Foreign Language Annals
Published by the American Council on the Teaching of Foreign Languages

Converting Data to Knowledge: One District's Experience Using Large-Scale Proficiency Assessment

Kristin J. Davin
Loyola University Chicago

Tania A. Rempert
Loyola University Chicago

Amy A. Hammerand
Murray Language Academy

Abstract: *The present study reports data from a large-scale foreign language proficiency assessment to explore trends across a large urban school district. These data were used in conjunction with data from teacher and student questionnaires to make recommendations for foreign language programs across the district. This evaluation process resulted in recommendations related to a need for consistency in curriculum and assessment, program articulation, and responsive placement strategies, as well as the need for greater emphasis in interpretive and interpersonal communication.*

Key words: *assessment, outcomes, proficiency, program evaluation, STAMP*

Particularly during the last decade, foreign language teachers have been largely uninvolved in the high-stakes testing movement initiated by such legislation as No Child Left Behind (NCLB, 2002) that has linked government funding to yearly test scores in math and language arts. However, even if not legislatively mandated, systematic documentation of student achievement is essential to the success and longevity of programs in all academic disciplines, including foreign languages (Donato & Tucker, 2010). In the United States, districts and individual programs or schools that implement large-scale foreign language assessments do so by choice, and often at great expense. Data from these assessments can be

Kristin J. Davin (PhD, University of Pittsburgh) is Assistant Professor of Foreign Language Education, Loyola University Chicago.

Tania A. Rempert (PhD, University of Illinois at Champaign-Urbana) is Evaluator, Loyola University Chicago.

Amy A. Hammerand (MEd, University of Illinois at Chicago; MA, University of Florida) is a Spanish teacher at Murray Language Academy, Chicago, IL.

Foreign Language Annals, Vol. 47, Iss. 2, pp. 241–260.
DOI: 10.1111/flan.12081

used for a myriad of purposes, such as monitoring the progress of individual students or groups of students, placing students into appropriate levels of instruction, diagnosing learning difficulties, allocating resources, and evaluating program effectiveness (Brindley, 2001; Newton, 2007). When the goal is to place students into the appropriate level of instruction, proficiency data may be all that are needed. However, when the goal is to evaluate the effectiveness of a program, scores from a proficiency assessment alone are not sufficient evidence upon which to make decisions. Rather, proficiency data serve as only one source of evidence that must be triangulated and contextualized to determine the strengths and weaknesses of a foreign langage program as a whole. Additional evidence may include information about students' cognitive development, general background knowledge, type and duration of prior learning experiences in the language, or practitioners' direct local knowledge and experiences (Datnow, Hubbard, & Mehan, 2002).

The present study reports results from a large-scale foreign language proficiency assessment and examines how proficiency scores were used in conjunction with other sources of data to inform programmatic decisions in one of the largest urban public school districts in the United States. Unlike research focusing on program evaluation at the university level (Norris, Davis, Sinicrope, & Watanabe, 2009; Pfeiffer & Byrnes, 2009; Walther, 2009; Watanabe, Norris, & González-Lloret, 2009), this work describes how a program evaluation was conducted by the central office of a public school district. Thus, the purpose of the present study was twofold: (1) to present data gathered from four instruments, including a pre-assessment teacher questionnaire, a student demographic questionnaire, a foreign language proficiency assessment, and a post-assessment teacher questionnaire; and (2) to explain how those data were used to evaluate the program's effectiveness.

Background

Assessing Student Learning and Program Effectiveness

The extent of students' learning in a foreign language is most often assessed using either district-created assessments or existing instruments such as the Oral Proficiency Interview (OPI) and the ACTFL Assessment of Performance toward Proficiency in Languages published by ACTFL, the Early Language Listening and Oral Proficiency Assessment and the Student Oral Proficiency Assessment published by the Center for Applied Linguistics, and the Standards-based Measurement of Proficiency (STAMP) created by the Center for Applied Second Language Studies and distributed by Avant Assessment. Using these assessments as well as others designed independently by individual districts (Fall, Adair-Hauck, & Glisan, 2007), a growing number of researchers have examined foreign language proficiency scores of high school students and correlated them with variables such as years of study or the age of the student when study first began (Fall et al., 2007; Glisan & Foltz, 1998; Huebner & Jensen, 1992). Huebner and Jensen (1992) examined the results of oral proficiency testing in a high school foreign language program in a small school district. They found that the majority of students enrolled in Level II of Spanish, French, and German scored between Novice Mid and Novice High on the OPI, while the majority of the students enrolled in Level III scored between Intermediate Low and Intermediate Mid. At Level IV, scores generally fell between Intermediate Low and Intermediate High. Scores at each level of language study typically ranged across at least four proficiency sublevels, suggesting that students enrolled in the same level of study varied tremendously in terms of proficiency. The work of Glisan and Foltz (1998) corroborated these findings. Based on a sample of students enrolled in Spanish at two high schools, they also found that students enrolled in Spanish II scored between Novice Mid and Novice High on the OPI. However,

