AAUSC 2014 Volume – Issues in Language Program Direction

Innovation and Accountability in Language Program Evaluation

John Norris, Georgetown University
Nicole Mills, Harvard University
Editors

Stacey Katz Bourns, Harvard University
Series Editor

CENGAGE Learning®

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

AAUSC 2014 Volume - Issues in Language Program Direction: Innovation and Accountability in Language Program Evaluation
John Norris, Nicole Mills, and Stacey Katz Bourns

Product Director: Beth Kramer

Product Assistant: Jacob Schott

Marketing Brand Manager: Christine Sosa

IP Analyst: Jessica Elias

IP Project Manager: Farah Fard

Manufacturing Planner: Betsy Donaghey

Art and Design Direction, Production Management, and Composition: Lumina Datamatics, Inc.

Library of Congress Control Number: 2014948811

ISBN-13: 978-1-305-27509-6

Cengage Learning
20 Channel Center Street
Boston, MA
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at **international.cengage.com/region.**

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit **www.cengage.com.**

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com.**

Instructors: Please visit **login.cengage.com** and log in to access instructor-specific resources.

Printed in the United States of America
Print Number: 01 Print Year: 2014

Chapter 2

The Development, Management, and Costs of a Large-Scale Foreign Language Assessment Program

Elizabeth Bernhardt and Monica Brillantes, Stanford University

Introduction

The Stanford Language Center has a 17-year history of designing, implementing, and managing an assessment program that spans two years of instruction across 12 languages. This assessment program is considered to be one of three key, intertwined elements associated with the success of the center. Alongside an elaborate professional development program and the extensive use of technology, the assessment and subsequent publication of student performance levels is often mentioned by university officials as helping to bring about enthusiasm for language learning and for reinforcing and sustaining ambitious learning goals, particularly in speaking and writing. The first part of this chapter will chronicle the scope of the assessment program, the assessment measures used, the relationship of these measures to the stated curricular objectives and outcomes based in the *National Standards* (National Standards in Foreign Language Education Project, 2006), and the assessment strategy's link to the professional development program. This first section ends with a discussion of how assessing student performance and providing public documentation of it has influenced the status and authority of the Language Center within the university.

The substance of the chapter, however, focuses on monetary and nonmonetary costs of establishing and maintaining this type of systematic and systemic assessment program and whether the costs engendered bring about benefits to students, to their teachers, and to the university. Adding to the general public narrative on assessment, the chapter illustrates a program that uses performance indicators in speaking and writing to document the efficacy of an institution's foreign language agenda. It describes in detail a funded mandate and tries to answer the question of how much each dimension of such an assessment program costs—from placement testing to exit testing to, perhaps most importantly, concomitant teacher development costs. Ultimately, the chapter poses the fundamental question of whether the investment in assessment provides a significant return in terms of student performance.

By providing financing information, we intend to offer the profession some baseline information not in terms of raw dollars, but rather in terms of size and scope for adopting and adapting successful assessment programs. This information is intended to provide a perspective, not a set of rules, as costs change and

are regionally determined and institutions may choose other types of directions. The purpose is therefore to acknowledge that the program outlined had real costs attached to it and that any postsecondary institution considering an assessment strategy must calculate its costs rather than simply impose a mandate on already overworked instructors. The intention is that the data displayed may offer guidance to those writing funding proposals and developing budgets for the establishment of assessment programs. Examining financial records provides concrete numbers, but perhaps more importantly, it reaffirms the extent of staff time devoted to many dimensions of language teaching and the incredible personal commitment that language teachers must make to their students even when funding is available.

Cost is often related to the negative reputation that *assessment* engenders in the popular and scholarly press. One side of the narrative refers to assessment as the documentation of a public good. This perspective tries to address the question of what the public return is for the investment of educational tax or tuition dollars. The counterpoint to this perspective is that public investment, that is, tax or tuition dollars, is increasingly limited, and therefore, the discussion is moot given that the public tends to show little interest in investing money in assessment. Assessment can then become just another unfunded mandate for educational institutions to shoulder. Another consideration within the assessment narrative is its impact on the profession of teaching (Bernhardt, 2000, 2006; Joyce & Showers, 1988). In this part of the story, merit pay for teachers is often attached to student outcomes. Student test scores become a measure of teacher performance as well as a gauge for contract renewal, potentially negatively impacting teacher morale (Ramirez, 2001). The counterpoint to this dimension of the narrative is that most professions, industries, and occupations have performance-based salary structures, and the question that is then raised is why teaching is considered as fundamentally different from other careers. We would claim that the entire argument can be captured in the word *accountability*. Among the myriad national and international discussions on educational accountability is a commentary funded by the World Bank:

> Universities should be accountable in various ways to students, parents, employers, and to the general public. There is a perception, widespread in some countries, that the university—and especially the classical university and the professorate—is insufficiently accountable, particularly to the first degree student. Accountability is difficult to achieve because the benefits are (appropriately) both multiple and hard to measure. But it is nonetheless essential to lay down transparent guidelines, to install better measures of outputs, or performance, and to better align both individual and institutional rewards with these performance indicators. (Johnstone, 1998, p. 6)

The report concludes that "the principal higher educational productivity problems lie not so much with excessive costs, but with insufficient learning" (p. 7).

A rash reaction to enhancing student performance, that is, solving the most critical of educational problems, often appears in the shape of restructuring strategies. In primary and secondary schools, various strategies have been instituted

in an array of settings. Such strategies include site-based management, class size reduction initiatives, an increased use of online supplements to an already overcrowded curriculum, stripping union contracts to restrict teacher privileges, or increased opportunities for private management of schools. Adams (2010) argues that the implementation of this confusing and complex array of strategies has been shocking and notes that "what's missing from the new landscape are improvements in student learning that match public expectations and that, after a survey of all the commotion, would make a 'tough but worth it' judgment more satisfying" (p. 4).

The establishment of a language center at Stanford was indeed an example of restructuring to solve an educational problem or, better said, a perceived lack of performance or productivity on the part of language programs. Educators often argue that centralization is just another administrative ruse to limit resources and to force faculty to do more with less. This was not the case in the instance outlined in this chapter. The principal feature of the institution's desire for the centralization of language programs and the vision provided in that centralization was a focus on *heightened student performance*. Undeniably, the concept of central leadership with a vision, that is, the Language Center, is critical for understanding both the development and the purpose of an extensive assessment program and its end results.[1] Centralization led to a substantive investment in technology, in teachers, and in the assessments they use to determine whether students are meeting high performance standards.

