# CONTENTS

chapter one

# ALTERNATIVES IN SECOND LANGUAGE ASSESSMENT

The primary goal of this project is to provide guidelines for designing performance assessments for second or foreign language classes at all levels of university instruction. Through our efforts towards this goal, we hope to achieve three fundamental objectives: (a) to provide means whereby student performance on real-world language tasks can be validly assessed in terms of real-world criteria, (b) to elucidate the potential for using task-based performance assessment to generalize about students' L2 abilities, and (c) to facilitate a direct link between classroom L2 instruction and real-world language use. Given these objectives and the consequent scope of this project (described in chapter 7), we have decided to report on our progress and findings in a series of publications. The following chapters have been included in this first volume:

Chapter 1 —Alternatives in Second Language Assessment

Chapter 2 — Performance Assessments

Chapter 3 — Task-Based Language Teaching

Chapter 4 — Task-Based Language Assessment

Chapter 5 — Test and Item Specifications

Chapter 6 — Item Prompts

Chapter 7 — Conclusions

Appendix — Example Items and Item Generation

In this first chapter, we will begin by exploring the general topic of so-called "alternative assessments" and how they contrast with our notion of *alternatives in assessment*. In chapter 2, we will turn to the issues involved in performance assessment as it is covered in general education circles as well as in the language testing literature. In chapter 3, we will explore the literature on task-based teaching and demonstrate its relevance for the current project. In chapter 4, we discuss the issues in the literature that are directly related to task-based language assessment. In chapter 5, we will provide test specifications that describe a number of variables which contribute to task difficulty and apply that knowledge to the process of grading the difficulty of prototypical performance assessment items in item specifications. In chapter 6, we will provide detailed descriptions of prototype item

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

prompts generated in this project. In chapter 7, we will summarize our project, discuss our immediate plans for research, and provide suggestions for other future research. In the appendix, we will demonstrate the generative process involved in creating test and item specifications.

## "ALTERNATIVE ASSESSMENTS"

A variety of so-called "alternative assessment" procedures have become popular in recent years: performance assessments, portfolios, student-teacher conferences, diaries, self-assessments, peer-assessments, and so forth. But what are "alternative assessments," and how are they different from more traditional assessment procedures?

Within the mainstream educational assessment literature, the characteristics of "alternative assessments" seem to differ depending on who is describing them. Aschbacher (1991) lists several common characteristics that all alternative assessment procedures seem to share: (a) problem solving and higher level thinking are required, (b) the required tasks are worthwhile as instructional activities, (c) real-world contexts or simulations are used, (d) focus is given to processes as well as products, and (e) public disclosure of standards and criteria is encouraged.

Herman, Aschbacher, and Winters (1992, p. 6) list six characteristics of alternative assessments, which:

1. Ask students to perform, create, produce, or do something.
2. Tap higher-level thinking and problem-solving skills.
3. Use tasks that represent meaningful instructional activities.
4. Invoke real-world applications.
5. Use human judgments, not machine scoring.
6. Require new instructional and assessment roles for teachers.

Huerta-Macías (1995) points out that one benefit of alternative assessments is that they are non-intrusive in that they merely extend and reflect the day-to-day classroom curriculum. She goes on to suggest that, more importantly, students are therefore evaluated on what they ordinarily do in class every day. Alternative assessment provides information not only on learners' weaknesses, but also on their strengths, as they are manifested in class over time. In addition, appropriately administered alternative assessment should be multiculturally sensitive (thus particularly suited for second or foreign language populations). The alternative assessment procedures listed by Huerta-Macías include checklists, journals, logs, videotapes and audiotapes, self-evaluation, teacher observations, and so forth. In addition, Huerta-Macías (1995) argues that:

> Alternative assessment should borrow terminology from qualitative research. Trustworthiness of a measure consists of its credibility and auditability. Alternative assessments are in and of themselves valid, due to the direct nature of the assessment. Consistency is ensured by the auditability of the procedure (leaving evidence of decision-making processes), by using multiple tasks, by training judges to use clear criteria, and by triangulating any decision-making process with varied sources of data (for example, students, families, and teachers). Alternative assessment consists of valid and reliable procedures that avoid many of the problems inherent in traditional testing including norming, linguistic, and cultural biases. (p. 10)

## ALTERNATIVES IN ASSESSMENT

While we are excited about the possibilities of developing new assessment procedures that provide opportunities for students to demonstrate their abilities to use language for meaningful communication (in ways that are consonant with the particular curriculum in which they are studying), we must take issue with the notion that "alternative assessments" are somehow completely new and different from all other testing that has gone before. We also reject the notion that *any* assessment procedure can be held to be inherently valid.

Alternative assessments, especially in the form of performance assessments, are nothing new. We feel that the use of the phrase "alternative assessments" may be somewhat counterproductive, as it is often taken to mean that these assessment procedures are somehow a completely new way of doing assessment, that they are somehow completely different, and that they are somehow exempt from the requirements of responsible decision making. We would like to view procedures like performance assessments, portfolios, conferences, diaries, self-assessments, peer-assessments, and so forth, not as alternative assessments, but rather as *alternatives in assessment*. Language testers have always done assessment in one form or another in conjunction with language teaching (including procedures like multiple-choice, composition, dictation, cloze, etc.), and the recent "alternative assessment" procedures are just new alternatives in that long tradition.

At the moment, others would seem to hold a different view. For instance, from the language testing literature, Huerta-Macías (1995) talks rather dismissively about "problems inherent in traditional testing including norming, linguistic, and cultural biases" (p. 10). Two problems jump to mind with such a statement: first, *norming* is not in-and-of-itself a bad thing, and second, the idea of *traditional testing* seems to be too narrowly defined.