the majority of students enrolled in Spanish IV scored lower than those reported by Huebner and Jensen (1992), only approaching the Intermediate Low level. A third study analyzed proficiency scores for more than 6,000 students in a large urban school district. The results found by Fall and her colleagues (2007) on a large-scale test of oral proficiency created by the Pittsburgh Public Schools were similar to the findings of Huebner and Jensen (1992) and Glisan and Foltz (1998). They, too, concluded that most students required sequential, well-articulated, five-period-per-week programs for at least four years to reach an Intermediate Low level of proficiency. Each of these studies revealed a wide variation in scores within each level and clearly demonstrated that any two students who have followed the same program of study for the same amount of time may reach different levels of proficiency, supporting previous findings that learners progress at varying rates, regardless of course boundaries (Magnan, 1986).

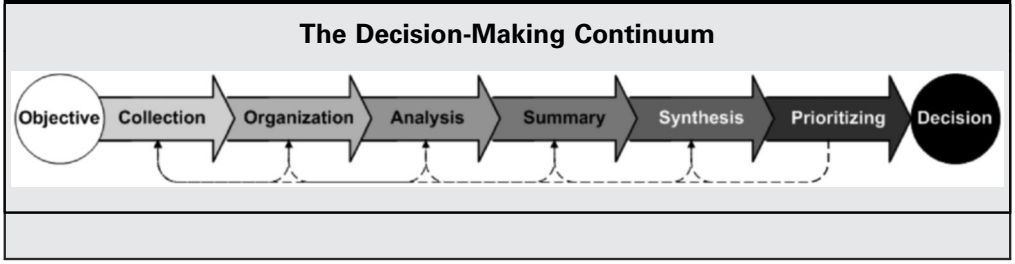
At the time of the present study, STAMP measured proficiency in reading, writing, and speaking, unlike the studies described above, which assessed only speaking. In 2010, Avant Assessment released a report detailing nationally aggregated STAMP results from 2008 to 2009. The sample consisted of 29,000 test takers in grades 7–16 from schools across the United States, as well as from international schools around the world. The aggregated data showed that students studying European languages (French, German, Italian, and Spanish) generally required at least four years of study to reach the Intermediate Low level in speaking. Three or four years of target language study were required to reach the Intermediate Low level in writing, and students did not generally reach the Intermediate Low level of proficiency in reading during a four-year sequence. However, variation in student performance existed across all three domains, with some students reaching higher levels of proficiency more quickly than their peers (Avant Assessment, 2010).

Data-Based Decision Making

While each of the studies cited above provided essential information about students' developing proficiency across levels of instruction, they did not address how those data were used. To understand how data informed program decisions, the present study drew upon the existing literature on data use (Ackoff, 1989; Mandinach, Honey, & Light, 2006; Mandinach, Honey, Light, & Brunner, 2008; Marsh, 2012; Marsh, Pane, & Hamilton, 2006). Within this framework, data, information, and knowledge form a continuum (Ackoff, 1989) in which raw data have no meaning in and of themselves. Initially, one must collect relevant data and organize them in a systematic way. Once these data are analyzed and summarized, they become information, which must be synthesized and prioritized in order to be transformed into knowledge. Once this conversion has taken place, stakeholders must prioritize knowledge to make decisions (Mandinach et al., 2008). At any point along the data–information–knowledge continuum, one can return to an earlier stage to collect additional data or try different forms of analyses, thus leading to feedback loops within the process (Marsh, 2012). This process is represented in Figure 1. However, at any point along the continuum, breakdown may occur and data can fail to become information or knowledge (Mandinach et al., 2008).