Institutional Context: The Funded Mandate

In 1994, Stanford University completed an extensive undergraduate curriculum review. Stanford's need to internationalize, that is, to have more students study abroad and to foster greater faculty involvement in research and development overseas, was identified as a key future direction. Within the context of faculty discussions regarding internationalization, language proficiency on the part of the student body was targeted as a concern. Faculty across the university—engineering, environmental science, biological sciences, anthropology, international relations—all identified interests in preparing students to live, work, study, and research in foreign settings. A consensus developed that the national literature departments, with their focus on preparing students for text analysis, were not appropriately configured to efficiently prepare a larger and more diverse student body. The Senate recommendation was to create a language center charged with broadening the language curriculum for this more diverse student body. The university Senate gave the Language Center the following directives: (a) ensure excellent language programs for all Stanford students; (b) launch a professional

[1] A key point is that in the restructuring of language programs and the creation of a language center, the institution sought a tenured full professor in the area of applied linguistics to take on the leadership role. The university administration recognized that only a person with full academic privileges and responsibilities as well as a knowledge base in linguistics, assessment, professional development, and finance could bring about significant and sustained systemic change.

development program for the teaching staff that will enable them to broaden the curriculum, making it more attractive and accessible to all students, not just humanities students; (c) infuse technology into language instruction; and (d) establish a research program in language teaching and learning.

Insuring the quality of the language programs meant, first of all, establishing what that quality should be; in other words, quality had to be linked to what the language programs were intending to accomplish. At the time, the answer from the teaching staff was generally that the first-year language programs should ensure that students *know* the material reviewed in their respective first-year textbooks; the second-year language programs should review the first-year content *(i.e., students will improve their grammatical ability)* and prepare students to read and translate literature. By language teaching standards in 1995, these were, of course, feeble goals. The university was demanding that students be prepared to live, work, study, and research abroad; these goals reflected the wider focus in the language teaching community of *proficiency* or of what learners can *do*. The real question to be answered was *what should students be able to do in listening, reading, writing, and speaking at the end of a first-year (30-week) sequence and at the end of an additional 30-week (second-year) experience?* These questions were the ones that were reflective of the university-wide concern about language programs meeting campus needs.

For a full year, the Language Center director was in communication with each language program to discuss *what students should be able to do*. Because many teachers had never considered student performance outside of textbook coverage, these were trying discussions. Others who were open to the discussion often possessed little professional vocabulary with which to discuss language-based performance outside of *grammar test* scores. Literature faculty, at the same time, were uncomfortable with the notion of *doing* because their needs had ostensibly been met through a focus on grammatical form. Ultimately, however, brief goals were set by all language programs in the four skill areas of reading, writing, listening, and speaking. Posing the question of what learners should be able to do enabled the staff to open a conversation about the intentions of language programs. The establishment of underspecified goals in the four skill areas was clearly insufficient, but it set the stage for the articulation of targeted goals focused on student needs.

Contemporaneous with the establishment of the Stanford Language Center and others across the United States was the publication of the first edition of the *National Standards* (National Standards in Foreign Language Education Project, 1996). A member of the original *Standards* task force and Stanford professor of Spanish and Portuguese, Guadalupe Valdes, was extremely influential in assisting the language programs in the articulation of objectives consistent with the *Standards* (Bernhardt, Valdes, & Miano, 2009). It was the need for the articulation of objectives and a process for establishing them through the *National Standards* that facilitated the requisite professional conversation that had been virtually nonexistent and created a growing atmosphere of mutual self-respect and peer reliance among the full teaching staff. Ultimately, each language program generated curricular documents that were organized according to 10-week quarters across two years of instruction and focused on interpersonal abilities, presentational

abilities, and interpretive abilities[2] (see http://language.stanford.edu). As the objectives of the curriculum became clearer, the need for an assessment program beyond a grammatical placement test and chapter quizzes and tests became more acute (Bernhardt, 2000). That need generated a grassroots demand for additional professional training in the arenas of oral and written proficiency testing.

Dimensions of the Assessment Program

Assessment processes have been institutionalized in three major categories at the Stanford Language Center. The concept of institutionalization rather than centralization is used to underline that all language programs proceed through particular processes, most of which have identical technical dimensions, without losing the individuality of languages with specific characteristics. In other words, some language programs, such as Chinese and Arabic, need to ask learners to handwrite at times to ensure that these learners can use characters and other orthographies appropriately. Other language programs such as Spanish and Italian prefer to have students move almost immediately into reading authentic, unedited material and to keyboarding because of orthographic overlap. In other words, all dimensions of the assessment program, though consistent, were customized by language.

Because the Language Center monitors and enforces the language requirement, accurate placement testing is critical. Equally critical is accuracy and reliability within high-stakes exit assessments. These exit exams carry with them, on the one hand, the integrity of the entire curriculum and, on the other hand, requirements for certain majors, such as International Relations, or the notation of Advanced Proficiency inserted on student transcripts. Exit assessments indicate whether the curriculum is indeed rigorous and producing active users of the foreign language. Finally, the interim assessment program, that is, somewhere between placement and exit, garners attention because its assessment of interim objectives must be consistent with the assessment of concluding objectives. In other words, formative assessment must be consistent with the values encompassed in summative assessments.

Recall that the overriding values held by the Language Center programs, consistent with those of the *National Standards* and the local university community, were that students should be prepared to live, work, study, and research in non-English-speaking settings. This value implied that students must be prepared to listen, speak, read, and write. To determine whether students were able to use the language productively in speech, the Simulated Oral Proficiency Interview (SOPI) was adopted as a primary assessment tool. This assessment protocol was originally developed and validated by the Center for Applied Linguistics (Malone, 2000).[3]

[2] Bernhardt et al. (2009) provide an extensive description of curriculum renewal and document the iterative process of objective setting, staff reaction and involvement, implementation, and revision.

[3] Criticism can be attached to any assessment instrument, and the SOPI, based in the OPI framework, is not immune.Yet, because published evidence of the reliability and validity of the SOPI protocol existed, this choice was considered to be a reasonable one.

