An Analysis of Foreign Language Achievement Test Drafts

Catherine Barrette
Wayne State University

Abstract: *In order to obtain concrete information to improve test writing, this study analyzes 13 achievement test drafts written by graduate teaching assistants, lecturers, and adjunct faculty teaching introductory and intermediate college Spanish. The analysis focuses on factors of test method and content likely to create variance in test performance unrelated to learners' abilities. Bachman and Palmer's (1996) rubric for describing target language use and test tasks functions as an organizational tool to describe and analyze the drafts. Results reveal patterns of inappropriate input and inadequately specified procedures, tasks, scoring criteria, and expected responses. Findings suggest common pitfalls for the novice test writer to avoid in test development, and provide a starting point for teachers of any language who want to improve their own achievement tests.*

Introduction

Foreign language teachers both want and need information about their students' current level of ability, progress, and readiness for subsequent levels of instruction. One common source of such information is an achievement test. Achievement tests are foreign language assessment tools directly tied to a particular curriculum and are used to evaluate student progress toward or mastery of course objectives (Hughes, 2002). As a result, such tests are necessarily context specific, and are influenced by the course materials and syllabus, the instructor's preferred teaching methodology, the student population, local administrative factors, and so forth. Many teachers choose or modify commercially produced tests distributed by their textbook publisher; others create their own. In each of these cases, teachers need adequate knowledge of language testing to be able to provide a reliable and valid instrument to measure their students' level of achievement. In many cases, however, teacher certification programs and university foreign language departments are unable to provide adequate professional development in language testing due to time constraints and competing priorities. Nonetheless, teachers often create their own tests and quizzes with which to make decisions about their students' achievement. In these circumstances, measurement error may arise due to the use of inappropriate test content or method that produces construct-irrelevant variance (i.e., variance unrelated to students' language abilities), potentially reducing the validity and reliability of the test.

The purpose of this study was to determine whether common sources of construct-irrelevant variance related to test content and method can be identified and targeted for improvement during the test writing process. This study analyzed 13 achievement test drafts for introductory and intermediate Spanish written by university graduate teaching assistants, lecturers, and adjunct faculty beginning professional development in test writing. The study identified several common types of problems related to construct-irrelevant variance that may reduce the validity or reliability of a test. These results point to aspects of test writing that can be improved through professional development for teachers. An awareness of the types of content and method

Catherine Barrette (PhD, University of Arizona) is Associate Professor of Spanish at Wayne State University, Detroit, Michigan.

problems found in test drafts written by these teachers-in-training may help other test writers to avoid common pitfalls during the test writing process.

Two research questions guided this investigation: (1) Which aspects of the test items written by instructors would become a source of construct-irrelevant variance if this test were administered? (2) Which sources of construct-irrelevant variance have the greatest impact in terms of either frequency or global effect?

To answer these questions, the present study utilized Bachman and Palmer's (1996) "Task Characteristics" and their model of language test usefulness as tools to describe problems of test content and method identified in the test drafts. The Task Characteristics allow for a structured evaluation of the drafts' content and test method, which in turn serve as the basis for the suggestions regarding professional development in test writing that conclude this study.

This study therefore provides an overview of factors of test method shown to affect language test performance, contextualizes these factors within Bachman and Palmer's model of language testing, gives background information about the testing context, and details the study of the 13 test drafts written by teachers undergoing professional development in test writing. The test writing process and the collection and analysis of data are described as they contribute to the search for patterns of content and method problems in the achievement test drafts. Results indicate several priorities for professional development in test writing for novice test authors.

Factors of Test Content and Method that Affect Test Performance

Language testing theory and research into language testing each contribute to the foundations of the present study. Theories of language testing offer explanations of the role of factors—both measured and hypothesized—that influence learners' performance on tests and their relationship to each other. In turn, empirical research provides evidence of the degree and type of influence different factors have on test performance.

Presently, the most comprehensive model of language testing that can be generalized to all four traditional skills (reading, writing, listening, and speaking) and to any language test content is that of Bachman and Palmer (1996). They integrated factors from empirical and theoretical perspectives in their model of language testing, and frame those factors within the larger consideration of test usefulness. These authors asserted that six qualities contribute to test usefulness: reliability, validity, authenticity, interactiveness, impact, and practicality (p. 38). For the purposes of the present study, the focus was on task authenticity, the "degree of correspondence of the characteristics of a given language test task to the features of a TLU [target language use] task" (p. 23) since its absence can be an important source of construct-irrelevant variance.

Bachman and Palmer's model provided an extensive approach to analyzing test tasks and the degree to which such tasks match target language use behaviors. This analysis is accomplished through a detailed description of the factors that can vary from one task to another. The interpretation of such an analysis provided evidence of test usefulness.

For this study, the set of TLU tasks within the relevant TLU domain was defined as the set of classroom activities in the introductory and intermediate Spanish courses at Wayne State University, Detroit, Michigan. These TLU tasks included a range of classroom activities employing the four skills. Representative classroom activities paralleled those in the students' textbook and workbook: listening to and reading narratives, essays, and dialogues followed by short answer or multiple choice comprehension tasks, contextualized grammar practice, guided and open conversations, and the writing of informal letters or compositions. Given this context-specific definition of TLU tasks, for the test tasks in this study to have authenticity, they had to parallel classroom tasks.