The concept of "sense-making" or "interpretation" (Spillane, Reiser, & Reimer, 2002) describes how raw data are converted to useable information to inform decisions about future programming. Ways in which data are organized, analyzed, summarized, synthesized, and prioritized depend on the interaction among the varying actors and their existing knowledge (Burch & Spillane, 2005; Coburn & Talbert, 2006; Marsh et al., 2006; Spillane, 2012). In addition, the organizational structure of a school district, a lack of sufficient resources, or a change in leadership can influence how data are interpreted (Coburn, Toure, & Yamashita, 2009).

FIGURE 1



Program Evaluation

Program evaluation refers to "the systematic collection of information about the activities, characteristics, and outcomes of programs to make judgments about the program, improve program effectiveness, and/or inform decisions about future programming" (Patton, 1997, p. 23). According to the evaluation standards published by the Joint Committee on Standards for Educational Evaluation (1994), evaluations should meet four criteria: utility (usefulness to a predetermined audience), feasibility (practicality and cost-effectiveness), propriety (fairness and ethical methods and actions), and accuracy (usefulness and accuracy of the information) (Patton, 1997).

Much of the research on foreign language program evaluation has taken place at the college level. Several case studies in particular, supported by the Foreign Language Program Evaluation Project, illustrated how evaluation informed and supported foreign language programs (Norris et al., 2009). These evaluations examined questions about the types of degrees universities should offer in foreign languages (Loewensen & Gómez, 2009), how to train graduate teaching assistants appropriately (Zannirato & Sánchez-Serrano, 2009), the extent to which study abroad experiences facilitated language and cultural development (Ramsay, 2009), and students' impressions of the language program (Pfeiffer & Byrnes, 2009). Walther (2009) specifically investigated students' learning by course and level using the Simulated Oral Proficiency Interview to guide course improvements, comparing students' performance with the performance of students at other universities. Because the evaluation process was time-consuming and required substantial resources, Walther (2009) noted that, in program evaluation, "goals and expectations must be carefully and realistically measured against available financial, institutional, and above all human resources" (p. 132).

A second body of research has focused on the processes involved in program evaluation. Elder (2009) investigated the processes of evaluation and relationships between evaluators and program participants within three elementary or secondary bilingual programs in Australia. She found that four factors influenced the extent to which evaluations led to program improvement: (1) the relationship between the evaluator and the stakeholders, (2) the degree of stakeholder understanding of the evaluation process, (3) the difference in cultural or linguistic affiliation of the evaluator and program insiders, and (4) the role of communication between the evaluator and program insiders. Harris (2009) examined two case studies focusing on the processes that took place between the initial evaluation findings and the final evaluation report drawn from a series of national-level evaluations of Irish language education programs, a minority language that the Republic of Ireland has been working to revitalize for 85 years. Harris (2009) highlighted the findings that were challenging or misunderstood and described how further statistical analysis and stronger contextualization of the data better shaped data users' understandings and the ensuing use of the evaluation.

The Present Study

Unlike the studies described above, the present study highlights data use at the district level. School district central offices can and should play a central role in instructional improvement (Coburn et al., 2009; Elmore & Burney, 1997; Hightower, Knapp, Marsh, & McLaughlin, 2002; Knapp, Copland, & Talbert, 2003), and this involves using data in meaningful ways to accomplish specific goals. Data have the potential to shine "a clear unambiguous light on how to strengthen school performance or at least where districts should direct their efforts" (Honig & Coburn, 2007, p. 582). To explore how one school district used data to strengthen the performance of foreign language programs, three research questions guided the present program evaluation:

1. To what extent are language offerings consistent across schools in this district?
2. Do students have the opportunity to continue to study the same language in articulated sequences?
3. Are foreign language programs in the district permitting students to reach Intermediate levels of proficiency in reading, writing, and speaking?

Methods

Context

In 2010, the foreign language coordinator of an urban school district that served a diverse population of more than 400,000 students, of whom 86% were classified as low-income and 12% were classified as limited English proficient (a federal designation), received funds to implement a large-scale foreign language proficiency assessment and program evaluation. Instruction across the district was offered in 12 different languages, listed in Table 1. Elementary school principals chose whether to include a foreign language program in their building, while high school principals were required to offer courses in at least one foreign language. Principals at both levels were given the freedom to choose which languages to offer.

TABLE 1

Number of Schools That Offered Each Language in 2010–2011

Foreign Languages	Elementary Schools	High Schools
Spanish	68	125
French	15	54
Chinese	25	17
Latin	4	10
Arabic	7	3
Italian	3	5
German	1	6
Japanese	4	3
Russian	0	4
Polish	0	2
Urdu	1	0
Korean	0	1

There are more than 600 schools in this district and approximately 500 foreign language teachers. The district foreign language coordinator provided year-round professional development opportunities and served as a language resource for principals and other school administrators.