Items on a SOPI are delivered in audio or video form via CD or DVD or on a computer, and students respond using a handheld digital recorder or directly into a computer interface. Tests usually contain 10 to 20 items of gradually increasing difficulty, and test takers are asked to provide 30- to 90-second responses to various situations. Responses are rated according to the proficiency guidelines developed by the American Council on the Teaching of Foreign Languages (ACTFL). In addition to the SOPI as an assessment instrument, individual instructors were trained and became certified in administering the Oral Proficiency Interview (OPI), a one-to-one administered assessment, also rated according to the ACTFL's Proficiency Guidelines (Breiner-Sanders, Lowe, Miles, & Swender, 2000). To determine whether the students were able to use the language productively in writing, writing proficiency assessments (WPA) were used. These assessment tools were developed in-house based on ACTFL's writing proficiency test (WPT) and included writing prompts consistent with ACTFL's Writing Guidelines (Breiner-Sanders, Swender, & Terry, 2002).

These two sets of instruments, based on the productive modes of speaking (S/OPIs) and writing (WPT/As), constitute the assessments used through the Language Center's assessment program. All assessments are cross-rated by certified proficiency raters on staff at the Language Center as well as by telephonic OPIs and computer-delivered WPTs administered to random samples of students by Language Testing International, an independent testing organization.[4] This procedure ensures the reliability and validity of the assessments used in the placement, formative, and summative dimensions of the program. The following paragraphs outline the use of the S/OPIs and WPT/As across the three assessment arenas within the Language Center.

Placement

The placement process has two parts for foreign language programs in the Language Center. Soon after new students are admitted to Stanford, they are able to sign up for an online placement test that includes reading comprehension, vocabulary, grammar, and in the case of the Romance and Germanic languages, composition and listening.[5] Placing these assessments online brought two major advantages. First, it enabled language coordinators to make preliminary placements that helped in planning section numbers and sizes, and second, it cleared the available testing time for the assessment of speaking (Bernhardt, Rivera, & Kamil, 2004). The university's New Student Orientation Office offers the Language Center only two 90-minute sessions to assess the placement of approximately 1,000 students. To meet the challenge of this restricted time period, the SOPIs are delivered to groups of students in classrooms and auditoria around campus. The SOPIs are then assessed by Language Center instructors who are

[4] Generally speaking, correlations range between .72 and .90 for SOPIs and OPIs across languages. When there are inconsistencies between SOPIs and OPIs, the SOPIs tend to receive lower ratings than OPIs. Insufficient data exist at present to report correlations between WPTs and WPAs.

[5] Placement test development and test specifications as well as relevant test statistics are included in Bernhardt et al. (2004).

certified OPI testers/raters, and results are posted by 5pm on the testing day. Both the online reading and grammar tests completed over the summer in conjunction with the on-campus speaking tests are used to place students in courses that fit their proficiency needs.

Formative Assessment

While each instructor has different ways to handle formative assessment, most programs deliver an agreed-upon set of oral diagnostic assessments (ODAs) and written online diagnostic assessments (WDAs) in Stanford's learning management system called Coursework©. These formative assessments are also grounded within ACTFL's Oral and Writing Proficiency Guidelines. They are, therefore, conceptually consistent with both placement and exit assessments and are delivered in technical formats familiar to students.

Instructors in each language program collaborate to create sets of items that incorporate combinations of text, image, audio, and video to ensure efficiency and consistency within and across programs. Students use an audio response applet in the learning management system to give spoken responses. These responses are then available online for instructor assessment and feedback. In the WDAs, written responses make use of a text box or upload function. Banks of items are available to instructors to enable them to easily construct a WDA. Each term, instructors import the items that they need into their courses for diagnostic assessments and they release them according to a schedule determined by the program coordinator. An academic technology specialist assists programs as they set up and deliver these assessments and a language lab service manager offers orientations in which students learn to use the specific assessment tools, as well as more general skills such as typing in languages other than English.

Formative assessments are also provided to students on their presentations, particularly in second-year courses. While most instructors agree that students learn the most from presentation practice when they have both thorough feedback from instructors and an obligation to reflect on their performance, it has proven to be difficult to facilitate both of these goals simultaneously. Traditional capture methods for student presentations, such as magnetic tape, require time- and material-consuming duplication, and digital capture necessarily means working with files that are so large that they are very difficult to share via email or other private online systems. To respond to this need, the language lab staff configured video cameras with directional microphones connected to laptops that capture video using settings that produce significantly smaller files. Currently, files can be immediately uploaded to the learning management system as formative assessment data for reflection by the student and review and feedback by the instructor.

Exit Assessment

At the end of the academic year, the Language Center administers speaking (SOPIs) and writing (WPAs) exit assessments for foreign language students completing the first- and second-year sequences. SOPIs are administered in the language lab cluster using software specially developed for this purpose. The SOPI software is an application installed on each of the cluster machines by the imaging team in the

language laboratory, but only for the few weeks during which the assessment is taking place. During the last two weeks of the academic year, students go to the language lab during one of their regular class meetings with their teacher to take their SOPI. In parallel to the placement testing procedure, items in the test consist of an oral description of a scenario (e.g., helping a classmate move items into storage before leaving for a study abroad experience) in English, the native language of most learners, often accompanied by a pictorial representation of the situation. Test takers are given a short 15- to 30-second preparation period, and then they hear a prompt in the target language such as "what types of things do you need to move into storage?" The software then automatically starts recording for a specified period of time, after which it moves on to the next item. The user interface of the software has only one button that allows students to move to the next item if they finish before the time is up. There is no way to pause the recording or listen to the item again. First-year students typically do nine of these items (a short-form SOPI), while second-year students do 18 (a long-form SOPI), with the difficulty slowly increasing through the proficiency levels as they proceed. The software stores each response on the computer until upload to the server has been confirmed, and the test cannot be exited until all items have been uploaded. The student responses are then available in a special tool in the learning management system, where raters listen to each response and ultimately determine a global rating for each student.

A writing proficiency test is also part of the exit assessment, and while the idea behind the test is quite simple, the delivery has several conditions that make it quite challenging. The use of writing prompts is quite straightforward when delivered on paper in a classroom, but attempting to deliver them on computers means locking down devices so that students do not have access to any resources that might help them compose text in the target language. The language laboratory imaging team was able to configure inexpensive laptops that provide access to the learning management system and block all other sites, including email, social networking, and online translation services. Individual classrooms are reserved for the delivery of these writing assessments.

The oral and written exit assessments are rated throughout the summer,[6] collated by the Language Center, and the ratings are reported to the university Senate. Each language program also has immediate access to the data, by individual student and by class section. This process enables a discussion of the curriculum and alerts the teaching staff to both strengths and weaknesses in the speaking and writing abilities of students.