### Norming

If the purpose of a test is to make norm-referenced decisions (for say proficiency or placement decision making), that test must be normed in one way or another because each student's performance will ultimately be compared with the

performances of other students. How the test is normed is another issue. Norming is only a statistical procedure. How it is done and who is used as the norm population are separate issues. There are many options from among which educators can and should choose, but there is nothing inherently bad about norming itself. Norming is a necessary part of developing norm-referenced tests, and such tests are often necessary for distinguishing among peoples' abilities for purposes of admissions or placement. There may be valid political reasons for arguing against making such norm-referenced admissions and placement decisions, but these are separate political issues rather than inherent characteristics of entire types of tests. In a perfect world, with unlimited resources, such decisions would indeed be unnecessary. In the world we live in, such decisions are, and will continue to be, made on a daily basis. We advocate that they be made in a responsible manner.

## Traditional testing

In 1962, another branch of *traditional* testing called criterion-referenced testing was born with the publication of an article by Glaser and Klaus (1962). This approach to educational testing has nothing to do with norming. Instead, in this line of work, strategies and statistical procedures have been developed for improving the relationship of tests to the curriculum actually being taught, to the objectives involved, and to the learning that students are doing. Criterion-referenced testing is not some peripheral movement that will soon disappear. As mentioned earlier, it began with Glaser and Klaus (1962), but it continued with Glaser's (1963) paper, and year by year, criterion-referenced testing has been gaining acceptance (for instance, see Popham, 1978 & 1981 and Berk, 1980 & 1984 for much more on the history of criterion-referenced testing; or see almost any recent issue of *Journal of Educational Measurement* or *Applied Psychological Measurement*).

Criterion-referenced testing has also become increasingly important in language testing circles. The concept of criterion-referenced testing first appeared in the language testing literature as far back as 1968 with an article by Cartier in the *TESOL Quarterly*. While criterion-referenced testing did not reappear until the 1980s, a considerable number of articles have appeared since then (e.g., Cziko, 1982, 1983; J. D. Brown, 1984; Hudson & Lynch, 1984; Henning, 1987; J. D. Brown, 1988; Bachman, 1989; J. D. Brown, 1989 a & b; Hudson, 1989 a & b; Bachman, 1990; J. D. Brown, 1990 a & b; Cook, 1990; Hudson 1991, J. D. Brown, 1992, 1993; Griffee, 1995; and J. D. Brown, 1995 a & b, 1996).

Given its decades-long history and general acceptance, criterion-referenced testing can now be considered rather *traditional*, and this branch of testing answers many of the problems that traditional norm-referenced multiple-choice tests did indeed create in language testing. In addition, criterion-referenced testing procedures, statistics, and theory in general are consonant with the various alternative types of assessment advocated by Aschbacher (1991), Herman, Aschbacher, and Winters (1992), Huerta-Macías (1995), and many others.

Indeed, the authors of this report are comfortable with the ideas and procedures of criterion-referenced testing and find nothing particularly new in the types of assessments advocated by those on the "alternative assessments" bandwagon. *Traditional* testing methods in language testing diverged from multiple-choice years ago with explorations of integrative tests (cloze, dictation, etc.) and performance tests (compositions, interviews, etc.). Research on integrative and performance tests dates back to the 1970s, so they too can be considered quite traditional. How is it that such *traditional* measures are therefore inherently problematic?

We also feel compelled to take issue with the notion that "alternative assessments" are somehow "in and of themselves valid, due to the direct nature of the assessment" (Huerta-Macías, 1995, p. 10). Such an attitude ignores the fact that, like all other forms of assessment, the so-called alternative assessments are used to make decisions (sometimes very high-stakes in nature) about students. Hence, as with all other forms of assessment, the designers and users of alternative assessment procedures must make every effort to structure such decisions so that they are shown to be reliable and valid. Indeed, we feel that designers of alternatives in assessment would do well to attend to and meet the guidelines for reliability and validity set forth in the *Standards for Educational and Psychological Testing* (APA, 1985, 1986).

Precedents exist for demonstrating the reliability and validity of such procedures in the performance assessment branch of the educational testing literature, and we want to adapt those reliability and validity procedures to the purposes of developing sound alternatives in language assessment. The existing techniques for showing reliability and validity of performance assessments, including certain statistical procedures, a variety of observational procedures, and techniques for triangulation of decision making, are not particularly new, nor are they particularly difficult from a technical point of view. Hence, we feel that it is no longer acceptable to take the view that alternatives in assessment are "in and of themselves valid." Such a stance can only lead to smug self-satisfaction and irresponsible decisions being made about the very language students that we claim to care so much about. The issues of reliability and validity must be dealt with for "alternative assessments" just as they are for any alternative in assessment — in an open, honest, clear, demonstrable, and convincing way.

The point of view represented here, then, is that using demonstrably reliable and valid alternatives in assessment can only expand our capacity to make responsible decisions in language programs (for admitting students to the program through aptitude or proficiency testing or for placing them into levels of language study once they are in the program) and classrooms (for diagnosing strengths and weaknesses, checking progress, or assessing achievement). In order to meet this wide variety of decision-making needs, we maintain that a variety of assessment approaches must be used. While our focus in this report will be on a particular decision-making role for L2 performance assessments, we hope that the guidelines set down here and in subsequent reports for test development and validation will also be more generally applicable to other alternatives in language assessment.