The Practicality and Efficiency of Web-Based Placement Testing for College-Level Language Programs

Elizabeth B. Bernhardt
Stanford University

Raymond J. Rivera
Stanford University

Michael L. Kamil
Stanford University

Abstract: *Articulation is one of the principal challenges of all foreign language programs. A key component of the articulation process is an assessment of student language abilities. On college and university campuses this process is usually conducted via a placement test. As developments in proficiency research have progressed, it is clear that programs need information about a student's grammatical command of a language as well as about their integrative use of the language specifically in speaking. This article examines the process of having students test online before their arrival on campus and provides insights into efficiencies brought about by such testing. The data for the article were generated by 679 learners of Spanish and 78 learners of German as well as by their 14 instructors and 2 language program directors.*

Introduction

Byrnes (1990) cautioned that as foreign language curriculum theory becomes more multifaceted, acknowledging the components of language proficiency, the need for more precise articulation from one curriculum into another would become more complex and more critical. At present, within the context of the Standards Movement and radical shifts in high school foreign language curricula characterized by greater inclusiveness, Byrnes's caution has become a prophesy. The front line of articulation, although admittedly by no means the only dimension of that articulation, is placement testing. Indeed, the placement of students into courses in an effective and efficient manner is one of the primary challenges faced by large-scale university foreign language programs. Students demand a timely placement that accommodates their foreign language knowledge and deserve a placement that is challenging without being frustrating. Teachers demand placement mechanisms that enable them to calibrate their instruction to the needs of entering students and deserve ones that facilitate the design of tasks that help students learn in the language.

Elizabeth B. Bernhardt (PhD, University of Minnesota) is Professor of German Studies and Director of the Language Center, Stanford University, Stanford, CA.

Raymond J. Rivera (PhD candidate, Stanford University) is a research assistant at Stanford University, Stanford, CA.

Michael L. Kamil (PhD, University of Wisconsin) is Professor of Psychological Studies in Education, Stanford University, Stanford, CA.

Program administrators demand efficient placement mechanisms in terms of personnel and time and deserve ones that are valid and reliable reflections of foreign language proficiency.

The essentials shared among stakeholders in the language placement process are often difficult to reconcile. Large-scale, college-level language programs frequently have thousands of students who need to be placed into course levels in a very short period of time. While a face-to-face assessment of each student might be desirable, under such conditions, it is tantamount to impossible. Even traditional paper-and-pencil tests scored by machine do not permit a convenient analysis of individual student performance, so critical to understanding learners' language strengths and weaknesses. The bureaucratic practicalities of running large-scale programs within colleges and universities are also important considerations. Foreign language programs are but one dimension of the undergraduate curriculum. During orientation periods, they compete for time and resources with other academic areas that require placement, such as English language composition and mathematics. In other words, the demands of efficiency often take precedence over, and at times even ignore, research findings about the construct of language proficiency. Research indicates that language proficiency is clearly not limited to grammatical ability and reading and writing abilities assessed in paper-and-pencil tests. Oral language assessment must also be included in any learner profile of language ability. Harlow and Caminero (1990) articulated the point: "If we pay lip service to the importance of oral performance, then we must evaluate that oral proficiency in some visible way" (p. 489). Yet assessment of learners takes time, as dedicated instructors need to examine at a fairly deep level of detail what students are and are not able to accomplish with a given foreign language. In fact, Harlow and Caminero found that 57% of language programs cite lack of time and complicated logistics (such as space and scheduling) as typical impediments to elaborate student assessments.

Statement of the Problem

The tensions among various linguistic and bureaucratic forces need to be reconciled in order to meet the needs of the student/teacher/administrator/researcher stakeholders in the language placement and articulation process. Yet efficiency, accuracy, and completeness are difficult goals to meet under real-world conditions. How can time for oral assessment and evaluation be created under circumstances in which a fixed amount of campus time is available for language testing that must include the assessment of grammatical and literacy skills? How can teachers and students have confidence in placements when they have had little if any time to exhibit and/or examine complete linguistic portfolios? If there are ways of gaining time and efficiency in language placement, what are the opportunity costs? For example, do the gains in time and efficiency come at the cost of reliability and validity of the assessment itself?