The Impact of Documenting Student Performance on the University

The university began its most recent regional accreditation process in 2009; the previous one had been completed in 2000 through the Western Association of Schools and Colleges (WASC). In the 2009 process, the language programs were

[6] This process became a job creator. Raters are generally offered between one and two months of summer salary for their assessment work. The opportunity for summer salary is widely sought.

chosen as a focus area because of the need to evaluate the university's decision to enhance the language requirement *and* perhaps more importantly because the Language Center had proficiency data available and was in a position to generate more data based on a national scale that enabled comparison. Succinctly, the Language Center had an assessment program as a systemic part of its portfolio and that catapulted the language programs into a position of prominence.

In its self-study, the university included three data points about Language Center programs.[7] First, the language data available from the inception of the Language Center indicated that student oral language performance increased statistically significantly over the 12 years of data available within the focus languages of French, Spanish, and Chinese, amounting to around 50% of the total enrollment in language courses. This finding was a powerful statement to make in an era of accountability. The second data point was an analysis of 12 years of student placement data and subsequent language acquisition. The placement data confirmed that any Stanford student who placed into a language section, and at some point enrolled and completed a foreign language course, increased in oral proficiency. This finding provided evidence against the accusation that students in some universities are merely good when they enter and good when they exit with little interim additional learning (Arum & Roksa, 2011). This was not the case in the Language Center. The final data point came from pre–post testing using OPIs with students studying abroad at campuses in Madrid, Spain; Santiago, Chile; Beijing, China; and Paris, France. Before embarkation, students took an OPI, and at the end of their stay another OPI was administered on site. This pre–post assessment documented that students were well prepared for their study abroad experience and that they increased in their oral proficiency at the site. The report notes that all students progressed at least one half step on the ACTFL/OPI scale (from Mid to High, for example) and some crossed major borders (e.g., from Intermediate High to Advanced Low).

Because the Stanford Language Center had a systemic and systematic assessment program, it was able to become a focal point of the university's review. It is unusual for almost any area of the humanities in the modern era to become a focal point of a university review, particularly at a university known for its technical prowess. The conclusion of the visiting team's report (WASC) underlines the significant positives accruing from the Language Center's assessment program:

> Two aspects of this examination of foreign language proficiency demand special praise at the outset. First is that the success of the language programs in creating a high level of foreign language proficiency is tied successfully to core aspects of Stanford's undergraduate curriculum. One cannot create global citizens without instilling linguistic fluency in languages other than English. We therefore commend Stanford for putting foreign language assessment front and center in this report and viewing it as a central, rather than a marginal University concern.

[7] The WASC Review (2010) contains the analysis of oral proficiency ratings over time, placement test scores, and pre- and post-test oral proficiency ratings of students studying abroad.

> Second is the extension of the analysis of the success of language teaching to the results of study abroad experiences in enhancing linguistic fluency. . . . The Stanford Language Center is already known as a national leader in the teaching of languages other than English, and both this study and the central role it has played in Stanford's re-accreditation effort will only enhance its deservedly high standing. (WASC Visiting Team Capacity and Preparatory Review, 2010, p. 14–15)

Linked to and, in part, stimulated by the regional accreditation process, was an internal two-year-long undergraduate program review. Published in January 2012, the report acknowledged the data submitted by the Language Center over the years and then concluded:

> Our conversations with students reinforced what the numerical data told us. Stanford students enjoy their language courses, recognize the value of the skills they acquire within them, and relish the opportunity to deploy those skills overseas. In stark contrast to conversations about most other general education requirements, no one we spoke to suggested abolishing or reducing the language requirement. (University Board of Trustees, 2012, p. 32)

For those affiliated with the Language Center, this statement, generated by faculty and staff across the university, reinforced how incredibly positive an assessment program can be for students, teachers, and the university. Our experience proved that evidence is more powerful than emotion in a university context.

Monetary and Nonmonetary Costs of the Assessment Program

This large and expansive program ultimately resulted in benefits to the students and their teachers and to the esteem of the university. The size of the language programs and increased attention to them on the part of university administrators underline the contention that the process outlined earlier in text was valuable. The question for the remaining section of this chapter is *how much did all of this cost?* Our answer to this question intends to offer guidance as well as cautions as educational excellence comes at a cost, both monetary and nonmonetary. That cost is usually on the shoulders of already overburdened teachers, and this situation simply should not stand. *What happens when the costs are examined separately?*

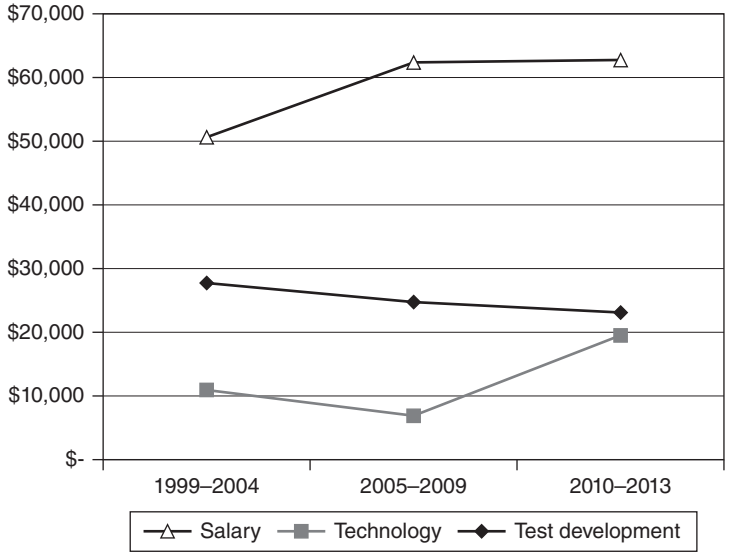
To answer these questions, two five-year periods—1999 to 2004 and 2005 to 2009—within the context of two commonly taught languages, Spanish and French, were examined. The two periods signify the periods used in the analysis for regional accreditation (WASC) launched in 2000 and 2009. They also signify, in the former instance (1999–2004), a period of building toward a comprehensive assessment program (i.e., attempting to assess all languages within the portfolio of the Language Center). The latter (2005–2009) refers to a period when the assessment program was fully normalized and included all languages. By choosing two languages, the Language Center was able to provide information about a

large sector of the enrollments (around 30% of total enrollment, which is generally 6,000 enrollments over an academic year) and some degree of anonymity with regard to the reality of the financial numbers offered.[8]

