g., narrative vs. expository texts), item types (e.g., multiple-choice vs. open-ended tests), and even test consequences (e.g., declared vs. undeclared purposes of tests) (Shohamy 1997). To address the limitations of a single testing instrument, several researchers have promoted the use of a battery of alternative assessment instruments in complementary ways (e.g., Liskin-Gasparro 1996; Lynch 1997). Spolsky (1997, 246) expresses it best when he states, “What we are starting to do, I am pleased to see, is accept . . . the inevitable uncertainty, and turn our attention to the way tests are used, insisting on multiple testing and alternative methods, and realizing that the results need cautious and careful interpretation.”

An early effort at demonstrating how a multiple-measure instrument can improve on measurement was that described by Shohamy, Reves, and Bejarano (1986). They report on the construction of an oral proficiency test to replace the existing English as a Foreign Language (EFL) Oral Matriculation Test administered by the Ministry of Education in Israel. The results showed that the experimental test had better linguistic, educational, and testing qualities than the existing Oral Matriculation Test; namely, it produced a better distribution of scores, showed reasonable rater reliability, tested a broader range of speech styles, and produced favorable attitudes on the part of the test
takers. A somewhat scaled-down variety of this multimeasure test is still being used in Israel a decade later.

Thus language assessment can involve various types of measures, from the more traditional formats to the more innovative ones, such as the following (adapted and expanded from Liskin-Gasparro 1996):

- Portfolios (sample materials plus reflective assessment/longitudinal)
- OPIs and SOPIs (face-to-face or tape-mediated interviews as described above)
- Mixed OPI and SOPI (students interact with each other and the tape)
- Computer adaptive testing (competency level modified as responses are entered)
- Tasks (news broadcast program, writing letters, preparing websites)
- Self-assessment (usually guided with specified criteria)
- Collaborative assessment (both teachers' and students' opinions)
- Learning logs (students quantitatively evaluate their own progress)
- Journals (students qualitatively evaluate their own progress)

3.0 Testing Effects on the Spanish Curriculum

In this section we analyze the relationship between testing objectives and program/course objectives. For that purpose we will identify and describe the pedagogical role of tests and, in particular, the washback effect of tests on teaching.

3.1 Testing Objectives

Tests are usually defined according to specific objectives such as achievement versus proficiency tests, formative versus summative tests, or process- versus product-oriented tests. Achievement tests focus on measuring whatever topics and components of language abilities were taught in a given course, whereas proficiency tests measure language abilities independently of the process of acquiring such language competence (Shohamy 1992). Achievement tests would be expected to restrict themselves to that material which is specified by course objectives, whether it be focused more broadly on communicative language abilities or more narrowly on grammatical functions. The challenge, however, is to clearly identify such program objectives.

For instance, assessing the achievement of the objective of “successfully paying a compliment in Spanish” may be easier said than done. For one thing, there are several dimensions to be taken into account to measure the successful accomplishment of complimenting. Notice that not only is verbal information relevant for meeting this objective but also paralinguistic (e.g., intonation) and physical cues (e.g., face gestures, hand movements). Second, as with any other speech act, there are various complex social cues that need to be taken into account to successfully compliment someone in Spanish (e.g., the relative status and relationship of the interlocutors, the setting, and the expected responses to compliments in that society.). Finally, we note that it is not easy to developmentally grade levels of complimenting. For instance, if a learner is not successful at complimenting someone else, should we conclude that this learner conveyed a less intense sense of the compliment? The problem with this approach is that it is not easy to measure what these different degrees of complimenting mean or how
they should be measured. In other words, can complimenting be graded? Or should we consider it a dichotomous category? In sum, the dichotomy between achievement and proficiency assessment is not a clear-cut distinction since the same test items and tasks could be assessing both. The only factor that might make a task involving complimenting an achievement item rather than a general proficiency one would be a very narrow definition and specification of what the teacher had specifically taught in class to compliment others.