These conflicting questions are probed in the present study, which gained time for oral language assessment by moving traditional placement tests into a computerized format. More crucial perhaps than the computerized format was the Web-based delivery mechanism used. While computerized testing, most notably, computer adaptive testing, has been validated and provides a useful and specific kind of efficiency (Chalhoub-Deville, 1999; Madsen, 1991), Web-based delivery permits a different form of delivery—one that can take place at any hour and that is not bound by geography. It also enables instructors to have access to student grammatical and interpretive performance well in advance of student arrival in classes. Finally, and perhaps most importantly, it frees up valuable time on campus during orientation periods to conduct critical oral assessments. Such a solution appears to resolve at least some of the dilemmas confronted by teachers and administrators in the language placement cycle. Yet, gaining time and efficiency in instructional settings might be useless if there is a negative impact either on the quality of the assessment or suggested placement. If assessment mechanisms do not meet standard criteria for test reliability and validity; if test delivery mechanisms interfere with student performance; if administrators, teachers, and students fail to perceive any advantage (i.e., react negatively to new assessment mechanisms delivered in alternative ways), then additional time and efficiency are to no purpose. In order to examine this complex of dilemmas, the present study reviews issues regarding placement testing as well as the process of computerizing tests formerly known as paper-and-pencil tests. It then probes whether computerized placement tests can be held to the same standards for validity and reliability as paper-and-pencil placement tests; whether test users report satisfaction with the placement capabilities of Web-based tests; and whether gaining time for oral assessment by using Web-based testing is perceived as a positive and useful development for language teachers.

Literature on Placement Testing

Significant work about the formal properties of tests—either large-scale, high-stakes tests such as the Scholastic Aptitude Test (SAT) or the Test of English as a Foreign Language (TOEFL) or smaller scale, lower stakes tests such as vocabulary quizzes or classroom-based chapter tests—indicates that all tests are to help decision makers make the best judgments they can regarding human performance under a set of constraints. Bachman's award-winning text (1990), as well as Bachman and Palmer (1996), provided the language learning and teaching field with important insights into testing processes and also raised complicated concerns and dilemmas. These important

publications appear at one level to be treatises on measurement and the statistics and theory implied by the concept of measurement. Indeed, they do provide critical language testing-specific information in contrast to the more general testing literature (e.g., Messick, 1989). Yet perhaps more importantly, they are, in the final analysis, reminders about the practicality of tests and why testing is so crucial to helping teaching and learning.

Chapelle (2001) offered an excellent complement to Bachman (1990) and Bachman and Palmer (1996) with a book-length discussion that extended fundamental concepts regarding language testing toward computer applications of language testing in teaching and research. Chapelle reviewed Bachman and Palmer's criteria for "test usefulness" and provided critical insights into the key features of high quality tests: "reliability, construct validity, authenticity, interactiveness, positive impact, and practicality" (p. 101). In her highly informative treatment of these key features, Chapelle worked through the manner in which each feature exhibits itself within the context of the assessment of the various language constructs, reading, listening, and writing. Indeed Chapelle's work, as well as earlier treatments such as Bachman and Palmer's, derive from Messick's (1989) theoretical orientation which emphasized both the evidence for validity and the consequences of validity for tests.

There is no doubt that construct validity is a critical part of test validity, but as Shohamy (1998) noted, "the validity of assessment procedures also depends on their purpose" (p. 252). Shohamy noted that there is another facet of test validity that can be equally important particularly for placement testing—the notion of predictive validity. The predictive validity of a placement test is uppermost in the list of criteria with regard to any given placement test's utility. That is, unless the test can correctly predict where a student should be placed in a sequence of courses (for optimal learning), the test is less than useful. Given this assumption, different procedures for establishing validity become important. Methods that relate the test to learning in the sequence of courses are to be preferred. In short, a placement test must be aligned with the curriculum to the extent that a student will improve both in language proficiency and in the score on the placement test as a result of having taken the optimal sequence of courses.

Language placement tests are usually constructed to yield one score (Anastasi & Urbina, 1997) from which examinees are placed into one of several categories. Well-conceived test specifications, therefore, are critical in assuring that scores can be interpreted according to tenable placement criteria, and that the test is sufficiently comprehensive so as to describe an examinee's capabilities. Alderson (1988) viewed test specifications as one of the most critical validity issues in language proficiency testing. In discussing the role of test specifications, Alderson suggested that test specifications should capture as much of the construct (i.e., language proficiency) as possible while simultaneously providing a template by which multiple forms of the test can be produced. Others suggested that documentation of test specifications provides a critical piece of validity evidence (Ebel & Frisbie, 1991). Practical issues of constructing comprehensive tests according to sound specifications and real-world uses have been treated in detail (Bachman & Palmer, 1996; Harrison, 1986; Hughes, 1986). Placement testing has also been an obvious issue for Spanish language programs in particular (Klee & Rogers, 1989; Wherritt & Cleary, 1990; Wherritt, Cleary, & Druva-Roush, 1990a). With the increasing number of heritage language speakers and the concomitant growth of Spanish programs in secondary schools, the need to place students into appropriate courses has been especially acute (Larson, 1989; Teschner, 1990).