Bachman and Palmer argued that for a test task to be authentic, examinees' performance must relate to the way language is used in the TLU domain to be measured. Therefore, they presented a framework for characterizing TLU and test tasks in order to identify parallels and disparities between them. Test tasks had to correspond in key ways as influenced by test purpose, context, etc., to TLU tasks. In order to adequately describe test tasks and make a determination about their correlation to TLU tasks, Bachman and Palmer elaborated their Task Characteristics[1]—an extensive outline for delimiting the test setting, test rubric, input, expected response, and relationship between input and response of a task. The subdivisions of these five categories with relevant data from this study are shown in Table 1 (see Bachman and Palmer, p. 49–50, for the complete list of subdivisions).

Many of the elements of the Task Characteristics have been empirically studied and shown to have an impact upon test performance (see Table 2). A representative sample of recent investigations provides empirical evidence to support the inclusion of a number of the Task Characteristics in Bachman and Palmer's model. For example, Tarone (1998) presented evidence that situational context plays a role in interlanguage variation, an issue related to Bachman and Palmer's "Characteristics of the setting" and "Characteristics of the input." Kobayashi's (2002) study indicating a significant effect of text organization on reading comprehension performance supported elements of the "Characteristics of the input." Differing applications of scoring criteria on the part of raters was shown by Lumley and McNamara (1995) to reflect significant variance in test scores. North's (2000) discussion of scale development further demonstrated the importance of constructing a clear and appropriate scale for scoring. Results such

Table 1

PROBLEMS NOTED BY TEST REVIEWERS AND TEST AUTHORS ON 13 ACHIEVEMENT TEST DRAFTS

Task Characteristics [a]	**Number of problems noted by test reviewers and authors / Total problems**
Characteristics of the test rubrics	
Instructions	
Language (native, target)	1/4
Specification of procedures and tasks	4/12
Structure	
Sequence of parts/tasks	0/1
Relative importance of parts/tasks	3/12
Time allotment	1/4
Scoring method	
Criteria for correctness	4/23
Procedures for scoring the response	0/2
Explicitness of criteria and procedures	1/4
Characteristics of the input	
Format	
Length	4/15
Degree of speededness	1/5
Language of input	
Language characteristics	
1. Organizational characteristics	
a. Grammatical (vocabulary, syntax, phonology, graphology)	3/6
b. Textual (cohesion, rhetorical/conversational organization)	2/10
2. Pragmatic characteristics: Sociolinguistic (dialect/variety, register, naturalness, cultural references and figurative language)	2/2
Characteristics of the expected response	
Format	
Length	1/3
Language of expected response	
Language characteristics	
1. Organizational characteristics: Grammatical (vocabulary, syntax, phonology, graphology)	0/4
Other[b]	2/13

[a]Source: Bachman and Palmer's "Task Characteristics" (1996, p. 49-50).
[b]"Other" category is not part of Bachman and Palmer's original model, but was added for the purposes of the present investigation.

Table 2

EMPIRICAL RESEARCH IN SUPPORT OF BACHMAN AND PALMER'S TASK CHARACTERISTICS

Task Characteristics	
Characteristics of the setting	Tarone (1998)
Characteristics of the test rubrics	Brown, Yamashiro, and Ogane (2001) Lumley and McNamara (1995) Mehnert (1998) North (2000) Skehan and Foster (1999) Wigglesworth (1998)
Characteristics of the input	Kobayashi (2002) Tarone (1998)
Characteristics of the expected response	Lumley and McNamara (1995) North (2000)
Relationship between input and response	Bygate (2001) Skehan (2001)

as these are reflected most directly in the subcategory "Scoring method" of "Characteristics of the test rubrics" and also indirectly in "Instructions" and the "Characteristics of the expected response." Relating to Bachman and Palmer's "Test rubric," Skehan and Foster (1999) reported that fluency is significantly affected by the presence or absence of structure in an oral task. Brown, Yamashiro, and Ogane (2001) revealed an effect of cloze test format on reliability. Relevant to "Time allotment," Mehnert (1998) reported that accuracy of speech improves with one minute of planning time, although Wigglesworth (1995) found no effect of planning time. Bygate (2001) and Skehan (2001) noted an effect on fluency and accuracy for monologue versus dialogue formats, a result relevant to the category "Relationship between input and response."