There are other distinctions of note as well, such as that between formative assessment, aimed at getting an ongoing picture of performance, and summative assessment, which is intended to assess learner achievement and program effectiveness after a determined period of time, such as a unit of instruction, a semester, an academic year, and so on. Associated with formative versus summative assessment is the distinction between assessing the process whereby learners perform a task and the product obtained. So assessment could, for example, look only at a finished written composition or measure incremental gain through a series of drafts (as would be the case with a writing portfolio). Additionally, tests may be created for placement purposes, including the awarding of advanced credit. For instance, several commercial tests in Spanish are geared toward this goal: College Board’s CLEP (College Level Examination Program), SAT II, and the Educational Testing Service’s Spanish AP (Advanced Placement) test.

It also is possible to gather information about the strategies that the learners use and evaluate their selection of strategies in the processing stage of the task (independent of how successful they are with the product) (see Cohen 1998a; in press). A final distinction is between internal assessment, aimed at giving feedback to the classroom teacher, participating students, and perhaps the parents, versus external assessment, which is meant to inform the school district, the language program, an association such as ACTFL, and even the federal government (e.g., the National Assessment of Educational Progress has recently generated a test of Spanish for administration nationwide in the United States).

3.2 Washback Effect
Whereas Bachman and Palmer propose six qualities to account for the usefulness of a test, Hughes (2003) would limit the list of qualities to four: validity, reliability, practicality, and a new category not explicitly mentioned by Bachman and Palmer: "beneficial washback." Washback refers to the impact of the language assessment measures on the teaching syllabus, the course materials, and the classroom management (Taylor 2004). Washback can be beneficial, as in the case where changing or instituting language measures leads to beneficial changes in teaching and curriculum. Washback can also refer to negative effects, such as when the testing program fails to recognize course goals and learning objectives to which the test is supposed to relate (Cheng and Curtis 2004). Taylor (2004, 143) points out that the impact is now often used to describe the washback or consequences that a test may have not just at the micro, or local, educational level, but also its impact at a macro, or
societal, level. Both the local and macro influence contribute to the consequential validity of the test.

The link between curriculum and assessment of the objectives pursued in a curriculum needs to be bidirectional. That is, changes in the curriculum should have a beneficial impact on testing content and on criteria as well. Consider, for instance, how notions and goals of cross-cultural awareness are routinely mentioned in Spanish L2 program objectives. Barnwell (1996, 185), for instance, points out, “It is not uncommon to hear language teaching justified in terms of rather tenuous notions of opening students’ minds to other cultures, imparting a more sophisticated awareness of the nature of human language itself.” The concern is for whether such lofty goals are actually implemented in course objectives and syllabi and, more important, whether they are carried out in actual classroom practices.

4.0 Testing Pragmatics and Cultural Knowledge

In this section we analyze in detail two aspects of Spanish L2 competence (pragmatics and culture) that tend to be neglected in the explicit testing of the target language.

4.1 Can We Test for Knowledge of Spanish Pragmatics?

An important component of communicative language competence that is frequently glossed over in the testing of Spanish is that of pragmatics and, more specifically, speech acts (e.g., apologizing, requesting, complimenting, and complaining). Pragmatics focuses on the functional use of language within a social, cognitive, and cultural context (see Koike, Pearson, and Witten 2003). The relevance of teaching and testing pragmatic knowledge cannot be overemphasized, given the importance of pragmatic abilities for communicating successfully in the second language and the daunting challenges facing learners in attempting to be pragmatically appropriate.

Numerous research studies have documented the role of pragmatics. Let us take, for example, a study that compared the linguistic expression of Spanish and English speakers in their own native language in six specific situations that prompted various speech acts such as requests and apologies (Fulcher and Márquez Reiter 2003, 335). In the first situation the participants had to borrow a book from a professor. The following were typical ways in which English and Spanish native speakers phrased their requests in their native language:

1a. I was just wondering if you have the book and if I could borrow it?
1b. ¿Me puedes prestar el libro? ‘Can you lend me the book?’

Fulcher and Márquez Reiter concluded that “English speakers used more conditional or embedded conditional sentences than the Spanish speakers. They also used softening devices and provided reasons for making a request.” Needless to say, full proficiency in Spanish would imply knowledge about how to express specific speech acts in ways that are nativelike. Yet nonnative speakers of a language may take years to master these speech functions, if at all. And in fairness to students, instructional programs usually do not provide adequate instruction in this area (see Cohen and Ishihara 2005).
Notice that important differences in pragmatic information may exist even in cases where the surface utterances are almost direct equivalents of each other across the languages. For instance, Koike, Pearson, and Witten (2003) describe how the direct translation of a suggestion in Spanish to English can be misleading at best or confrontational at worst. Spanish suggestions expressed in the negative form acquire a much stronger connotation in English:

2a. ¿No has pensado en leer este libro?
2b. Haven’t you thought about reading this book?