Dyer (1947a, 1947b) found medium to high correlations between written tests of language ability in German and French, and course grades. In comparison to the verbal measures of Williams and Leavitt (1947), Dyer (1947b) echoed their assertion that language test scores were more valid among higher achieving students, suggesting that the French tests were more accurate in measuring the reading ability of students who had completed several quarters of instruction. In somewhat of a contrast to the preceding authors, Goodman, Freed, and McManus (1990) found that scores from the short form of the Modern Language Aptitude Test (MLAT) did not have much predictive capability and had a weak association with first-year language course grades.

Much of the attention in computer-administered placement testing has been focused on areas other than languages. An example of this is found in Day (1999). Although treating general curriculum students, Day found in the placement of postsecondary students that computer-adaptive testing provided satisfactory placement results for students entering remedial or developmental programs in algebra. It is interesting to note that these results were observed among lower achieving students, in contrast to Williams and Leavitt (1947) and Dyer (1947b), who asserted better predictability among higher achieving students.

The testing validation literature urges mindfulness of the purposes for which tests are used and that "assessment is not just about writing tests; it involves a host of factors that affect the learning of languages" (Shohamy, 1998, p. 258). Hence, a placement test even with precise technical qualities (e.g., high reliability, substantial construct validity) that does not appropriately place students into courses will affect instruction negatively by causing confusion, disorder, and frustration. The literature indicates that an effective way to establish the validity of a placement test for language is to do an intervention study (Shepard, 1993). One way to do this is to have students take the placement test

before being placed in an appropriate course, complete the sequence of instruction, and then retake the placement test upon completion. If the test is effective in assisting instructors to make valid decisions, students' scores on the placement test should improve as a result of the intervention of the instruction in the courses. Shohamy (1998) noted:

> Assessment is shaped by its specific context, its purpose, the type of knowledge it addresses, the procedures . . . This multiplicity may seem overwhelming at first, but it opens new avenues for matching assessments to contexts and for making quality choices that are likely to have greater benefits for learners and teachers precisely because assessment and learning are seen not as separate activities but as intimately related to each other. (p. 258)

Research Literature about Computerized Testing

The advantages and disadvantages of computer-based testing and paper-and-pencil testing have been considered extensively (e.g., Boo, 1997; Bunderson, Inouye, & Olsen, 1989; Hambleton, Swaminathan, & Rogers, 1991; Hamilton, Klein, & Lorié, 2000; Kumar, 1996). The discussion of advantages and disadvantages is familiar to users and administrators of computer-based tests. The dramatic increase of educational and psychological measurement instruments capable of being administered by computer keeps fresh the issue of whether scores obtained from one medium are comparable to scores obtained from another. The question is whether there are general differential effects of test scores by delivery medium, and what testing features are most likely to give rise to potential differential effects.

In general, three variables have been identified as potential sources of variance caused by a computerized delivery mechanism: computer-display variables, chronometric variables, and computer experience variables. Early on in computerized testing, limitations of monitor display capabilities and poor legibility were issues of concern (Jonassen, 1986; Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1991; Van de Vijver & Harsveld, 1994; Wildgrube, 1982). In like manner, in the earlier days of computer usage, poor resolution and clarity led to slower rates of reading and processing. However, with the technical developments in screen clarity and resolution, the issues have faded. Similarly, concerns regarding computer user comfort level led some researchers to examine the impact of computers versus paper-and-pencil tests on test-taker affect. Wise, Boettcher, Harvey, and Plake (1987) and Dimock and Cornier (1991) found no evidence to support the finding of any significant relationship between examinee computer anxiety level and test performance. Further examinations of affective concerns involved investigating whether examinees with more computer experience were more likely to be positively disposed towards computer-based testing (Burke, Normand, & Raju, 1987; Levin & Gordon, 1989; Powers & O'Neill, 1992; Ward, Hooper, & Hannafin, 1989). Regardless of experience, researchers reported that examinees have generally positive attitudes toward computerized testing (Bresolin, 1984; Harrel, Honaker, Hetu, & Oberwager, 1987; Vispoel et al., 1997).

Significant evidence has come from the Educational Testing Service in this regard. In an extensive study that compared paper-and-pencil and computer-based TOEFL tests, Taylor, Jamieson, Eignor, and Kirsch (1998) found no evidence indicating that the medium of testing made a difference in test-taker performance. The study examined 1,100 computer users at two different computer-experience levels across 60 TOEFL items. Computer administration did not seem to interfere with the ability or comfort level of subjects with high or low levels of computer experience from completing the TOEFL exam.