Placement Testing

The Stanford Language Center began online placement testing in June 1996; this was the first recorded use of the Web for placement testing (Bernhardt et al., 2004). The initial setup of the Stanford placement tests came at the cost of a PowerMac 7100 computer priced at $2,600. A staff member was anxious to have the experience and responsibility of designing Web applications and offered to set up the placement testing system for the cost of this computer. The initial setup included French, German, Spanish, and Chinese and remained in place until 2004. At that time, other regularly taught languages, including Italian, Russian, and Japanese, were added. These additions signified a need for some additional technical abilities, plus the management of a much wider range of students who would log into the system. To accommodate these additions and revisions, a webmaster was contracted, who maintained an 800 number in order to answer students' technical questions over the summer in addition to regularly updating the tests and technical arrangements. These costs are captured under "technology" in Figure 2.1 and represent expenditures for the Spanish and French portion of the placement

Figure 2.1 Costs related to managing placement testing activities for the French and Spanish language programs.



[8] The cost data compiled for this chapter originated from the examination of financial statements. The two periods discussed in the chapter, 1999–2004 and 2005–2009, were used to calculate expenditures for test development, technology purchases, teacher professional development activities, and teacher salaries. This chapter provides collated data whenever possible for the programs in Spanish and French, not precise cost accounting, and intends to provide useful estimates to readers.

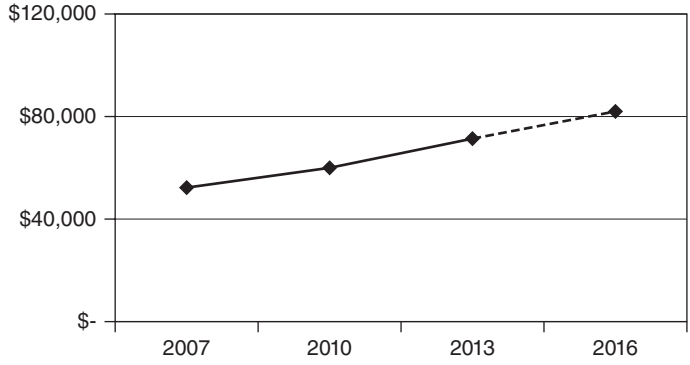
testing program. Summer placement testing also required that instructors be on staff throughout the summer to conduct any manual grading, that is, reading and assessing each student's writing sample, and to make preliminary placements. Each staff member charged with monitoring and rating online assessments began receiving a one-ninth salary summer stipend ("salary" in Figure 2.1). The cost of new SOPI items for the final element of the placement testing process completed during new student orientation, as well as refined artwork and new writing and grammar prompts, are also captured in Figure 2.1 under "test development." Costs have been relatively stable over the years with the exception of technology. A major purchase of digital recorders, moving from cassette recorders and tapes, accounts for this rise in costs. After this relatively recent outlay, technology costs for placement tests will also stabilize for a period of years.

Formative Assessment

In some sense, formative assessment is part of the cost of doing language teaching. Formative assessments need to be consistent with the objectives that the teaching staff establish for all courses and sequences and that, in the case of the Stanford Language Center, rely heavily on interpersonal and presentational language. To meet formative assessments that were consistent with the *Standards*-based objectives, the teaching staff developed ODAs, as mentioned earlier in the text. These were first developed for Spanish in 2007 and other languages followed suit. To reiterate, an ODA provides a prompt either in writing or via video that targets language functions such as narration and description in particular contexts, and students are then asked for an oral response. The ODAs (prompts and responses) are carried through the university's course management system. Instructors log in to the system to provide students with qualitative feedback on their oral responses (hence, *diagnostic*). Students are asked to complete an ODA every two to three weeks. Approximately a year after the development of ODAs, the staff began to develop WDAs. These assessments are also carried through the course management system and were developed as part of the expected instructor workload. Students are expected to complete three or four WDAs during the course of a quarter. These formative writing assessments target particular language functions, and students receive qualitative feedback from their instructors regarding their ability to use written functions proficiently.

As mentioned earlier in text, the ODAs and WDAs were developed out of the daily workload of instructors. At some level, formative assessments may appear to be cost free, but they are actually embedded in salaries. In reality, explicit costs relate to the price of technology used in the development and administration of the ODAs and the WDAs. The cost estimates that are available and relevant relate to computer replacement costs in the language lab. The university requires technology replacement every three years, and the Language Center is responsible for the equipment. The university libraries, charged with guidance in instructional technology, pick up the costs of technical staff for the language laboratory. Figure 2.2 illustrates the costs of the replacement of a 60-station laboratory equipped with dual-boot iMacs over the periods targeted in this chapter. The graph includes the projection for the next replacement period, which begins in 2016.

Figure 2.2 Costs related to language laboratory equipment replacement (including projections for the 2016 replacement cycle).



Admittedly, the equipment in the language lab is used by students for other dimensions of their learning such as foreign language homework and language projects in the laboratory. Informal student surveys conducted by the language laboratory staff evidence students' claims that the laboratory is a quieter and more comfortable place in which to work than their dormitory rooms even though, technically, they could complete many of their assignments from their personal computers or tablets.

Exit Assessment

Figure 2.3 delineates the financial investment in assessing student performance at the end of language sequences in Spanish and French over the two periods covered in this chapter. In the initial period, a SOPI developed by the Center for Applied Linguistics was used. Hence, the costs incurred were simply for a one-ninth salary stipend to cover the assessment of each oral performance ("salary"). In the second period under examination, a serious effort at SOPI development was undertaken locally, ultimately creating four additional equivalent forms ("test development"). The content of the forms was created in-house as part of general instructor responsibilities. Expenses incurred in test construction revolved about the development of artwork for the new forms as well as for the conduct of quality assurance and additional technical assistance for the placement of the forms into Stanford's course management system.

To facilitate the administration of writing assessments at the end of each language sequence, it was critical to enable students to write with a computer without language assistance provided by dictionaries, translation programs, and the like. In addition, the writing assessments needed to be administered conterminously with the oral assessments. This double challenge was met by reserving two classrooms for writing assessment and purchasing highly portable personal computers *without* internet connections. Netbooks were purchased for $33,000 along with carts to store them so that they could be easily moved from classroom to classroom for assessment. In parallel to the construction of four new SOPIs, WPAs consisting of two prompts, again targeted at particular language functions, were developed in-house by ACTFL-certified WPT raters.