The small subset of research cited above provides evidence of the importance of the characteristics included in Bachman and Palmer's model and illustrates the appropriateness of this model for research projects similar to the present study. Indeed, Bachman and Palmer suggested using their Task Characteristics framework and modifying it to complement individual contexts for:

(1) describing TLU tasks as a basis for designing language test tasks;

(2) describing different test tasks in order to insure their comparability, and as a means for assessing reliability; and

(3) comparing the characteristics of TLU and test tasks to assess authenticity. (p. 47)

The authors further stated that:

> the characteristics of the tasks used are always likely to affect test scores to some degree . . . since we cannot totally eliminate the effect of task characteristics, we must learn to understand them and to control them so as to insure that the tests we use will have the qualities we desire and are appropriate for the uses for which they are intended. (p. 46)

One key step in gaining the control and understanding mentioned by Bachman and Palmer lies in identifying the areas in which greater control is required. Therefore, in the present study, the Task Characteristics were used for the second and third purposes cited above: to analyze tasks in order to evaluate the comparability between the TLU tasks practiced in the classroom and the tasks found in 13 test drafts scrutinized herein. Disparities between the TLU tasks and test tasks pointed to areas requiring greater control in the test development process to avoid potential construct-irrelevant variance due to limited authenticity or low reliability. The Task Characteristics framework provided an organizational tool for identifying areas of test method and content that lack comparability with classroom tasks. Patterns of disparities for particular characteristics are indicators of factors that could affect test performance due to test content or method rather than learners' knowledge. Therefore, those patterns suggest areas that require attention during test writing, and thus serve as the foundation for recommendations for prioritizing professional development in test writing.

The remainder of this article describes the test writing context and process, reports the analysis of the achievement test drafts written by the introductory and intermediate Spanish teachers, and discusses some implications of the results.

Test Writing Context and Professional Development

In fall 1996, Wayne State University offered approximately 45 sections of introductory and intermediate Spanish (101, 102, and 201). All sections of each level shared a common syllabus, materials, and departmental tests, but each teacher created his or her own quizzes. Prior to fall 1996, the departmental tests were created and distributed by individual course supervisors. Upon my appointment as Language Program Director and Spanish Applied Linguist in fall 1996, the teachers for each level began to collabora-

tively create, evaluate, revise, and edit the departmental tests under my supervision and in conjunction with ongoing professional development in test writing. The goals of this policy change were twofold: (1) to provide teachers with professional development in test writing that they could apply to quizzes and other assessment tools, and (2) to improve the quality of the achievement tests used.

During the semester of the study reported herein, teachers for each of the three levels of Spanish wrote, revised, and administered four to five common tests during the 15-week term. These departmental achievement tests served both formative and summative goals. The tests provided students with feedback about their strengths and weaknesses, and gave teachers information regarding students' content mastery throughout the semester. While such use of tests was pedagogically sound, it created an ongoing dilemma: Postadministration test security was nonexistent because tests were returned to the students with feedback and grades. As a result, original instruments had to be developed for subsequent semesters so that the tests for new and repeating students would not be compromised. In order to keep pace with this constant need for new tests, all instructors contributed to test development either through item writing and revision or item and test evaluation. The majority of these teachers had received no formal training in language testing prior to fall 1996, and therefore began professional development and supervision in test writing at that time.

As mentioned previously, one purpose of educating these teachers about test writing was to help them create tests and quizzes that more accurately measured the language behaviors about which the instructors wanted information. The focus of their professional development was necessarily on the development of the best test items possible through a drafting and revision process since post-hoc analyses of test items was not a useful approach in this context where tests were not reused. Based upon this focus, efforts at professional development in test writing consisted of three main activities:

(1) a four-hour workshop that provided definitions of reliability, validity, and some of the factors that contribute to reliability and validity, and hands-on practice evaluating and revising existing tests with attention to the curriculum and methodology for the introductory and intermediate Spanish courses;

(2) individual consultations between the test authors and myself to discuss specific test content and methods; and

(3) group meetings by instructional level to evaluate test drafts and suggest revisions.

As the primary focus of these professional development activities, teachers collaboratively developed the tests for their level. Each test included a section comprised of multiple items on reading and listening comprehension, writing, conversation, and grammar, which is tested both discretely and integratively in the skill sections (Davidson & Lynch, 2002). In the lowest level course (Spanish 101), students completed all five test sections within two class hours. The other two courses (Spanish 102 and 201) tested listening, reading, and grammar together in one class hour, with a separate process used for evaluating writing and conversation skills. The writing and conversation sections for these two courses were excluded from this study due to distinct administration processes.

Design of the Study

Two main activities were carried out during the fall 1996 semester as part of this study: (1) test writing (in conjunction with the professional development[2] and supervision already described), and (2) data collection and analysis.

Test Writing: Participants and Process

Each of the instructors in the introductory and intermediate Spanish courses authored between two and four test sections (subtests) for their instructional level over the course of one semester to constitute the tests given at that level. Twenty-four teachers were test authors in the fall 1996 semester and participated in the supervised writing of common tests for the first time, working in groups of five to create each exam.

The educational background of these teachers ranged from a bachelor's degree through doctorate. Ten instructors were working towards a master's degree, twelve had already completed it, and two had earned doctoral degrees. Just over one third of the instructors were male, but the group was evenly divided between native and nonnative Spanish speakers. One of the instructors was teaching for the first time; 5 others had between 2 and 3 years of previous language teaching experience, and the remaining 18 instructors had between 3 and 10 years experience in a language classroom. Their ages ranged from 23 through early 50s. Four of the instructors were lecturers (full-time, non-tenure-track), 6 were graduate teaching assistants, and 14 were adjunct faculty hired on a single-semester basis. Although all of these instructors had previously written tests, only the three instructors who were completing the final stages of their doctoral degrees had ever received any formal education in test writing prior to the workshop offered during this study. Their prior exposure was limited to methods of scoring writing samples, however, and thus bears only peripherally on the data in this study.