The overwhelming majority of evidence regarding the computer administration of tests suggests that scores obtained from computer-based tests are comparable to those obtained from paper-and-pencil tests. There appears to be little or no inherent and systematic differential functioning of one testing medium over the other. This suggests that one delivery medium is not preferable to the other in so far as test validity is concerned. Either medium can be used depending on the testing situation. Any advantages from computer-based testing are not offset by compromised scores or validity owing to media differences. Indeed, any of the differential effects found by researchers since the mid-1980s will be reduced or disappear in the future as examinees become more familiar with computers and have more experience with them in learning and assessment situations.

Language Placement at Stanford University

The Stanford Language Center was established in 1995 with the responsibility for establishing and maintaining language performance standards, encouraging excellence in foreign language teaching, providing professional enhancement activities for the teaching staff, and establishing a research program about language teaching and learning. This charge, of course, entails enforcing the language requirement: "one year of university language study or its equivalent." To do this, the Center directs and coordinates all proficiency testing and collects and analyzes student performance.

Each language program has written and oral dimensions to its placement/exit testing. All written examinations test grammatical features as well as the ability to write connected text and to read authentic prose material. These written examinations are administered online via the Language Center Web site.

To access the placement examinations, students must enter a valid seven-digit Stanford University ID number. Each online exam contains several sections of multiple-choice or fill-in-the-blank questions testing knowledge of grammar and vocabulary. Additional sections request a short biographical paragraph in the language as well as the comprehension of a short passage assessed according to multiple choice or immediate recall. In order to complete the placement procedure, students must also take the oral portion upon arrival at Stanford.

At the bottom of each section of each placement examination, are buttons labeled "BACK" and "PROCEED." When all the questions in a section are answered, the examinee presses PROCEED to go to the next section. Pressing BACK enables the review of a previous section. Testees may use these buttons to go back and forth between sections as often as they wish while taking the exam. Once they are satisfied with all of the answers and have logged out of the exam, they are not granted access again. The recommended time to take the online tests is one hour in a single sitting. Netscape or Microsoft Internet Explorer 4.0 (or a later version) are recommended. Other Web browsers that handle forms work, but problems with formatting may make the tests difficult to read. For the writing sample a chart of keystrokes to produce accent marks on Mac and Windows machines is provided. Advisory information given to the students at log-in time is provided in the Appendix.

Language program directors charged with placement decisions have access to a scoring Web site. Each test taker's performance is available along with his or her demographics. Calculations of items that are automatically scored are listed, along with the answer that the test taker chose for a particular item. In addition, writing performances that must be assessed individually are available with a scoring and comment grid for each student. All data are fed into an MS Excel spreadsheet for large scale reporting and analysis. In fact, the preliminary placement is already made when students arrive for the oral portion of the examination. For this portion, German, Japanese, Chinese, Italian, Spanish, French, and Portuguese instructors use the Simulated Oral Proficiency Interview (SOPI) created by the Center for Applied Linguistics to assess oral proficiency (Kuo & Jiang, 1997; Stansfield & Kenyon, 1992a; 1992b). A SOPI is an audiotaped interview with an accompanying booklet. Students are asked questions and to describe events or pictures in the language, and their responses are recorded. Trained raters then assess each interview. The SOPIs used at Stanford were developed at the university and are delivered during new student orientation periods via cassette tape with an accompanying booklet.

Examining Practicality and Efficiency

Formal analyses of language program developments and modifications are always critical. In order to examine the impact and efficacy of online placement testing, two sets of data were collected. First, a quasi-experimental evaluation of the validity and reliability of the placement instruments delivered via the Web was conducted. Second, cognizant language instructors and program directors were interviewed regarding their perceptions of the accuracy of placement as well as their reaction to modifications in the administrative and instructional processes involved.

Scores were collected from the summer 1999 and 2000 administration of the German and Spanish placement examinations. These scores were generated from freshman and transfer students required to take a language placement test before their fall matriculation. Examinees took their placement exams at their homes, in libraries, schools, or any other location not on the Stanford campus.

The German Placement Test contains two parts: Part I has 30 items and Part II, 39 items. The multiple-choice items in Part I focus on morphology; the items in Part II, on syntax (principally, dependent and independent word order) as well as the past and perfect tenses and voice. A total of 30 students took the 1999 placement exam; 48 took the examination in 2000. The Spanish Placement Test contains Part I (32 multiple-choice items principally focused on present tense forms and vocabulary); Part II (35 fill-in-the-blank items assessing noun/pronoun replacement as well as imperfect and preterite tenses); and Part III (25 multiple-choice items assessing the comprehension of two passages). Two hundred and eighty-six students were placed in 1999 and 393 in 2000. These test scores were queried in raw string format, and then converted to Excel files, where the data were examined for consistency and correctness. Responses were then recoded as 0/1, and imported into SPSS where a number of statistical analyses were performed.