Figure 2.3 Fixed and one-time costs related to managing exit assessment components for the French and Spanish language programs.

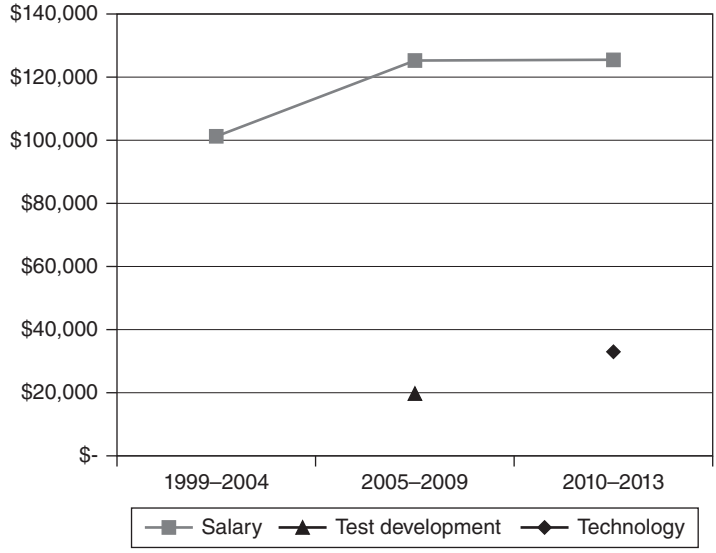


Figure 2.3 provides insight into regularly incurred, relatively fixed costs, such as summer salary for the rating of individual SOPIs and WPAs. It also illustrates one-time costs such as test development and technology replacement. The four forms of the SOPI will ultimately turn into many additional forms by interchanging tasks. In other words, the tasks on Forms A and B can be interchanged with those on Forms C and D to make new Forms E and F and so forth that use tasks from the earlier forms. This strategy was done deliberately so that the staff would not be in constant "exit assessment development" mode. Hence, only one data point, "test development," is plotted in Figure 2.3. The cost of technology for writing assessment will more than likely be placed on a three-year replacement cycle. As a result, only one data point, "technology," appears in Figure 2.3 plotted against ongoing salary costs.

The data included in Figure 2.3 reflect, on average, 500 assessments in both writing and speaking. Prorating these total costs (a little more than $380,000) across the years indicates an outlay of around $29,000 per year for exit assessments in Spanish and French.

Professional Development Costs

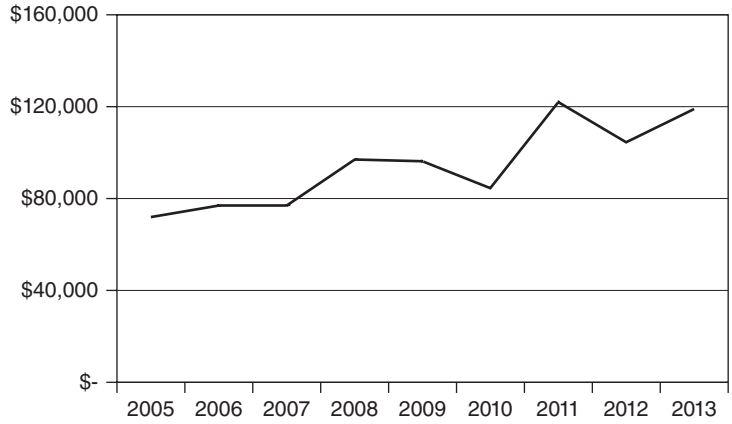
This chapter has referred repeatedly to the importance of having teachers highly knowledgeable about national foreign language curriculum initiatives and the development of assessment tools and skillful in rating assessments. In our experience in conducting more than 150 interviews for various lectureships across languages, it is clear that this array of skills is rarely found even in very experienced instructors and must be developed over time. In addition, one instructor on a staff with a well-established and elaborate set of professional skills among other

teachers with less-extensive credentials is a recipe for resentment and jealousy (Chen & Kristjansson, 2011). The teaching staff as a *professional* staff must have comparable professional experiences (Bernhardt, 2000). An integrated and consistent professional development program that includes all instructors enables a collegial conversation consisting of a common vocabulary and framework as well as a support system for that continued development (Bernhardt et al., 2009).

To achieve this coherent framework, the Language Center finances annual two-day OPI training sessions across multiple languages. The two-day workshops, known as Modified Oral Proficiency Interview workshops (MOPIs), focus on the lower range of the ACTFL/FSI (Foreign Service Institute) scale and enable participants to begin the process of tester certification. Staff members who receive limited certification or the ability to test and rate through the Intermediate High range then continue through to full certification by taking the following two-day OPI training workshop. These remaining days are regularly offered at regional or national meetings, and this portion is also financed by the Language Center. After receiving full OPI certification, or the ability to rate across the full ACTFL/FSI spectrum, most staff members continue through the writing certification process, consisting of a workshop and a rating protocol. Beyond these intensive experiences, teaching staff members are funded to attend national conferences as well as local and regional language teaching-related events and workshops.

Figure 2.4 illustrates the annual outlay for professional development expenses across all languages. It captures total expenses in recent years for workshops, certifications, travel to conferences, and so on. Given that different staff members are at different levels of the certification process (i.e., either beginning, in process, or recertifying) and that graduate students participate in all workshops but rarely complete the full certification, composite figures across all languages are offered, rather than limiting the costs to Spanish and French. The graph also illustrates the impact on staffing of the market crash in 2009 and the impact of the recovery on total instructional staff in 2011. On average, 75 full-time staff members and

Figure 2.4 Investments in teacher professional development.



around 25 graduate students participate in professional development activities annually, meaning that the per-person outlay for professional development each year is around $1,200.