It is open to question whether Spanish learners can actually offer suggestions with this particular phrasing of the question in the negative form.²

Ironically, the type of test that the teaching profession was hoping would help assess language abilities beyond grammar or vocabulary, the real-life performance test exemplified in the ACTFL-OPI, failed to deliver on its promise. For instance, Raffaldini (1988) pointed out that the ACTFL-OPI evaluated numerous aspects of grammatical competence and certain aspects of discourse competence, such as grammatical and lexical features in cohesive discourse. But the OPI, she argued, did not properly or thoroughly evaluate sociocultural and sociolinguistic abilities. Raffaldini proposed that one particular component of the traditional OPI, the role play, be used more extensively and across all levels of proficiency to address this serious shortcoming of a test of overall communicative language ability.

Raffaldini’s Oral Situation Test was intended to assess more areas of language proficiency in a wider range of language-use situations than the ACTFL/ETS Oral Proficiency Interview. It was aimed at college-level study-abroad students to France and added something that other measures were not assessing, namely, tone (e.g., courteous, regretful, persuasive). The following are two examples from the test:

**Tone:** persuasive; **Stimulus:** You will be leaving France in a few weeks and all the students in the program would like to get together for a final party. The only place big enough is the house where you are living. You ask the parents if you can have the party there. You say:

**Tone:** annoyed; **Stimulus:** The parents of the family with whom you are living have gone away for the day and left you in charge of their little boy. He went out to play and disappeared for quite a while. You went out looking for him but couldn’t find him. When he finally returns you are upset at what he has done and tell him not to do it again. You say:

The rating scales for the oral test were comprehensive, including ratings for discourse competence, sociolinguistic competence, and linguistic competence. It would appear that the call to assess pragmatic tone fell largely on deaf ears because now, many years later, Beebe and Waring (2002, 2004a, 2004b) are once again raising the issue since tone has continued to be neglected in measurement.

Along the same lines, Cohen (2004) has offered a basic framework for the teaching and testing of speech acts:
1. Keep the speech act situations realistic for the students and engaging.

2. Check for the sociocultural (= sociopragmatic) appropriateness of the strategies in the given situation and the appropriateness of the sociolinguistic (= pragmalinguistic) forms used with regard to level of formality, degree of politeness, and amount of language used.

3. Have a discussion afterward with the students as to whether the setting was clear and as to the factors that most contributed to the students’ responses.

4. Use verbal reports to help in reconstructing why the students responded as they did.

Cohen (2004, 320–21) described how learners could be asked to reconstruct retrospectively (while viewing their own videotaped speech act performance) the processes that they went through while responding to prompts that required the use of specific speech acts, and to describe the strategies that they selected in performing the given speech acts. García (1996, 2001) has specifically adapted some of Cohen and Olshtain’s (1993) recommendations for the teaching of Spanish speech acts. In her earlier publication she described how results from sociolinguistic research studying a group of Spanish speakers declining an invitation, along with models from several researchers, led to the design of listening and speaking activities for developing students’ ability to communicate and their avoidance of cross-cultural miscommunication (García 1996).

4.2 Can We Test Cultural Knowledge in Spanish?

An even more elusive target than the testing of pragmatically appropriate uses of Spanish language is the testing of cultural knowledge. As argued by Byram (1997), second language speakers are attempting to attain several goals at once when they engage in a sociocultural analysis of the uses of the second language:

- To be pragmatic, by attempting to communicate appropriately with native speakers of the target language community
- To be critical, by trying to understand others
- To be hermeneutic, by getting to understand oneself in the process

But is it possible to teach and test culture and in so doing to assess the degree of critical thinking brought about by second language learning?