Posttest scores were obtained from Stanford students who had completed three quarters of language instruction. Participants were recruited from among Stanford language students completing their third quarter of language study at the time of testing, and were administered the same tests as are administered to incoming Stanford students planning to be placed within their intended language programs. In the posttestings, 14 German and 41 Spanish language students participated. These volunteer participants were each paid $10 to retake the online placement examination under conditions parallel with summer administrations. That is, as with the pretests, students were able to take the posttests at any location with Internet access, such as on-campus residences, campus libraries, departments, or any other location not on the Stanford campus.

In order to access both the pre- and posttests, Stanford students were required to log in to a secure server environment using their Stanford ID numbers. Both administrations required an electronic signature invoking the Stanford Honor Code before being allowed to view any test items,

Figure 1

SCREEN SHOT OF A WEB-BASED PLACEMENT TEST

University

German Language Placement Test

PART I

(Suggested Time: 30 minutes)
Click the button next to the word or phrase which correctly completes the sentence.
When you have answered all the questions, click the READY button at the bottom of the page.

1. Ihr _____ zuviel.

A. spricht
B. redet
C. sprechen
D. reden

2. Er _____ es sicher nicht.

A. wißt
B. weißt
C. wissen
D. weiß

3. Wo _____?

A. essen wir heute
B. wir essen heute
C. heute wir essen
D. wir heute essen

4. Ich will _____.

A. nach Berlin morgen gehen
B. morgen fahren nach Berlin
C. morgen nach Berlin fahren
D. fahren nach Berlin morgen

5. Ich kenne _____ Fräulein schon lange.

A. diese
B. diesen
C. dies
D. dieses

and students were given the option of completing a brief tutorial familiarizing them with the test interface and navigation. Upon submitting an electronic signature, examinees were then issued the first part of the testing sequence. All items were in multiple-choice or one-word, fill-in-the-blank format, and numbered clearly.

Figure 1 illustrates a screen shot from the German examination.

Test Score Reliability

In order to assess reliability, the scores for the German and Spanish placement tests were tabulated and analyzed to obtain Cronbach's alpha. This is a conservative procedure that is the average of all possible split-half reliabilities for a set of scores. The placement tests for Spanish and German in both 1999 and 2000 were used in this analysis. For German, 78 students participated; for Spanish, the number was 679 (data from heritage Spanish-speakers are not included in the analysis). Descriptive statistics are presented in Table 1.

Table 1 also contains the number of items, the number of subjects, mean scores and variances. Critical are the Cronbach's alpha data. Reliability estimates were calculated for each section of the tests and for the tests as a whole and range from a low of .81 to a high of .94. These numbers indicate that the tests render scores that are reliable and hence capable of providing reliable information to test administrators.

Test Score Validity

Employing Shepard's notions about using experiments to help to establish validity, students studying in the language were solicited to take the placement test near the end of the academic year 2000–2001. Students were offered a modest stipend for participation. There were 14 students who participated in the German portion of the study and 41 who participated in the Spanish portion.

Scores on the original placement tests were compared to the scores on the second placement test. This follows from the assumption that students should do significantly better on the second administration of the placement test, if the original decisions were appropriate. Students placed too high or low would not learn very much, either because of being exposed to materials that were too difficult or because they were exposed to what they already knew.

The analysis was relatively simple. Means and variances were observed and effect sizes calculated. These data are displayed in Table 2.

There were significant differences between the first and second administrations for all students who took the origi-

Table 1

DESCRIPTIVE STATISTICS OF PLACEMENT TEST SCORES AND RELIABILITY ESTIMATES

Test name	N (items)	Mean	Variance	Cronbach's alpha (α)
German 1999 (n = 30)				
Part 1	30	18.43	27.98	0.83
Part 2	39	17.97	50.24	0.85
Part 1 + Part 2	69	36.40	138.11	
German 2001 (n = 48)				
Part 1	30	20.77	40.56	0.90
Part 2	39	19.38	101.86	0.94
Part 1 + Part 2	69	40.15	245.70	
Spanish 1999 (n = 286)				
Part 1	32	24.17	24.40	0.81
Part 2	35	24.46	32.16	0.81
Part 3	25	14.98	25.77	0.85
Part 1 + Part 2 + Part 3	92	63.60	203.16	
Spanish 2000 (n = 393)				
Part 1	32	22.80	28.36	0.83
Part 2	35	22.49	41.10	0.86
Part 3	25	14.03	35.95	0.89
Part 1 + Part 2 + Part 3	92	59.40	264.77	
Posttests				
German 2001 (n = 14)		46.57	181.19	
Spanish 2001 (n = 17)		75.18	302.03	