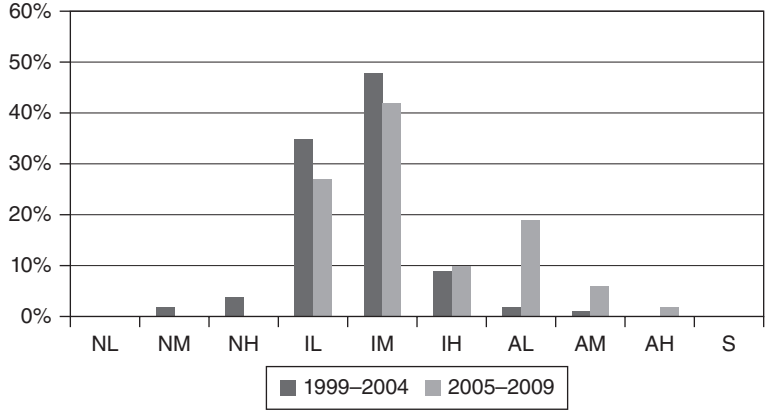
Monetary Costs and Student Performance

Of course, expending dollars is meaningful only if there is an understanding of student performance linked to the outlay of funds. Figure 2.5 illustrates student performance ratings in oral proficiency in the first- and second-year Spanish and French programs over the two focus periods discussed in this chapter. In the earlier period, 1999–2004, 48% of students exiting these programs were rated at the Intermediate Mid level in their oral proficiency. By the following five-year period, 42% were rated at the Intermediate Mid level, with the remaining students moving into higher ranges on the ACTFL/FSI proficiency scale. Nine percent of the Spanish and French students were at the Intermediate High range in speaking in 1999–2004, and in the following period, that rating increased to 10%. Importantly, a substantial increase is found in the Advanced range. In 1999–2004, only 3% of students were rated in the Advanced range. By 2005–2009, the rating rose to 27%.

Examining investments in the teaching staff over the same time periods lends insight into the relationship between professional development and student performance. In 1999–2004, $171,302 was expended in summer salaries for rating assessments, in instructors' professional memberships, for OPI and WPT training and certification processes, and for work on individual assessment development. In 2005–2009, those costs rose to $237,723. Salary increases rose from 0% to 3% over these periods. Teachers were not "paid" to produce results, but rather were taught to do so.

The link between professional development expenditures and increases in student performance illustrated in this chapter is admittedly tenuous. Yet it is consistent with the research on investment in teacher professional development to enhance learner proficiency. Data on restructuring in schools indicates that

Figure 2.5 Exit simulated oral proficiency ratings for first- and second-year French and Spanish.



curriculum leadership and professional development programs have a "greater impact on student learning and are relatively inexpensive compared to maintaining small class sizes" (Levenson, 2012, p. 49). A review funded by the Regional Educational Laboratory Southwest (REL Southwest, 2008) corroborates and expands these findings. Following the examination of nine rigorous studies, REL concluded that "teachers who receive substantial professional development—an average of 49 hours in the nine studies—boosted their students' achievement by 21 percentile points" (p. i). These data are also consistent with the findings of the National Reading Panel Report (NICHD, 2000), which indicated that professional development interventions almost always lead to greater student achievement. Importantly, professional development should never be defined as short-term workshops or talks, but rather as systematic experiences in which teachers expend considerable time and into which they are able to make a personal and intellectual investment.

Nonmonetary Costs of the Assessment Program

The nonmonetary costs in a large assessment program revolve around both teaching staff and students. The teaching staff must be prepared to *buy into* the concept of assessment in general and to the measures chosen. The process of buying into any policy in any organization can be time consuming and frustrating. Several disparate responses characterize the reactions of the teaching staff to the genesis of the large-scale assessment program described in this chapter. First, instructors claimed that assessment meant *Big Brother was watching* and that instruction and teaching style should remain private. They claimed that assessment was too invasive. Next, teachers asserted their experience as sufficient documentation for students' performance as well as their own. Having spent more than 20 years in the workplace was deemed sufficient to document performance. Grades were considered to be adequate judgments of student language abilities. Third, teachers did not like the thought of having their students sit for examinations that they (their teachers) had not written and for which they could not directly prepare their students. Even staff who were fully certified as oral proficiency testers did not seem to grasp the concept of proficiency—a concept of what any given student can do with a language regardless of curriculum and training (Bernhardt, 1997, 2000; Bernhardt et al., 2009).

Psychological cost cannot be overestimated. It became clear throughout the move toward a systematic assessment program that the teaching staff had to be on the same page regarding the scoring schemes and scales for any assessment. This involved an extensive individual investment of person hours as well as professional training. The teaching staff also had to accept the concept of external assessment. This situation led to an environment that was seen as quite threatening by instructors who were insecure regarding their own and their students' performance. Many instructors spent months, if not years, trying to overcome anxieties—anxieties that perhaps still exist a decade later.[9]

[9] A full description of teacher reactions both to centralization and to restructuring is found in Bernhardt (2000).

Students must also buy in to an assessment program. Students are overstressed and exhausted by the many tests that they have been subjected to in secondary and postsecondary years. This perception of stress, fueled by the popular press that is often ill-disposed toward tests and measurement, can lead to the modern undergraduate attitude of "blowing this off" (Hancock, 2001). This attitude implies that the assessment might not accurately reflect what students are actually able to do, or it might also unduly stress test takers. This is, of course, an important cost to consider. It is our belief that students must be convinced that any assessment program is really about improving their performance as well as that of their peers.

Conclusion

This chapter provides financial information on each area of assessment within the Language Center's programs while trying to preserve the critical sense of interconnectedness between and among the components of the program. Each entity relies on the other components to bring about success. Probably the most significant interconnected notion in the assessment program outlined in the previous pages is the *who* and the *what* of it. Determining objectives (the *what)* and having agreement among staff (the *who*) on these objectives is the most challenging part of the process. If determining community objectives is not done carefully and critically, no amount of funding will matter. As the chapter indicates, at least five years were spent in determining and articulating what is to be taught in language programs within the Language Center. Admittedly, educational goal setting should not take five years, but we learned that recognition of professional knowledge, self-esteem, and an attitude about student performance must be considered and respected. This process takes time. Sitting in groups discussing goals and aspirations for students and *making some decisions* about that is an enormous and often threatening challenge. It is enormous because one decision often changes the landscape of what entire programs are about; it is threatening because of the realization of the coming workload and facing the possible finding that students are not meeting rigorous learning standards. The *when* of assessment is also an important dimension. Placement is critical in order to carry out an efficacious curriculum. Maximal heterogeneity within individual courses leads to chaotic instruction that cannot facilitate the meeting of rigorous learning goals for all. Hence, developing effective assessment strategies for placement is paramount in providing sufficient knowledge of individual performances that enables instructors to help students advance in their foreign language proficiency. This chapter indicates that an ongoing investment in placement testing and in hardware for both Web-based testing and on-campus testing is extremely important. Formative assessments and the associated costs are by and large the cost of doing business, and the *when* of assessment happens each day and practically moment by moment. Assessment also happens in terms of assignments and activities, and hence financial resources in support of the language lab are integral to the success of the formative assessment strategies outlined here. Furthermore, exit assessments that indicate the proficiency levels achieved at the end of a language-learning sequence

are perhaps the most essential items on this list. These are the assessments that denote the success level of language programs that is so important for institutions and for individual students.