Moore (1994, 164) argues that “testing culture has traditionally measured the knowledge of bits and pieces of information, rather than insights and awareness of the essence of a culture or society.” In fact, in a review of various proficiency rating scales, North (2000, 95) concludes that “inter-cultural skills are an aspect of Socio-cultural Competence not found in any of the scales analyzed.” Echoing the opinions held by Moore and North, Storme and Derakhshani (2002, 663) stated that “the profession has only begun to give serious thought to developing the requisite measures to cultural proficiency.” Arguably, at least part of the blame for the lack of adequate focus of classroom and curricular practices on the development of cultural competence—beyond
the trivial and superficial facts—can be associated with testing practices and logistical concerns. Byram (1997, 111), for one, recognizes the difficulties in assessing the overall range of complex competences that make up intercultural communicative competence: “It is the simplification of competences to what can be ‘objectively’ tested which has a detrimental effect: the learning of trivial facts, the reduction of subtle understanding to generalizations and stereotypes, the lack of attention to interaction and engagement because these are not tested.”

Identifying the intercultural skills that should be taught and implementing appropriate assessment measures that would hopefully have a beneficial washback effect on the curriculum may be a challenging task. In order to understand the true nature of this challenge, let us define in more detail the actual objective of teaching and testing culture. For one thing, we need to remember that learning a second language cannot be simplistically reduced to becoming a monolingual speaker of that language because, in fact, learners add a second language to the first (not to mention the fact that a second language could also have effects on the first one). With regard to culture in particular, North (2000, 95) argues, “The curriculum aim of ‘intercultural skills’ is to create ‘150% persons’ who understand (empathy), find value, and have positive sentiments towards (favorableness) both cultures.”

Second, the assessment of intercultural skills is multifaceted. For instance, the National Standards in Foreign Language Education Project (1996, 439) established that learners “must be able to participate appropriately in a range of social relationships and in a variety of contexts.” It posits further that the “capacity to communicate requires not only an awareness of the linguistic code to be used, but also an understanding of the cultural context within which meaning is encoded and decoded.” Third, there are no right or wrong answers when it comes to the testing of culture, and although this same point can in principle be made about other aspects of a second language (such as morphology or syntax), it is clear that such variability is most noticeable when it is predicated on the assessment of cultural knowledge. With regard to the testing of intercultural pragmatic ability (which clearly involves the intricate interweaving of language and culture), coauthor Cohen (2004, 322) explains that “sociocultural and sociolinguistic behavior are by their very nature variable. Thus, there will be few ‘right’ and ‘wrong’ answers in comparing L2 to L1 responses, but rather tendencies in one direction or another.” Therefore, he concludes that “the variable nature of speech act behavior has made tested outcomes less reliable and valid than those for more circumscribed language performance and helps explain why such measures do not abound in the field.”

There are, nevertheless, proposed solutions to the problem of testing cultural knowledge. Byram (1997, 103), for instance, argues that for the assessment of intercultural competence, language testers need to take into account metacognitive capacities for self-analysis, proposing that “neither the testing of knowledge nor the evaluation of observable performance are sufficient. It is in the self-analytical and often retrospective accounts by a learner of their interaction, their savoir faire and savoir s’engager, that the main evidence will be found.” Storme and Derakhshani (2002, 663), however, conclude
that some of Byram's recommendations for the testing of intercultural knowledge may not be appropriate or practical for a second language classroom. They, in turn, suggest that the program of action outlined in the National Standards project may be more suitable for the teaching and testing of cultural knowledge.

5.0 Spanish Developmental Sequences and Testing

We turn now to the discussion of the benefits of taking into account potential developmental sequences in the acquisition of Spanish in an effort to make language testing procedures more congruent with learning processes.

5.1 What Do We Know about Developmental Sequences in Spanish?

North (2000, 13) suggests that scales of proficiency "have the potential to exert a positive influence on the orientation, organization and reporting of language learning." North sees the potential for scales in that they provide learners with

1. explicit goals and descriptions of them,
2. coherent internal links for curriculum development and testing,
3. behavioral evidence of progress,
4. a means for increasing the reliability of subjective ratings, and
5. a common metric for comparisons among different populations of learners.

North acknowledges, however, that scales of proficiency have serious limitations if they do not conform to actual developmental stages of acquisition (see Brindley 1988, for an extensive discussion). The problem is that we still know very little about developmental sequences in language acquisition in Spanish as the chapters in Lafford and Salaberry (2003) show.