Based on the textbook and syllabus for each course, the authors of each test met to decide collectively on the distribution of the vocabulary and grammar content among the subtests. Subsequently each author was to create a subtest valued at approximately 20 points, requiring a maximum of 10 minutes for students to complete, and modeled on the types of activities found in the textbook

and workbook. Using this information and the textbook as references, participants then individually developed their test items. Afterwards, the group of authors for each test, the other instructors for the same level, and I met to discuss and revise all drafted sections in a staff meeting. I contributed comments and guided discussions as a way to provide professional development. Typical comments included concerns about vocabulary, clarity of questions or instructions, and scoring criteria. Once the original authors made the changes recommended by the test reviewers to their sections, I reviewed the revisions and compiled them, and then instructors administered the revised test to students.

Data Collection and Analysis

At each of the staff meetings to discuss the test drafts, I acted as a participating observer, discussing and noting comments from all participants as a record of the test method and content problems found on the drafts. Instructors who were unable to attend the staff meetings submitted their written comments on the test drafts.

To provide a more complete description of the drafts, I reviewed each test again in the following semester. In this way I attempted to reduce any halo effect in the comments elicited during the staff meetings that was potentially caused by the teachers' state of knowledge at the time, my own goals with respect to teachers' professional development, and priming resulting from earlier comments on similar problems.

In creating a principled organization of these comments, I utilized Bachman and Palmer's (1996) Task Characteristics. In order to search for a pattern of method and content problems found in the test drafts, I sorted all comments on the test drafts from the instructors and test authors into the divisions of the Task Characteristics (p. 49–50). (See Table 1 for data and see Appendix for a sample test draft section, comments, and their categorization.) In addition, I organized my own comments from the meetings and my subsequent second review in the same way. Several of the comments did not fall clearly into any of the established categories however, leading to the addition of a combined category related to test content and tasks, initially labeled "Other." With this slight modification to the Task Characteristics, the organization of the comments was complete. Comments which identified any characteristic of an item or section which would potentially increase construct-irrelevant variance due to a mismatch between test task and classroom tasks were highlighted. I then tallied the number of comments from instructors and authors as well as the total number of comments from the combined reviews (the instructors', authors', and my own review) from which to calculate the ratio of problems identified by test authors and reviewers to the total number of problems in each category. This compilation of the teachers' and my own comments served as the data for this analysis (see Table 1).

Results

While each of the 13 achievement test drafts under scrutiny contained many well-designed elements, mismatches between classroom and test content and tasks appeared in every one, each mismatch a source of construct-irrelevant variance. Table 1 contains a tally of these problems, and provides a response to the first research question: Which aspects of the test items written by instructors would become a source of construct-irrelevant variance if this test were administered? The number of negative comments reflects the problems identified by instructors for Bachman and Palmer's Task Characteristics and the appended Other category. Fifteen of the 36 categories over which test writers had control contain negative comments. Due to this large number of categories, each cannot be discussed individually. However, several are given detailed attention below in addressing the second research question.

The second research question: Which sources of construct-irrelevant variance have the greatest impact in terms of either frequency or global effect? is discussed extensively in this section. The most frequent problems on these 13 achievement test drafts dealt with (1) the clear and unambiguous specification of procedures and tasks, (2) the relative importance of each section and item, (3) the explicitness of criteria for correctness, and (4) the length of the input. A less frequent but significant problem presents itself in relation to (5) the language of the expected response. Examples of the problems corresponding to each area are discussed in detail in the remainder of this section, and a brief description of the category added to Bachman and Palmer's Task Characteristics ([6] Other) concludes this section.

Specification of Procedures and Tasks

Problems arose on 7 of the 13 drafts with regard to the Task Characteristic category: "Specification of procedures and tasks." Spread among those tests were 12 sections with ambiguous tasks for either the examiner, the examinee, or both; instructors identified only 4 of the 12. The listening comprehension section of test 102-3 (the third test for Spanish 102) exemplified this problem. The instructions for the task read, "Answer the following questions with a complete sentence," yet there were no questions on the page and no indication that this activity was part of the listening comprehension task. The unexplained intent was for the examiner to read the questions aloud; the examinees were to then answer the questions in writing. Most problems in this subcategory were not as severe as this example, but rather omitted information that might seem obvious to the test author although not to examinees. Frequently the listening and reading comprehension subtests' instructions failed to require the examinee to respond based on information from the text, for example. This omission became an issue of importance for some inferential questions that could be perceived by students as opinion-based, and oth-

ers that could be answered correctly using information from personal experience contrary to that contained in the text. Instructors' procedural guidelines for presenting the listening comprehension text to the examinees were regularly absent, raising comments regarding the number of repetitions and speed at which the text should be read by the instructors.