Instruments and Procedures

Pre-Assessment Teacher Questionnaire

The pre-assessment teacher questionnaire was emailed to all 500 foreign language teachers in the district in February 2011 using Survey Monkey.[1] In addition to demographic data, teachers were asked about the average number of minutes allocated to instruction each week by foreign language and grade level, curriculum and materials used, methods of assessments, and impressions of the school's methods of placing students into the appropriate level of instruction. One hundred fifty-four of the 500 foreign language teachers representing 89 different schools participated, representing 30.8% of the district's foreign language teachers.

Assessment of Speaking, Reading, and Writing

During March 2011, 20 elementary schools and 30 high schools were invited to administer STAMP.[2] The intended sample for STAMP administration encompassed approximately 50 schools, 400 teachers, and 6,000 foreign language students across the district. In order to be eligible for participation in the study, teachers needed to (1) attend a training webinar, and (2) have access within their school building to a sufficient number of computers and headsets with microphones for the number of students to be tested. Of the 400 foreign language teachers who were invited, 120 volunteered to participate in STAMP, representing 20 of the 50 invited schools in the district. Of the 6,000 intended foreign language students, 3,881 students in grades 7–12 participated in the proficiency assessment, which was offered in Chinese, French, Italian, Japanese, and Spanish (see Table 2). Students represented 10 elementary schools and 10 high schools across the district.[3]

Students completed the speaking, reading, and writing portions of STAMP over a three-day period during regularly scheduled foreign language classes during April, May, and June 2011 on dates that were selected by the foreign language teacher(s) in each building. As shown in Table 3, the sample of students who participated in STAMP administration was statistically different ($p < 0.05$) from the district's population in every demographic listed, limiting the ability to draw generalizations across all students but sufficiently large to permit general conclusions to be drawn.

TABLE 2

Number of Students by Language Assessment Taken

Language	*n*	% of Total Test Takers
Spanish	2,166	56
Chinese	1,058	27
French	606	16
Italian	27	0.7
Japanese	24	0.6

Note: *Test takers* refer to those students who initiated the test but may or may not have finished each test portion (reading, writing, or speaking). *Test completers* refers to those students who finished the test portion being discussed.

Student Demographic Questionnaire

All 3,881 students who took STAMP completed the demographic questionnaire on the Internet during STAMP administration. The questionnaire asked students to respond to questions such as: "At what age did you begin studying the target language?" "How long have you studied the foreign language?" "What was your first language?" and "How often do you speak the target language with a relative?"

Post-Assessment Teacher Questionnaire

All of the teachers who participated in STAMP administration attended a professional development workshop during September 2011. At this workshop, they were asked to respond to a post-assessment questionnaire about their experiences proctoring STAMP and how they were using the resulting data. Responses were recorded for all 120 teachers.

Data Analysis

Pre-Assessment Teacher Questionnaire

Descriptive statistics were used to analyze the pre-assessment teacher questionnaire data. To determine the weekly minutes of foreign language instruction that students received in each language at each grade level, the mean, median, standard deviation, minimum, and maximum were calculated for each category. The computer program Survey Monkey calculated the number and percentage of teachers who used each type of curriculum or assessment listed in the

TABLE 3

Demographics of All District Schools Compared to Those Included in STAMP Administration

	% White	% Black	% Hispanic	% Asian	% Meets Standards	% Limited English Proficient	% Low Income
All District Schools	7*	66*	26*	0.40*	25*	6*	89*
Schools in STAMP Administration	12*	35*	46*	2*	42*	3*	81*

questionnaire, as well as each teacher's impressions of their schools' placement strategies. Because some questions were optional, not all of the 154 teachers who responded to the pre-assessment questionnaire answered every question.

STAMP Scores

Avant Assessment provided summary charts that showed the number of students who completed each portion of the test as well as the percentage of students scoring at each sublevel (i.e., Novice Low, Novice Mid, etc.) for each language in speaking, reading, and writing.[4] These summary charts represented all students in grades 7–12 who took STAMP. In addition to these summary charts, the evaluator conducted additional analyses to disaggregate the data into useable information. By correlating STAMP scores with student demographic data, scores for all students who indicated that their first language was the same one in which they took STAMP, as well as all students who indicated that they spoke the target language on a daily basis with a relative, were removed.