Such paucity of clear research findings about stages of acquisition should make us wary of blindly following traditional and almost categorical sequences of acquisition typically espoused by publisher's textbooks. For instance, there are well-attested developmental patterns and developmental processes in the acquisition of Spanish past tense morphology in both classroom and naturalistic environments (Andersen 1986; Salaberry and Ayoun 2005; Salaberry 2000b; Schell 2000; Shirai 2004). Learners in classroom environments are clearly focused on the acquisition of morphological markers of Spanish tense aspect. Furthermore, there appears to be a tendency to use some default markers of past tense during beginning stages of development, gradually incorporating a past tense marking system that is shaped by frequency-based distributional tendencies, and eventually using more sophisticated notions of viewpoint aspect marked by discursive grounding (e.g., García and vanPutte 1988). This is one of the areas of L2 development more akin to Pienemann's (1988) agenda of profiling L2 development, and an area where testing procedures could match developmental stages of acquisition that are now becoming better known to researchers. Other components that make up any one of the models of L2 competence reviewed above are
also subject to similar analyses of developmental stages of acquisition. For a more in-depth analysis, the chapters in Lafford and Salaberry focus on developmental trends in the acquisition of pragmatics, pronunciation, the subjunctive, object pronouns, tense and aspect, and vocabulary.

5.2 Developmental Nature of Institutionalized Rating Criteria

We now turn to the analysis of developmental criteria used in one specific test of Spanish as a second language. The ACTFL guidelines constitute a widely recognized standardized criteria used to measure second language proficiency in U.S. universities. The guidelines, although successful in bringing the attention of teachers to the use of standards in language assessment, are nevertheless problematic with respect to developmental criteria. Arguably, the more troublesome aspect of the guidelines is the notion of an implicit degree of complexity (linguistic or cognitive) assigned to particular registers, discourse genres, or topics of discussion. Our discussion will lead us to the conclusion that Spanish teachers should view such a set of scales with caution.

The ACTFL guidelines assign particular language abilities to specific stages of development (or levels of proficiency): “Each major level subsumes the criteria for the levels below it” (ACTFL 1986, 2–5). For instance, the guidelines propose that only Superior level speakers are able to “explain and defend opinions and develop effective hypotheses,” that not until the Advanced level can speakers “narrate and describe in major time frames with good control of aspect,” and that not until Intermediate level can speakers “obtain and give information by asking and answering questions.” Along the same lines, the ACTFL hierarchy proposes that the formal registers of the language are not managed successfully until learners are at the Superior level, and that speakers cannot control most informal and some formal registers until they are at the advanced level. More specifically, some informal registers are assumed to be controlled at the Intermediate level and only the most common informal settings are successfully controlled by novices. Similarly, only Superior level speakers are able to manage abstract and unfamiliar topics whereas concrete and factual topics are controlled at the Advanced level.

To the best of our knowledge, however, there is no empirical evidence that shows that classroom learners, for instance, will develop proficiency in informal registers of the language before they achieve such proficiency in formal registers—or, for that matter, that they will be able to discuss concrete and factual topics before they manage more abstract topics or from familiar to unfamiliar. For instance, it would not be out of the question to propose that formal registers of the language may actually be easier to learn for classroom learners for several reasons. For one thing, native English speakers learning Spanish have access to a lexicon based on both Germanic and Latin roots, the latter mostly represented in formal registers of English (compare “liberty” and “velocity” [Latin roots] with “freedom” and “speed” [Germanic roots]). Furthermore, the standard register of the language, the academic subjects of discussion, and academic discourse in general (not surprising given the university setting in which this type of instruction is embedded) are typically favored in classroom instruction (see, for exam-
ple, a learner’s testimonial in a diary study with regard to university-level Japanese instruction in Cohen 1997). More important, the more standard forms of vocabulary and syntax are typically preferred in classroom instruction in order to avoid dealing with any type of dialectal variation.

More important, proponents of ACTFL acknowledge this significant limitation about the hierarchical (developmental) distribution of communicative functions. For instance, Lowe (1985, 47) states, “In everyday life we tend to speak at Level 3, with forays into higher levels as required for technical topics.” Thus one could argue that the levels may not be hierarchical but complementary, as they represent, according to the distinction made by Lowe, different registers of the language. Following this line of argumentation, one could reasonably expand Lowe’s claim and argue for the complementarity of discourse genres, registers of the language, and other components of language competence, as opposed to assuming a hierarchy as proposed by the ACTFL scale. In practice this means that teachers can introduce learners to the linguistic features of several registers and genres of Spanish without concerns about the students not being developmentally ready to put such sociolinguistic information into use. In fact, it may even be more pedagogically sound (in keeping with the proposal of the National Standards) to explicitly point out to learners the linguistic and interactional contrasts that different genres and registers exhibit in Spanish.