Relative Importance of Parts

In the instructional goals of all of the introductory and intermediate Spanish courses, equal weight was assigned to developing language proficiency through the four traditional skills and grammatical accuracy. As such, each part of the test (representing the skill areas and grammar) should have been in close balance with the other parts. The instructions provided to the test authors included guiding information to encourage this balance. Nonetheless, 11 of the 13 tests revealed the test developers' lack of awareness of this relationship and of the effects of dismissing it since in all 11, at least one section reflected a disparate distribution of points. Instructors noticed this problem in only 3 of the 12 occasions in which a disparity was present. Some tests also contained sections requiring an inordinate amount of time to complete. For example, in test 101-4, 78 points were allocated to grammar, 30 to reading comprehension, and 15 or less to each of the three other sections. In another example, test 102-2 contained an exceptionally long reading comprehension section, giving this section extra weight in terms of time required for completion, although not in points.

Criteria for Correctness

The most common omissions made by the group of test developers in this study pertained to this topic. Examples of inadequate criteria for correctness were abundant, yet instructors recognized this as a problem in only 4 of 23 cases. Only test 201-4 provided instructors with complete information on the criteria for correctness for all subtests. The other 12 tests contained tasks lacking accurate or complete information concerning minimal requirements for full credit, availability of partial credit, or length, type, or language of the expected response. Many of the reading comprehension subtests, for example, informed examinees that their answers would be scored for grammatical accuracy, with no notice of credit to be given for the content of their answers, the true goal of that section. In reality, credit was only given for appropriate content in the response since examinees could not accrue points for good grammar, yet students could be penalized for poor grammar. Criteria for scoring writing sections were also vague; while instructors possessed a general format for grading compositions, instructions for the application of that information to particular tests was consistently absent from the test drafts.

Length of the Input

In terms of the input on each test, length can only be used as a comparison in the listening and reading comprehension activities because of the changing nature of the tasks in other subtests. While acknowledging that many task and text factors contribute to comprehension processes and cannot be ignored, a look at the length of the texts chosen for these tests can nonetheless be revealing. As Bachman stated, "While length in itself may not be a critical facet affecting performance, the longer the language sample, the greater the potential effects of the other characteristics . . ." (1990, p. 130).

As language students progress, one might expect them to be capable of reading longer and more difficult texts of a single genre in the same amount of time by utilizing their broadening linguistic repertoire. The reading and listening passages present in the Spanish introductory and intermediate courses' instructional materials reflected this expectation, as should have the tests. Even a cursory scan of Tables 3 and 4, however, reveals the inconsistencies among the comprehension tasks intended for students on these test drafts. In the 101 and 102 tests, the listening and reading passages varied almost randomly in length from long to short and back again. The lowest instructional level (101) did not have the shortest texts in most cases, and at times even had the longest ones. In no case was there a consistent progression

Table 3

LENGTH OF LISTENING PASSAGES (IN NUMBER OF WORDS)

Instructional Level	Test 1	Test 2	Test 3	Test 4	Test 5
101	340	224	80	300	—
102	165	60	180	465	—
201	130	128	160	135	400

Table 4

LENGTH OF READING PASSAGES (IN NUMBER OF WORDS)

Instructional Level	Test 1	Test 2	Test 3	Test 4	Test 5
101	340	224	225	410	—
102	240	480	370	275	—
201	190	238	368	364	720

in the drafts over time into longer texts, although the 201 reading passages approached this pattern. Fifteen comments criticized test sections based on length, making this the second most common problem in these test drafts; teachers noted this problem in four comments.

Language of the Expected Response: Language Characteristics

Construct underrepresentation is at the root of many of the problems noted with respect to the language elicited on the tests.[3] Messick defined this concept as a test which "is too narrow and fails to include important dimensions or facets of focal constructs" (1996, p. 244). This negative attribute was common to all levels as evidenced by the tendency to create tasks eliciting only the most common vocabulary items and the most regularized grammatical forms. While there were only four comments made on this topic (and none noted by the teachers), each occasion reflected a very global issue. Therefore, in order to set professional development priorities, the impact of each category in the Task Characteristics must be assessed for its range in addition to the frequency of problems.

For example, in the grammar section of test 101-4, the goal was to measure knowledge of the contextualized use of reflexive verbs. Nonetheless, all 13 items elicited only third person forms of regular verbs, and 11 of the items used Class I verbs, leaving only 1 item each for the other two regular verb classes. Also, in probing examinees' knowledge of other verb conjugations (among other goals), the writing section of the same test only elicited first person singular forms. In another example, the majority of the 201 grammar sections permitted meaningless answers and repetitive responses if they were grammatically accurate, where the intent was to test knowledge of specific vocabulary domains and grammatical forms. Such omissions of content and a focus on form over meaning were common despite test authors' possession of a list of concepts to include in their section.

Another source of error under this topic was the construction of tasks which required examinees to produce unknown grammar in their responses. In test 101-2, for instance, the writing task instructed examinees to describe plans for an upcoming weekend in order to convince a busy friend to accompany them on a trip. The instructions appeared in English, thus avoiding the use of any unknown Spanish forms. The section author's intent was to elicit leisure activity vocabulary and present tense verbs in an informal context. The vocabulary goals could be met by examinees in this case, but not the grammar goals since no means of expressing future events appeared in the curriculum prior to this test, nor would the topic's constraints sample students' present tense usage. Test 102-2 demonstrated a more subtle aspect of this sampling problem. The focus of the grammar section was the distinction between two past tenses, the major grammatical concept taught prior to this test. The remainder of the test nonetheless utilized the present tense for the reading and listening passages, and consequently for the responses associated with the comprehension activities. The opportunity for greater breadth and depth of sampling of the past tenses was available, but not taken.