Student Demographic Questionnaire

Descriptive statistics were used to determine the relationship between STAMP scores in each of the three domains (reading, writing, and speaking) and the number of years that students had studied the foreign language. Because only a small number of students completed STAMP in Japanese ($n = 23$) and Italian ($n = 28$), these scores were omitted from analysis.

Post-Assessment Teacher Questionnaire

Descriptive statistics were used to analyze the post-assessment teacher questionnaire data. To determine teachers' opinions of their school's placement strategies for foreign language students, the number and percentage of respondents for each category were calculated.

Feeder School Analysis

For the feeder school analysis, the patterns of movement from elementary schools to high schools within the district were analyzed by tracking the number of students feeding from one school to the next as well as the extent to which students could complete a sustained sequence of instruction in the language(s) that were offered at the elementary level. Given that many educational options were available to students within and outside of the school system, for purposes of this analysis any elementary school that sent at least 10 students to a high school was considered a feeder school for that high school.

Results

There were four primary findings related to (1) program design, (2) curriculum and assessment, (3) placement and articulation, and (4) proficiency scores.

Program Design—Minutes of Weekly Instruction

Table 4 shows the number of teachers who responded to the pre-assessment questionnaire by language and grade of instruction, as well as the average and median number of minutes of instruction for each language and grade per week, the standard deviation, the minimum, and the maximum number of minutes taught. The number of minutes of weekly instruction varied substantially even within the same grade band and target language: A student studying French at one elementary school received a minimum of 40 minutes of weekly instruction, while a student attending a different elementary school received a maximum of 640 minutes of instruction in French. On average, students received more minutes of weekly instruction in grades 9–12 than in grades 6–8 or K–5, with students in grades 6–8 receiving the fewest minutes of instruction. In grades 9–12, students studying Chinese or French received approximately 100 to 150 more minutes of weekly instruction than students who were studying Spanish.

Curriculum and Assessment

Teachers' responses on the pre-assessment questionnaire indicated that there was no mandated curriculum or assessment in foreign language programs across the district. Seventy-four teachers (48%) reported using their own combination of activities and using no particular textbook series, 70 teachers (47%) reported using one or more published textbook series, and 10 teachers (5%) did not respond to the question. When asked to identify which textbook series they were using, Spanish teachers identified 17 series, French teachers identified 3, and Chinese teachers identified 5.

Teachers were also asked to indicate which forms of assessment they used within their classrooms. Permitted to select more than one response, 52% ($n = 105$) of respondents indicated that they created their own assessments, 30% ($n = 61$) of respondents indicated that they used the assessments from the textbook series, 12% ($n = 24$)

TABLE 4

Weekly Instruction Minutes By Language and Grade

Language	Grades	*n*	Mean	Median	SD	Min	Max
Chinese	K–5	8	603.8	560	382.2	80	1200
	6–8	7	263.4	200	212	120	724
	9–12	10	779.2	840	287.8	180	1250
French	K–5	5	293.0	225.0	224.8	40	640
	6–8	4	291.3	282.5	198.3	120	480
	9–12	16	732.9	890.0	240.4	225	960
Spanish	K–5	17	568.6	540.0	366.0	120	1500
	6–8	14	325.4	285.0	171.4	40	675
	9–12	68	620.8	595.0	595.0	135	1020

indicated that they used a standardized assessment, and 5% ($n = 11$) indicated that they did not give any assessments.

Placement and Articulation

The feeder school analysis revealed that less than half of elementary schools that offered a world language matriculated more than 10 graduating students into a high school offering the same language. In this district, most students went directly from elementary school to high school because elementary schools typically consisted of pre-kindergarten to 8th grade, unlike other other districts that had separate middle schools for grades 6–8. Only 40 of the 91 elementary schools in which instruction in a foreign language was provided sent at least 10 students from 8th grade to a high school that provided instruction in the same foreign language. Despite the variability in elementary school language programs, many teachers indicated a lack of strong placement strategies within their schools. Table 5 shows teachers' responses on the post-assessment questionnaire to the question, "What are your impressions of the quality of your school's current foreign language placement strategy?" Responses indicated dissatisfaction with schools' procedures for placing students at the correct instructional level.

Proficiency Test Results

Avant Assessment provided the district with figures displaying the distribution of all students' scores in grades 7–12 in reading, writing, and speaking. However, the district undertook additional analyses to gain a deeper understanding of whether its foreign language programs permitted students to reach Intermediate levels of proficiency. To better use the data provided by Avant Assessment, the evaluator correlated demographic data with STAMP scores to examine how many years of study were required to reach an Intermediate level of proficiency.