Furthermore, the developmental hierarchy of skills and functions proposed for listening and reading tasks by the ACTFL guidelines has been questioned on both theoretical and empirical grounds. For instance, Phillips (1988, 138) concedes that the hierarchical skills or competencies described in the guidelines for reading may not be developmental after all. Lee and Musumeci’s (1988) findings lend empirical justification to making the same argument against the hierarchical nature of the reading scale. Similarly, the developmental progression for listening abilities is also questioned by Valdés et al. (1988). In principle, the hierarchy of developmental stages proposed by the guidelines (1986, 1999) may represent a possible theoretical hypothesis.

In sum, despite their conspicuousness in the profession, the ACTFL criteria for testing Spanish in the classroom are not particularly helpful because they are nondevelopmental (see above), vague, and disregard the relevance of the learning context. The main problem with using scales such as the ACTFL guidelines for testing Spanish learning in the classroom is that “the progressions described ... could appear to present a picture of universal patterns of second language development” (Brindley 1988, 133). At the same time, we do recognize that the ACTFL guidelines have become the only widely available set of language proficiency criteria. Thus in our opinion, a practical solution to this dilemma is to adapt the current rating criteria of proficiency (from ACTFL or other models) as necessary, as it is feasible to incorporate changes apparent in our current understanding of SLA.

6.0 Recommendations for Classroom Testing

Even though it is true that throughout the 1980s and 1990s language teachers were particularly concerned with standardized measurement of proficiency (Liskin-Gasparro
1996, 173), the day-to-day operation of a language program relies heavily on language tests developed by textbook publishers. Barnwell (1996, 188) explains that “in the foreign language classroom many, if not most, of the tests used come with the textbook that has been adapted [sic]. Hence, in one sense the most influential testers are those who write the test manuals for publishers and provide teachers with entire banks of tests and items to measure the progress and achievements of their students.” Given the importance of tests developed by publishers, Barrette (2004, 68) suggests “putting pressure on publishing companies to provide testing programs . . . that are written by the text’s authors to increase the level of comparability between the teaching and testing approaches, thereby providing a quality model for other tests developed by the textbook adopters.” In the meantime, however, language teachers need to be aware of some practical procedures at their disposal to improve and expand the tests made available to them by publishers.

We believe that the following considerations regarding the design and writing of tests for the Spanish classroom will address some of the issues we raised in previous sections of this chapter and will hopefully provide a clear framework for the assessment of interactional communicative competence.

6.1 Giving Primary Importance to the Context of Learning

Classroom interactions will likely have a significant effect on the acquisition of the L2, not only in terms of content areas but also in terms of preferred language learning and language use strategies. For instance, students from various regions of the country not only relate more or less to different topics that may or may not be relevant locally but also bring with them specific predispositions toward learning languages related to academic specializations, social networks, and so on. As Chalhoub-Deville (2003, 377) points out, the “language user has a set of preferred abilities that are typically activated in contexts with particular features. The more familiar the language user is with these ability structures—contextual features, the more efficient and fluid learners become at activating them: combining and recombining knowledge structures as needed to engage in a given situation.” She further notes that “variation is inevitable if we view ability within context as the construct” (Chalhoub-Deville 2003, 379). Johnson (2001) would underscore the importance of drawing on local rather than general models of language ability and use these for measuring sociocultural competence. For instance, regions where Spanish-English bilingualism is common may bring about specific dialectal features of Spanish that cannot be ignored from our model of language competence and ideally should be incorporated to it.