Other

This ad hoc category holds comments that either did not clearly pertain to other elements of the Task Characteristics paradigm, or else combined elements of more than one category. Comments placed into this category reflected the presence of ambiguous or "trick" questions, excessively and inadequately demanding tasks, and tangential requirements. Each type related to Messick's "threat to validity known as construct-irrelevant variance (which jeopardizes directness): the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct" (1996, p. 244). Twelve of the 13 tests in this study contained examples of these problems, yet the effects of each were idiosyncratic and limited to a local impact. Only two of the problems were noted by instructors, however, as indicated in the sections below.

Ambiguous or trick questions were most common in multiple choice and true–false activities. One multiple choice section of test 102-2 required examinees to select the "logical" phrase to complete the stem. If examinees understand "logical" to mean "grammatical," one of the options would be correct; if, instead, "logical" were under-

stood to mean "most likely in the real world," another answer would be the better choice. There was no option that would blend these two interpretations, however. In a second example from test 101-4, the listening comprehension text described an exam day of a student, "Claudia." The text at various points described her moods throughout that day. One of the true–false items required examinees to judge the accuracy of the statement "Claudia is afraid." While this was true at one point in the text, by the end of the text she was no longer afraid, thus creating difficulty in determining the correct response.

A second common criticism falling into this category arose from task difficulty. Some tasks in and of themselves were overly demanding, such as the listening section of test 101-2. The assigned purpose of this section was to measure comprehension of time expressions. The task required beginning language students to listen for and list *in order* a series of ten times embedded in a text replete with colloquialisms and a high concentration of unfamiliar vocabulary and grammar. Not only was the language in this text difficult, but the times followed quickly after each other, were unusual (e.g., 11:59 rather than 12:00), and as a result were more difficult to process, a problem one instructor noted in her comments. Examinees' failure to capture any single time created a domino effect, moving all subsequent times out of position in the list and confounding the scoring of this task. The limited processing time for examinees between items added a further complication. At the opposite extreme of task difficulty, other authors provided models of the expected responses which included the target vocabulary or grammatical structure that was to be elicited. Test 102-3, for instance, contained a grammar section meant to elicit the cultural distinction between the use of formal and informal commands, yet both the model and each item provided a note stating the formality required in the response. As a consequence, the difficulty of the task was minimal; rather than testing both sociolinguistic and linguistic competence, the items merely required a correct verb conjugation.

Finally, tangential requirements included those unrelated to the purpose of the test. Throughout many of the 102 test sections, authors required examinees to answer with complete sentences regardless of the goal of the task or the sociolinguistic appropriateness of this demand. As noted previously, in some cases the instructions to students implied that responses containing complete, grammatical sentences would receive credit regardless of content or relevance to the task. Compliance with these instructions did not demonstrate command of the skill purportedly tested. Equally problematic was the fact that incomplete sentences that were contextually appropriate resulted in lower scores. An interpretation of such a lowered score as reflective of the construct itself would be misleading and invalid. In another example, two comprehension sections (101-2 reading, 201-1 listening) incorporated items requiring mathematical calculations to generate correct responses. Incorrect answers did not necessarily indicate a lack of comprehension, but could be the result of an error in calculation, a separate skill. One instructor commented on this as a potential problem.

In summary, five key areas of test method and content accounted for 55.5% of the problems on these 13 test drafts: (1) specification of procedures and tasks (10%); (2) relative importance of parts (10%); (3) criteria for correctness (19.5%); (4) length of input (13%); (5) language characteristics (of the expected response) (3%). Only in some areas did the drafts improve as the semester of the study progressed despite the professional development teachers underwent. Issues related to the relative importance of parts (e.g., length of input, assignment of points per section) failed to improve over time, but tasks and procedures became clearer and more explicit as teachers gained experience. Regarding criteria for correctness, test authors provided greater detail as the semester continued, but the criteria themselves did not necessarily better match the goals of the test. As noted earlier, the length of the input varied randomly across test drafts, suggesting that teachers did not perceive the importance of articulation from one test to the next, but rather saw each test as an independent exercise. Lastly, the characteristics of the expected response improved over time, with fewer attempts to elicit aspects of the language which students had not yet studied.

Discussion

Language Proficiency and Knowledge of Testing: Separate Competencies

The comments gathered from the 13 test drafts in this study revealed several strengths and weaknesses. Among the strengths were instructors' language proficiency and expertise in the content that they teach. Their implementation of testing issues, however, appeared to be an independent element that must be developed.

Evidence of these instructors' strengths lay in part in the comments that were not made with frequency. Few of the tests, for example, presented ungrammatical input (at times due most likely to typographical errors), although arguments over dialectal differences in grammaticality were not uncommon. For subtests that were contextualized, sociolinguistic and rhetorical constraints were rarely violated and in some cases were tested, such as in the elicitation of appropriate formality and turn taking in the 101 conversation tasks. Of related importance is the fact that the contexts which were provided were realistic and within the world knowledge of most examinees, and parallel to the classroom activities that students had practiced, lending a degree of authenticity to the tasks in these contexts.