Reading Performance

The data displayed in Figure 2 show the distribution of reading scores across all three languages in percentages. Across all languages, the majority of students scored at the Novice Low level in reading, with an especially high percentage for Chinese at 86%. Few students reached an Intermediate level of proficiency in reading, with only 3% in Chinese, 12% in French, and 11% in Spanish scoring in the Intermediate range.

Chinese

Table 6 shows the mean, median, standard deviation (SD), minimum score, and maximum score for those who completed the

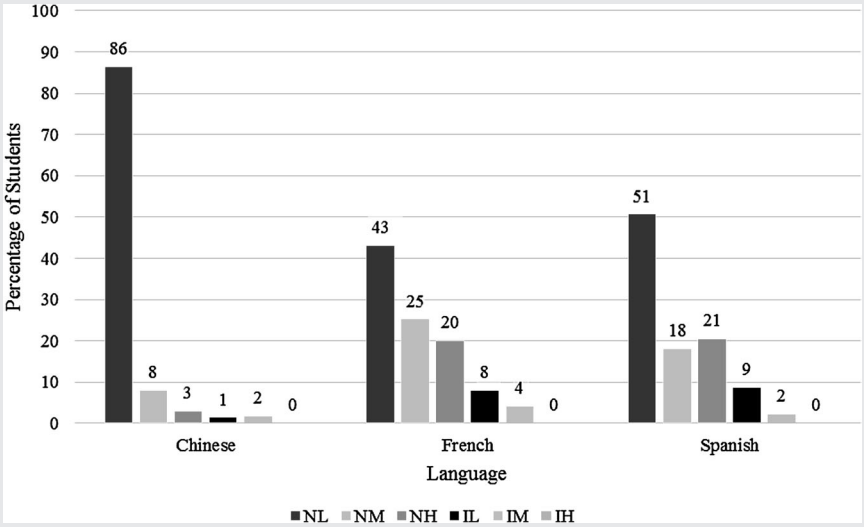
TABLE 5

Teachers' Impressions of Schools' Placement Strategies

Response	*n*	% of All Respondents
I am not sure if it is a good strategy, but I would like to know of other strategies that work.	31	25.83
We have a good placement strategy in place that works for us.	23	19.17
There isn't any strategy to speak of.	22	18.33
Not a good strategy; our students would benefit from another strategy.	25	20.83
We have a good placement strategy that I think other schools would benefit from.	3	2.50
Other	16	13.33

FIGURE 2

Distribution of Reading Scores in Percentages By Language



Note: NL = Novice Low, NM = Novice Mid, NH = Novice High, IL = Intermediate Low, IM = Intermediate Mid, IH = Intermediate High

reading portion of STAMP in Chinese. The mean generally increased with years of study but never surpassed Novice Mid (a STAMP score of 2), even with five years of study. The median student who had studied Chinese for as many as four years remained at the Novice Low level (a STAMP score of 1) but increased to Novice High with five years of study.

French

Table 7 presents student data on the reading portion of STAMP in French based on years of study. The mean and median score remained at Novice Low for students throughout their first three years of study. However, a sharp increase occurred for students who had studied French for four years, achieving a mean score of Intermediate Low (a STAMP score of 4).

Spanish

The mean score in reading for Spanish generally increased with the number of years of study (Table 8). The mean remained at Novice Low after two years of study but reached

TABLE 6

Reading Scores in Chinese Based on Years of Study

	n	Mean	Median	SD	Min	Max
Year 1	363	1.04	1	0.25	1	3
Year 2	258	1.19	1	0.56	1	5
Year 3	98	1.66	1	1.06	1	5
Year 4	24	2.04	1	1.43	1	5
Year 5	8	2.43	3	1.40	1	4

TABLE 7

Reading Scores in French Based on Years of Study

	n	Mean	Median	SD	Min	Max
Year 1	199	1.79	1	0.79	1	5
Year 2	203	1.79	1	1.01	1	5
Year 3	88	1.55	1	0.91	1	5
Year 4	33	4.00	4	0.95	2	5
Year 5	0					

Novice Mid after three years. The median student followed the same pattern, achieving the proficiency level of Novice High in reading only after three years of study. Unlike French, with four years of study Spanish students' mean score in reading still did not reach the Intermediate level of proficiency.