Table 2

TESTS OF STATISTICAL SIGNIFICANCE OF DIFFERENCES OF MEANS, VARIANCES, AND EFFECT SIZES

	German 2001 Posttest		
	t	*F*	*d*
German 1999	2.72 **	1.29	0.76
German 2001	1.72 *	1.29	0.48
	Spanish 2001 Posttest		
	t	*F*	*d*
Spanish 1999	2.67 **	1.57	0.67
Spanish 2000	3.63 ***	1.21	0.97

* $p = .05$
** $p < .01$
*** $p < .005$

nal in either 1999 or 2000. The variance issue required a technical analysis to ensure that the scores did not come from different distributions. We concluded that they all came from the same distribution, as there is no significant difference in any of the cells.

Finally, we calculated effect sizes. This is a way of converting raw scores into standard scores to determine how effective the treatment is. In this case the effect sizes ranged from .42 to .81. What these effect sizes represented were how much the instruction benefited the students over a theoretical control condition. An effect size of .42 is equivalent to an 16% increase (putting students at the 66th percentile compared to a control group at the 50th percentile); one of .81 is equivalent to an increase of 29% (79th percentile).

For the purposes of this study, we viewed the first administration of the placement test as the baseline and instruction in the courses a student took as a result of placement as the treatment. The gain in the second administration of the placement test is a measure of the effectiveness of the instruction as a function of placement. The larger the effect sizes, the more powerful the treatment. Effect sizes were used instead of normal test scores because they represent a common metric, allowing comparisons to be made between different tests.

User Perceptions

In order to probe the effect of online testing on teachers and program administrators, two sets of data were collected and examined: enrollment patterns and individual interviews with program teachers and program administrators. Table 3 recaps the total number of incoming students in German and Spanish who participated in online placement during 1999 and 2000. Of that total, it was difficult to sort out those who were placed and those who actually pursued a particular piece of advice. Given that placement testing is advisory and no placement is compulsory, this state of affairs portrayed a typical set of challenges faced by language program directors at the beginning of academic semesters and quarters, and most particularly in the fall, when large numbers of new students populate college and university campuses. Table 3 also reveals the fall quarter enrollments in Spanish and German courses into which students were placed. Seven Spanish instructors teaching courses into which students were placed (accounting for 18 sections, approximately 300 students) reported that they recommended to somewhere between 10 and 15% of students in any given section that they move to other sections—generally to higher level classes. Second-year instructors (N = 4) who receive students exiting from the language requirement reported more dissatisfaction, frequently suggesting that students advance to higher level courses in the second year rather than beginning at first semester/quarter, second year, where they were placed. German teachers (N = 3), with far fewer students to handle, reported satisfaction and claimed that they had not recommended that students move to other levels/courses.

Instructors were also interviewed about perceptions of the placement testing procedure in relation to their teaching. They all agreed that the most important aspect of the procedure was the opportunity to hear the oral interviews of students. Teachers contended that these interviews gave them insight into what they could expect from students and direction in terms of what they as teachers need to prepare. At one level, they reported perceiving the written sections of the online placement examination as making room for oral testing; they generally did not perceive the individual grammar scores as especially critical data.

The supervisors of the Spanish program and the German program (each of whom monitors the online placement testing process and scores the writing sections) remarked in their interviews that there are two principal efficiencies inherent in online placement testing. First, examining student grammatical performance in the summer, well in advance of the beginning of classes, affords the

Table 3

NUMBER OF PLACEMENTS AND TOTAL ENROLLMENT FOR SPANISH AND GERMAN STUDENTS

	Student Placement	Fall Enrollment
Spanish		
1999	332	560
2000	400	592
German		
1999	31	93
2000	48	101

opportunity to think about the curriculum as it is forecast for any given academic year, and grants time to fine tune teaching plans. This information is particularly useful in working with new teaching assistants and getting them socialized into examining individual learner performance. It also permits some time for teaching staffs to examine, in a relatively global fashion, the skills with which secondary school learners enter college. Second, they noted that summer placement provides the efficiency of examining the total number of sections needed for the year with time to open and close scheduled sections. Not only does this assist in budget planning, but it also makes requests to instructors to change sections, and perhaps course levels, easier with advance warning.

Conclusion

Placement testing via the Internet can be reliable and can validly reflect a foreign language curriculum. In this study, no adverse effects seemed to be at play either in terms of how students interacted with tests on the Internet or in terms of what the tests actually measured. The placement tests used in this study were sensitive to instruction and provided a good match between placement decisions and the curriculum. Succinctly, the tests led to appropriate instructional decisions that enabled students to benefit from instruction.