If the *who/what* dyad of assessment is the most significant key, then the *how* of assessment is more than likely the more complicated issue. The *how* must flow from the *what*. The more complicated the *what*, whichever way stated—as goals in listening, speaking, reading, and writing or as performance indicators in interpretive, interpersonal, and presentational language modes—the more complex the assessment needs. The construction of assessments is a complicated matter and the delivery of them likewise, and the hidden cost within assessment construction is professional development. A highly knowledgeable and skillful teaching force that has training in both development and rating is a nonnegotiable prerequisite for the success of any systematic assessment program and the Stanford Language Center expends the overwhelming majority of its discretionary funds on teacher professional development.

In the earlier pages of this chapter, Adams was quoted in posing the serious challenge of whether trying to match public expectations with student performance is a worthy goal. Adams used the word "tough." Examining what was implemented within the assessment cycle at the Language Center renders the word *tough* as an understatement. Yet the case of the Stanford Language Center provides a convincing answer to the question, *Is there significant value in establishing a systemic assessment program?* Internally, the university has recognized the language programs within public rhetoric and, perhaps more importantly, through salary increases for the teaching staff. Though teachers are never paid enough, salary increases as well as some significant bumps for some staff members indicate that the university does recognize the importance of the accomplishments of the language programs. Beyond salary increases, the university continually supports the renewal of technology—computer replacements for individuals and for the language lab—as well as increased funds for professional development. Having the lead role in the university's accreditation system means that, for the foreseeable future, the language programs are jewels in the crown of the university.

Undoubtedly, the most convincing reason for investing in and maintaining a consistent and coherent assessment program is that the program has helped to push students to high-proficiency gains in the productive language skills of speaking and writing. Without the explicit articulation of goals, students would be unsure of what it is they are to learn and how they are to achieve that; without an infusion of technology, teachers could not offer students the formative experiences with the language that contemporary learners need; and without a substantial and influential investment in professional development, students' instructors would not all share a sophisticated understanding of what higher registers in speaking and writing are and how to bring students toward them.

This chapter documents the creation and development of a sustainable infrastructure for establishing high-level foreign language performance at the university level. The infrastructure illustrated is in large part rooted in the important questions of *what students are able to do with what they learn in foreign language classrooms* and *whether that learning is of a high quality*. Only rigorous

assessment based in consistent and accessible standards and metrics can answer these questions with research-based evidence. Modern universities run on evidence, rather than on emotion and hope. In following an assessment orientation, the university diverges from the path of claiming excellence without evidence. It concretizes what students can do and, consequently, provides confidence and esteem within a university setting.

References

Adams, J. E., Jr. (2010). Smart money and America's schools. In J. E. Adams, Jr. (Ed.), *Smart money: Using educational resources to accomplish ambitious learning goals* (pp. 1–26). Cambridge, MA: Harvard University Press.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago: University of Chicago Press.

Bernhardt, E. B. (1997). Victim narratives or victimizing narratives? Discussions of the reinvention of language departments and language programs. *ADFL Bulletin, 29*(1), 13–19.

Bernhardt, E. B. (2000). The professional development of highly experienced and less experienced teachers: Meeting diverse needs. In B. Rifkin (Ed.), *Mentoring foreign language teaching assistants, lecturers and adjunct faculty* (pp. 41–54). Boston: Heinle & Heinle.

Bernhardt, E. B. (2006). Student learning outcomes as professional development and public relations. *Modern Language Journal, 90*(2), 588–590.

Bernhardt, E. B., Rivera, R. J., & Kamil, M. L. (2004). The practicality and efficiency of web-based placement testing for college-level programs. *Foreign Language Annals, 37,* 356–366.

Bernhardt, E. B., Valdes, G., & Miano, A. (2009). A chronicle of standards-based curricular reform in a research university. In V. Scott (Ed.), *Principles and practices of the standards in college foreign language education* (pp. 54–85). Boston: Heinle & Heinle.

Breiner-Sanders, K., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines: Speaking, revised 1999. *Foreign Language Annals, 33,* 13–18.

Breiner-Sanders, S. E., & Terry, R. (2002). ACTFL proficiency guidelines—writing revised 2001. *Foreign Language Annals, 35,* 9–15.

Chen, Y., & Kristjansson, K. (2011). Private feelings, public expressions: Professional jealousy and the moral practice of teaching. *Journal of Moral Education, 40*(3), 349–358.

Hancock, D. R. (2001). Effects of test anxiety and evaluative threat on students' achievement and motivation. *The Journal of Educational Research, 94*(5), 284–290.

Johnstone, D. B. (1998). *The financing and management of higher education: A status report on worldwide reforms*. Retrieved July 27, 2013, from www.worldbank.org/html/extdr/educ/postbasc.html

Joyce, B., & Showers, B. (1988). *Student achievement through staff development*. New York: Longman.

Levenson, N. (2012). *Smarter budgets, smarter schools: How to survive and thrive in hard times*. Cambridge: Harvard University Press.

Malone, M. (2000). *Simulated oral proficiency interviews: Recent developments*. Online resource digest. Retrieved July 10, 2013, from www.cal.org/resources/digest/0014sumulated.html

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for*

reading instruction: Reports of the subgroups (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.

National Standards in Foreign Language Education Project. (1996). *Standards for foreign language learning: Preparing for the 21st century*. Yonkers, NY: American Council on the Teaching of Foreign Languages.

National Standards in Foreign Language Education Project. (2006). *Standards for foreign language learning: Preparing for the 21st century*. Yonkers, NY: American Council on the Teaching of Foreign Languages.

Ramirez, A. (2001). How merit pay undermines education. *Educational Leadership, 58*(5), 16–20.

REL Southwest. (2008). *Reviewing the evidence on how teacher professional development affects student achievement*. Retrieved February 15, 2011, from http://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_2007003.pdf

WASC Visiting Team Capacity and Preparatory Review. (2010). *Report*. Retrieved August, 2012, from wasc.Stanford.edu

University Board of Trustees. (2012). *The study of undergraduate education at Stanford University*. Stanford, CA: Stanford University Board of Trustees.