Writing Performance

Figure 3 shows the distribution of writing scores in percentages for all test completers in grades 7–12. The majority of students scored at the Novice High level of proficiency across all three languages. Scores in writing proficiency were generally higher than scores in reading proficiency. Thirty-seven percent of all students scored at the Intermediate level in Chinese. The percentage was slightly higher for Spanish, with 43% of students reaching an Intermediate level, and slightly lower for French, with only 25% of students reaching an Intermediate level.

Chinese

Table 9 shows the mean, median, SD, minimum score, and maximum score for those who completed the writing section of STAMP in Chinese. The mean score generally increased with the number of years of study. The mean was Novice Mid after one year, Novice High after two years, and was approaching Intermediate Low after three years. The median student scored at the Novice High level after one and two years of study and approached the Intermediate Low level after three years of study. Students in their fourth and fifth year of study reached a mean and median score of Intermediate Low in writing, with the mean approaching Intermediate Mid after five years of study.

French

Table 10 illustrates proficiency data on the French STAMP writing section based on years of study. Like Chinese, the mean and median were Novice High, or approaching Novice High, after one and two years of

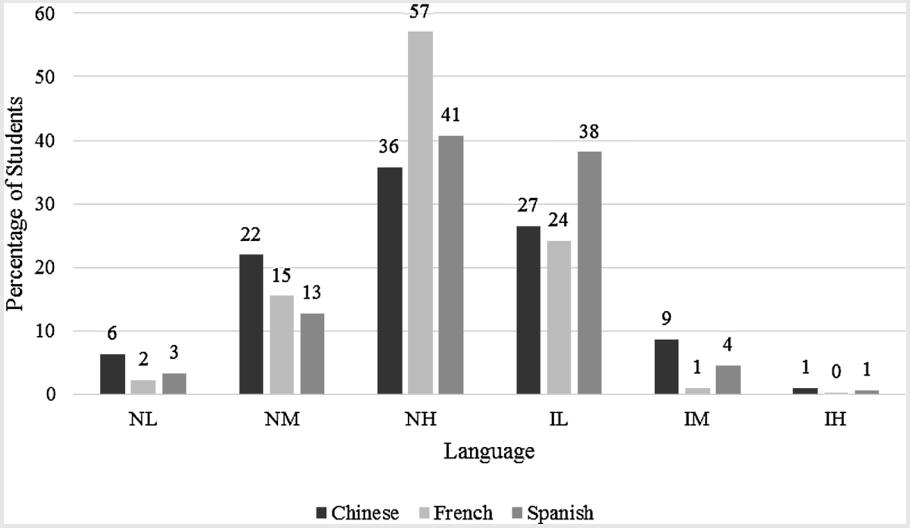
TABLE 8

Reading Scores in Spanish Based on Years of Study

	n	Mean	Median	SD	Min	Max
Year 1	671	1.50	1	0.81	1	4
Year 2	394	1.61	1	0.89	1	4
Year 3	10	3.11	3	0.93	1	4
Year 4	57	3.45	3.5	0.87	2	5
Year 5	0					

FIGURE 3

Distribution of Writing Scores in Percentages By Language



Note: NL = Novice Low, NM = Novice Mid, NH = Novice High, IL = Intermediate Low, IM = Intermediate Mid, IH = Intermediate High

study. The mean and median scores for students who had studied French for three or four years were Intermediate Low.

Spanish

Writing scores in Spanish generally followed the same patterns as those of French (Table 11). The mean after one and two years of study was approaching Novice High, but the median student reached Novice High after only one year. While the median student scored at the Intermediate Low level of proficiency after three years of study, four years were required to reach a mean score of Intermediate Low.

Speaking Performance

Figure 4 shows the distribution of speaking scores in percentages for all students in grades 7–12 who took STAMP. The majority of students scored at the Novice High level of proficiency across all three languages. The patterns in speaking scores were similar to the patterns in writing scores. Twenty-five percent of students studying Chinese, 16%

TABLE 9

Writing Scores in Chinese Based on Years of Study

	n	Mean	Median	SD	Min	Max
Year 1	262	2.60	3	0.90	1	5
Year 2	217	3.01	3	1.05	1	5
Year 3	94	3.88	4	0.91	1	6
Year 4	19	4.17	4	0.71	3	6
Year 5	8	4.86	4	1.07	4	6

TABLE 10

Writing Scores in French Based on Years of Study

	n	Mean	Median	SD	Min	Max
Year 1	137	2.76	3	0.66	1	4
Year 2	152	2.99	3	0.58	1	4
Year 3	88	3.89	4	0.85	1	6
Year 4	33	4.06	4	0.56	3	6
Year 5	0					

of students studying French, and 29% of students studying Spanish reached an Intermediate level of proficiency in speaking.