Test Writing Recommendations

While results indicated that the language proficiency of these instructors was high, they also brought to the foreground the importance of familiarity with testing issues.

The most frequently noted problems on these 13 tests dealt with the importance of each subtest and item, the clear and unambiguous specification of procedures and tasks, the explicitness of criteria for correctness, and the length of the input. The language of the expected response was an additional source of concern of a more global, but less frequent nature. All of these categories represented issues related to test method or content. The nature of most comments suggested only a cursory understanding of the link between teaching and testing, between a task and the information gathered from it, and between consistency of testing procedures and useful outcomes.

In keeping with these patterns of problems, a suggested approach to test writing for novice test writers is to focus on reviewing a test draft from two perspectives: the students and the scorers. By putting instructors into each of these roles as they read a test draft, they may be more likely to move beyond a superficial proofreading and into a true analysis of test tasks that locates the most common problems seen in this study.

Results of this study suggested that taking the test drafts as would a student and discussing students' possible (mis)interpretations of the tasks and procedures were effective strategies for helping the teachers in this study to improve the clarity of tasks and procedures These strategies also enabled the teachers to revise tasks in order to elicit appropriate output from examinees. Each discussion of the tasks and procedures allowed the participating teachers to understand some areas of ambiguity that could cause confusion among students or among the teachers themselves had they administered each test draft. These types of problems were quite easy for teachers to recognize once they began to think about taking tests from a student's perspective and used their creativity to predict behaviors and output that their pupils might produce if presented with the draft unrevised. Little time was needed to guide teachers to recognize problems in these categories, and the impact was substantial.

Less effective were the efforts to better the criteria for correctness on the test drafts. Although test writers' instructions for completing and scoring each section became more explicit after discussions of areas of ambiguity, their quality did not necessarily progress. Instructors seemed to find it difficult to rethink the relationship between a task's criteria for correctness and the information they hoped to gather about their students. This problem appeared to stem from two sources: the newness of a communicative approach for these teachers at the time of this study, and the "testing how you were tested" cliché. Instructors intuitively graded tasks with a set of internal criteria that may or may not have been representative of their teaching approach or the criteria set forth in the task. Although teachers in this study were directed to focus their attention more on communicating content than on grammatical accuracy, their testing approach generally tended toward more traditional scoring. In such tasks, linguistic features and accuracy were the primary or sole source of points for students, despite the fact that teachers were encouraged to model their test tasks on communicative activities in the textbook and workbook which placed a higher value on the expression of meaning. Therefore, greater attention should be paid to helping teachers understand that the information they would get from test items that deviate substantially from their teaching approach would not give them the information they desired. To achieve this understanding, a greater focus on the information teachers obtained from test tasks would be beneficial. Returning to the example of the test section purportedly eliciting formal versus informal commands, a juxtaposition of the original draft which stated the required formality for each item and an alternate form without the indicator of formality may have helped teachers to better recognize the knowledge required of learners (linguistic alone vs. linguistic and sociolinguistic knowledge). While the test writing workshop included examples of tasks with these types of problems, no juxtaposition of "better" and "worse" examples was provided to help teachers understand the relationship between teaching and testing, a task and the usefulness of the information obtained. Early professional development in test writing must therefore place a higher priority on recognizing the match between teaching and testing method, and between a task and the information gathered from it.

The remaining two predominant areas of problems, relative importance of parts and length of input, also reflected the need to spend more time attending to the link between teaching and testing approaches. A better strategy for raising awareness of this link may be the elicitation from teachers of descriptors of their teaching materials and methods rather than providing them with such a characterization in the test-writing guidelines. While this elicitation would certainly require a greater proportion of the available workshop time, it could have greater impact because teachers would be actively involved in delineating the key features of their teaching–testing context. To improve articulation across tests, attention to the progression of materials in particular may be a successful strategy. In contrast to the other areas requiring early professional development, these two areas would be likely to benefit from ongoing attention with a focus on comparing the previous test to the current draft with respect to both relative importance of parts and length of input.

In summary, a two-pronged approach to professional development in test writing seems indicated by the data in this study. Initial efforts at avoiding the many repetitive problems that teachers in this study found easy to recognize should focus on (1) predicting alternate interpretations of tasks, scoring, and expected responses, and (2) identifying key features of the teaching context that should be present in the testing situation. As a second strategy, ongoing dialogues with colleagues or applied linguists focused on ascertaining that the key features of the teach-

ing context actually are reflected in the tests would help reduce the frequency of the more resistant problems encountered in this study.

Conclusions

On a daily basis, decisions about student progress, course content, and curricula are made by parents, teachers, and administrators utilizing the results of tests, quizzes, homework assignments, and other activities in the language class. It is our responsibility as professionals to gather the most accurate and appropriate information possible as the basis for those decisions. To do so, we must empower teachers to create and use better tests.

This study highlighted language instructors' need for professional development in test writing *in addition* to teaching methods, the target language, classroom management and other pertinent aspects of their role as educators. The test developers' task is demanding: Not only must they be knowledgeable about testing, but also about the many features of language and language use that can have an impact on a testing situation. Further investigation into the characteristics of tests developed by instructors with little or no professional development in testing would help to confirm or revise the patterns found here and potentially increase the effectiveness of professional development in test writing.