At some level, the key advantage of online testing is not the online testing itself, but what it enables supervisors, instructors, and students to do. For students, accessing a test at their convenience without making an extra summer trip to campus for placement testing is seen as an incredible time saver. At the same time, having students participate in an academic exercise prior to arriving on campus sends a very positive message regarding the importance and prestige of the language program among many other university programs. In like manner, administrators report that the savings from eliminating the extra step throughout a summer orientation period in particular are significant. Fewer scheduling logistics make orientation programs run much more smoothly. Supervisors and instructors are able to make more effective decisions when they have time to contemplate a given student's performance. More individualized instructional decisions allow for more confident judgments about where students should begin their university language experience and send the message to students that their individual performance and ability to use language in all its facets is at the forefront of instructional decisions. Instructors report a greater confidence level in their curriculum when they encounter students during the first class hours.

Technology should never be and will never be a substitute for good teachers. This study provides evidence that technology can bring efficiencies into instruction that enable teachers to focus on what is really important (i.e., students and their language development). In this study, a normally onerous task was moved to a Web-based electronic environment. Moving the task to the Internet environment afforded the scoring and analysis of those scores at a relatively leisurely pace. It also opened up time for instructors to focus on the really challenging part of language learning—how students put the language together, principally in oral speech.

References

Alderson, J. C. (1988). New procedures for validating proficiency tests of ESP? Theory and practice. *Language Testing, 5*(2), 220–32.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Upper Saddle River, NJ: Prentice-Hall.

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

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

Boo, J. (1997). *Computerized versus paper-and-pencil assessment of educational development: Score comparability and examinee preferences*. Unpublished doctoral dissertation, The University of Iowa, Iowa City, IA.

Bresolin, Jr., M. J. (1984). *A comparative study of computer administration of the Minnesota Multiphasic Personality Inventory in an inpatient psychiatric setting*. Unpublished doctoral dissertation, Loyola University, Chicago.

Bunderson, V. C., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In L. R. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 367–407). Phoenix: Oryx Press.

Burke, M. J., Normand, J., & Raju, N. S. (1987). Examinee attitudes toward computer-administered ability testing. *Computers in Human Behavior, 3*, 95–107.

Byrnes, H. (1990). Priority: Curriculum articulation. Addressing curriculum articulation in the nineties: A proposal. *Foreign Language Annals, 23*(6), 281–92.

Chalhoub-Deville, M., (Ed.). (1999). *Issues in computer-adaptive testing of reading proficiency*. Cambridge: Cambridge University Press.

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge: Cambridge University Press.

Day, C. L. (1999). A predictive validity study of computer adaptive placement tests for Tennessee higher education institutions. (Doctoral dissertation, University of Texas, 1999.) *Dissertation Abstracts International, 59*(7-A), 2464.

Dimock, P. H., & Cornier P. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement and Evaluation in Counseling and Development, 24*, 119–126.

Dyer, H. S. (1947a). Validity of C.E.E.B. placement test in French. *College Board Review, 1*(1), 12–15.

Dyer, H. S. (1947b). Validity of the German placement test. *College Board Review, 1*(1), 24–26.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Goodman, J. F., Freed, B., & McManus, W. (1990). Determining exemptions from foreign language requirements: Use of the Modern Language Aptitude Test. *Contemporary Educational Psychology, 15*(2), 131–141.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE.

Hamilton, L. S., Klein, S. P., & Lorié, W. (2000). Using Web-based testing for large-scale assessment. RAND Education Document No. IP-196.

Harlow, L. L., & Caminero, R. (1990). Oral testing of beginning language students at large universities: Is it worth the trouble? *Foreign Language Annals, 23*(6), 489–501.

Harrel, T. H., Honaker, M. L., Hetu, M., & Oberwager, J. (1987). Computerized versus traditional administration of the multidimensional aptitude battery-verbal scale: An examination of reliability and validity. *Computers in Human Behavior, 3*, 129–137.

Harrison, A. (1986). *A language testing handbook*. London: Macmillan.

Hughes, A. (1986). A pragmatic approach to criterion-referenced foreign language testing. In M. Portal (Ed.), *Innovations in language testing*, (pp. 31-40). Philadelphia: NFER-NELSON.

Jonassen, D. H. (1986). *Effects of micro-computer display on a perceptual/cognitive task*. Paper presented at the annual meeting of the Association for Educational Communications and Technology, Las Vegas.

Klee, C., & Rogers, E. (1989) Status of articulation: Placement, advanced placement credit, and course options. *Hispania, 72*, 264–74.

Kumar, D. (1996). *Computers and assessment in science education*. ERIC Digest (ERIC Document Reproduction Services No. ED395770).