6.2 Identifying, Describing, and Operationalizing the Goals of Learning as Objectives

The use of any one of the communicative ability models briefly described in section 2 (or a modification of them) represents a good first step to determine the specifics of what a particular Spanish course is intended to achieve in terms of learning goals. For instance, programs which are intent on developing learners’ awareness of the prag-
matic, social and cultural aspects of the target language group cannot remain oblivious
to the most recent changes in the definition of communicative ability, especially with
regards to the notions of interactional competence, situated language use and overall
pragmatic abilities. On the other hand, programs that for specific reasons prefer to
maintain a focus on more traditional aspects of language learning such as the structure
of the language should specifically and explicitly identify such goals as part of their
objectives. In our view, the recent trend to operationalize almost any conceivable learn-
ing goal as a part of a program’s objective is not only unrealistic but also detrimental to
the achievement of the specific objectives that are indeed targeted by any program.

6.3 Testing the Course Objectives
We would concur with Hughes (2003, 13–14) in his recommendation that test con-
tent be based on course objectives (as opposed to course content) in that this “will
provide more accurate information about individual and group achievement, and it
is likely to promote a more beneficial backwash effect on teaching.” In addition, we
would recommend incorporating the testing of abilities/strategies/processes that will
further develop the L2 outside of the classroom environment/after the course. For
instance, conversational management techniques and awareness of register and
dialectal differences could also be assessed.

6.4 Obtaining a Robust Sampling of the Course Objectives
Not only should course objectives be sampled, but those language interactions that
take place regularly in classroom interactions should be included in testing tasks as
well. This would assure better congruency between actual classroom language and
what is measured on tests.

6.5 Writing Good Tests
Despite all good intentions, tests are not necessarily well written. Barrette (2004) has
recently identified five areas in which a series of draft language achievement tests
written by college-level instructors were lacking:

1. Criteria for correctness: Criteria for scoring writing sections were vague.

2. Weighting of the test components: There was a lack of correspondence in the
weight assigned to the different components of the test with regard to the goals
of the program (e.g., assigned points tended to favor the grammar component
of the test).

3. Length of the input: Inconsistency in the length of both listening and reading
passages, which, as Bachman (1990) noted, could heighten the potential effects
of the other characteristics on performance.

4. Representation of the construct: The intended language component was under-
represented—tests elicited only “the most common vocabulary items and the
most regularized grammatical forms.”
5. Extraneous factors: Ambiguous questions and overly demanding tasks created what would amount to error variance in the measures.

Among various recommendations for improving the writing of tests proposed by Barrette, we highlight the following:

- Reviewing test drafts both from the perspective of students and from that of scorers (see also Hughes 2003, 62–65)
- Making explicit and unambiguous the scoring criteria and the standards for correct target language use
- Making judicious judgments about what is taught in class and what is tested, especially with regard to the relative importance of different components of the test and the relative length of aural and written texts
- Engaging in networking (Barrette suggests sharing the burden of materials development by sharing test materials among institutions using the same textbooks or similar programs of study)
- Making ample use of verbal reports as a means of validating the measures (verbal reports [see Cohen 1998b, chap. 3] may be useful to assess the effectiveness of testing materials and to determine whether the objectives of the testing program are actually met and which changes may be needed)

7.0 Conclusion

In closing, we summarize the major points that we have made in this chapter. We have argued for the importance of using a clear definition of teaching and testing objectives as exemplified by the many models of communicative competence. We have briefly described some of the most relevant formats of tests for classroom learning, and we have advocated reaching a balance between efficient and valid measures of language proficiency. Further, we have highlighted the symbiotic relationship that exists between teaching and testing and the implications of changing testing procedures without considering concomitant changes in teaching processes. We have advocated the expansion of the traditional objectives of language tests to explicitly incorporate central aspects of later communicative ability models such as culture and pragmatics, components that to date, in our opinion, have been vaguely incorporated into the assessment of language proficiency. In addition, we have assessed the consequences of the use of developmental criteria for the creation of proficiency scales and have warned readers about the potential liabilities in blindly following the criteria from models that may disregard the most recent findings from SLA research. Finally, we have outlined some practical recommendations for the actual design and implementation of classroom-based language tests.

Notes
1. We should note that although the study carried out by Dandonoli and Henning (1990) was claimed to be the first to quantitatively validate the OPI as a testing instrument, it appears to contain various significant methodological weaknesses pointed out by Fulcher (1996).
2. Even though Koike, Pearson, and Witten (2003) report that only a few learners thought the above Spanish question expressed a rebuke, their analysis was based on results from a listening comprehension task.

References