Increasing teachers' knowledge of test development is an important goal. An understanding of testing issues allows teachers to put their knowledge of testing into practice in their daily activities (quizzes, homework assignments, in-class activities, selecting testing materials) as well as in developing tests of their own. It also has the potential for use as a tool in reflective teaching, providing instructors with a different perspective from which to evaluate the relationships among their instructional goals, curriculum, teaching methods, and philosophies. Greater knowledge of testing issues also empowers educators who must use externally developed tests by providing them with the terminology of testing with which to demand specific improvements in products from test developers. For instructors interested in research in the classroom, learning about testing equips them with skills useful for action research.

Addressing the need to improve test development by identifying key areas for professional development is only a small piece of the solution to a much larger problem, however. Graduate students and other introductory and intermediate language program faculty are often put into the position of undertaking complex, specialized tasks without adequate professional preparation. In some departments and at some universities, graduate students serve as aides (i.e., graduate *assistants*) to experienced faculty or participate in internships before becoming the primary instructor in their own classes. In contrast, in many foreign language departments graduate students are immediately assigned as the primary instructor, and only sometimes with concurrent professional development opportunities. In the long term and as a profession, we need to move language departments in the direction of consistently providing new teachers with a semester of professional development and observation with limited responsibilities before assigning them a class to teach. In the short term, we need to find better ways to support introductory and intermediate course faculty, particularly because it is generally not practical or always necessary to allow lecturers and adjunct faculty a full semester of professional development before becoming a primary instructor. One possibility for better support of these teachers is putting pressure on publishing companies to provide testing programs with introductory and intermediate language textbooks 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. A second approach could be selective collaboration among similar institutions using the same materials to exchange tests and other activities developed onsite and thus share the burden of materials development. Through collaborative efforts and further research into preparing faculty, students and teachers alike will benefit from better testing and test writing.

Notes

1. Bachman and Palmer noted that these characteristics are neither exhaustive nor appropriate for all contexts, but rather suggest users modify them to fit their own purposes and contexts (1996, p. 47).

2. Faculty had previously undergone professional development in communicative language instruction using a four-skills approach with culture embedded in all skills. Moderate attention to grammatical accuracy was encouraged since the introductory and intermediate courses feed the upper level major–minor sequences in which linguistic accuracy is highly valued.

3. The extent of construct underrepresentation in this study is hidden somewhat by the lack of connections among Bachman and Palmer's Task Characteristics. For example, both the content and the length of the expected response contribute to adequate sampling, yet this interaction cannot be addressed by the individual Task Characteristics appropriately.

References

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

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

Brown, J. D., Yamashiro, A. D., & Ogane, E. (2001). The emperor's new cloze: Strategies for revising cloze tests. In T. Hudson, & J. D. Brown (Eds.), *A focus on language test devel-*

opment: Expanding the language proficiency construct across a variety of tests (pp. 143–161). University of Hawaii at Manoa: Second Language Teaching and Curriculum Center.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23–48). Malaysia, PA: Pearson Education.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven: Yale University Press.

Hughes, A. (2002). *Testing for language teachers*. Cambridge: Cambridge University Press.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*(2): 193–220.

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

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*, 83–108.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–56.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang Publishing.

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

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(2): 93–120.

Tarone, E. (1998). Research on interlanguage variation: Implications for language testing. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 71–89). New York: Cambridge University Press.

Wigglesworth, G. (1998). The effect of planning time on second language test discourse. In A. J. Kunnan, (Ed.), *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 91–110). Mahwah, NJ: Lawrence Erlbaum.

Appendix

Sample Comments and Categorization: Test 101-2: Listening Comprehension Subtest

Comments made by test reviewers:	**Classification into Bachman and Palmer's Task Characteristics (1996: 49–50)**
Number of points for the section not specified	Characteristics of the test rubrics: Structure: Relative importance of parts
Unclear expectations for students whether to write out words, use numbers, and which language if in words	Characteristics of the test rubrics: Instructions: Specifications of procedures and tasks Characteristics of the expected response: Language of the expected response
No indication of number of times and pace at which teacher should read the passage	Characteristics of the test rubrics: Instructions: Specifications of procedures and tasks Characteristics of the input: Format: Degree of speededness
No indication of partial credit (e.g., for hours vs. minutes)	Characteristics of the test rubrics: Scoring method
Some times are unusual (e.g., 1:22, 11:59); test should stick to more common ones	Other
Times are too close together sometimes, and students may miss the second in a series—this is too difficult	Other
If student misses one time, then others will be out of order and will therefore be wrong—how should we grade this?	Characteristics of the test rubrics: Scoring method Other
The test is too long for the sixth week of the first semester	Characteristics of the input: Format: Length
Too much unknown vocabulary and grammar—past tense, colloquialisms, vocabulary from much later chapters	Characteristics of the input: Language of input: Language characteristics: Organizational characteristics: Grammatical Characteristics of the input: Language of input: Language of input: Language characteristics: Pragmatic characteristics: Sociolinguistic
	Other
Why would students ever have to do this task? A different context would be better.	Other