Kuo, J., & Jiang, X. (1997). Assessing the assessments: The OPI and the SOPI. *Foreign Language Annals 30*(4), 503–512.

Larson, J. W. (1989). S-CAPE: A Spanish computerized adaptive exam. In W. F. Smith (Ed.), *Modern technology in foreign language education: Applications and projects* (pp. 277–289). Lincolnwood, IL: National Textbook Co.

Levin, T., & Gordon, C. (1989). Effect of gender and computer experience on attitudes toward computers. *Journal of Educational Computing Research, 5*, 68–88.

Madsen, H. S. (1991). Computer-adaptive testing of listening and reading comprehension: The Brigham Young approach. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice*. (pp. 237–257). New York: Newbury House.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations*, (College Board Report 91-5). Princeton, NJ: ETS.

Messick, S. (1989). Validity. In R. L. Linn (Ed.)., *Educational measurement* (3rd ed.) (pp. 13–103). New York: American Council on Education.

Powers, D. E., & O'Neill, K. (1992). *Inexperience and anxious computer users: Coping with a computer-administered test of academic skills* (RR-92-75). Princeton, NJ: Education Testing Service.

Shepard, L. E. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.

Shohamy, E. (1998). Evaluation of learning outcomes in second language acquisition: A multiplism perspective. In Heidi Byrnes (Ed.), *Learning foreign and second languages: Perspectives in research and scholarship* (pp. 238–261). New York: Modern Language Association.

Stansfield, C. W., & Kenyon, D. M. (1992a). The development and validation of a simulated oral proficiency interview. *Modern Language Journal, 76*, 129–141.

Stansfield, C. W., & Kenyon, D. M. (1992b). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*, 347–64.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (Report 61). Princeton, NJ: Educational Testing Service.

Teschner, R. (1990). Spanish speakers semi- and residually native: After the placement test is over. *Hispania, 73*, 816–22.

Van de Vijver, F. J. R., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79*, 852–859.

Vispoel, W. P., Bleiler, T., Boo, J., Steger-May, K., Lin, C., & Turhan, A. (1997). *Computer versus paper-and-pencil assessment of self-concept: Psychometric equivalence and respondent preferences.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Ward, T. J., Hooper, S. R., & Hannafin, K. M. (1989). The effects of computerized tests on the performance and attitudes of college students. *Journal of Educational Computing Research, 5*(3), 327–333.

Wherritt, I., & Cleary, T. A. (1990). A national survey of Spanish language testing for placement or outcome assessment at B.A.-granting institutions in the United States. *Foreign Language Annals, 23*, 157–65.

Wherritt, I., Cleary, T. A., & Druva-Roush, C. (1990a). Development and analysis of a flexible Spanish language test for placement and outcome assessment. *Hispania, 73*, 24–29.

Wherritt, I., Cleary, T. A., & Druva-Roush, C. (1990b). The development of a foreign language placement system at the University of Iowa. In Richard Teschner, (Ed.), *Assessing foreign language proficiency of undergraduates* (pp. 79–92). Boston: Heinle and Heinle.

Wildgrube, W. (1982). Computerized testing in the German Federal Armed Forces: Empirical approaches. In D. J. Weiss (Ed.), *Item response theory and computerized adaptive testing conference proceedings* (pp. 353–359). Minneapolis, MN: University of Minnesota.

Williams, S. B., & Leavitt, H. J. (1947). Prediction of success in learning Japanese. *Journal of Applied Psychology, 31*, 164–168.

Wise, S. L., Boettcher L. L., Harvey, A. L., & Plake, B. S. (1987). *Computer-based testing versus paper-and-pencil testing: Effects of computer anxiety and computer experience.* Paper presented at the annual meeting of the American Educational Research Association, April, Washington, DC.

Appendix

Advisory Information

This exam is advisory to you and your professors. We want to recommend a course for you which is appropriate to your level of [foreign language]. We want to make sure that you make wise use of your time at Stanford. You shouldn't take a course which is too elementary for you—that is, one in which you already know the material. You also shouldn't take a course that is too advanced—this leads to frustration and to the complications of switching sections or to dropping out and starting over.

While taking this test, remember that you are now a Stanford student and subject to the Stanford Honor Code. It is very important that you do your best on this exam AND that you do not receive any help with it. In order to make good judgments about YOUR placement, we need good information about YOUR knowledge of [foreign language].

The results of your exam will be known to you and to members of the Language Center who direct placement testing. They will not be reported to all members of the department, to the registrar, or recorded on your transcript. Remember this is a placement test. It is not graded in the traditional sense, but used to match you to an appropriate level of [foreign language].