Chinese

Table 12 shows the mean, median, SD, minimum score, and maximum score on the Chinese STAMP speaking section. The median student reached a score of Intermediate Low after three years of study, but five years were required for the mean to reach this level. The median student achieved a score of Intermediate Mid (represented by a 5) after five years of study.

French

Table 13 shows student data on the French STAMP speaking section based on years of study. After one and two years of study, the mean remained at Novice Mid, while the median student achieved a score of Novice High after both one and two years of study. After both three and four years of study, the mean was Novice High, while the median reached Intermediate Low.

Spanish

Table 14 shows the mean, median, SD, minimum score, and maximum score for the Spanish STAMP speaking section. Similar to French, the mean after one and two years of study was Novice Mid. The mean increased to Novice High after three years of study and was approaching Intermediate Low after four years of study. Similar to Chinese and French, the median student reached the Intermediate Low level of proficiency after three years of study.

Discussion

Results of this program evaluation shaped four district recommendations.

Curriculum, Assessment, and Time Allocation

Results of the pre-assessment teacher questionnaire and STAMP data revealed a lack of consistency in curriculum goals, assessments, and time allocation across the

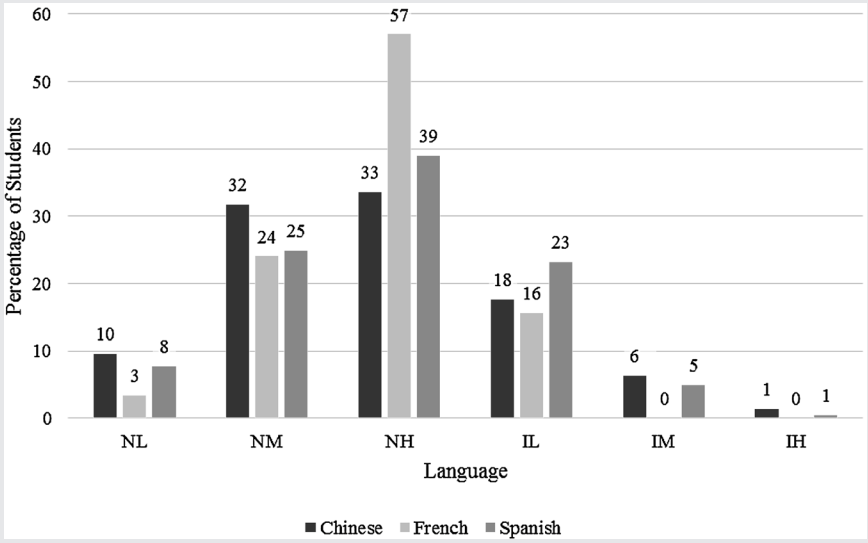
TABLE 11

Writing Scores in Spanish Based on Years of Study

	n	Mean	Median	SD	Min	Max
Year 1	481	2.85	3	0.80	1	5
Year 2	280	2.92	3	0.81	1	5
Year 3	8	3.71	4	0.76	2	4
Year 4	52	4.08	4	0.59	4	6
Year 5	0					

FIGURE 4

Distribution of Speaking Scores in Percentages By Language



Note: NL = Novice Low, NM = Novice Mid, NH = Novice High, IL = Intermediate Low, IM = Intermediate Mid, IH = Intermediate High

TABLE 12

Speaking Scores in Chinese Based on Years of Study

	n	Mean	Median	SD	Min	Max
Year 1	258	2.21	2	0.94	1	5
Year 2	193	2.78	3	0.95	1	6
Year 3	86	3.59	4	0.88	2	6
Year 4	18	3.71	4	0.77	3	5
Year 5	8	4.29	5	0.95	3	5

TABLE 13

Speaking Scores in French Based on Years of Study

	n	Mean	Median	SD	Min	Max
Year 1	109	2.54	3	0.66	1	4
Year 2	94	2.66	3	0.63	1	4
Year 3	88	3.60	4	0.88	2	6
Year 4	29	3.71	4	0.53	2	4
Year 5	0					

TABLE 14

Speaking Scores in Spanish Based on Years of Study

	n	Mean	Median	SD	Min	Max
Year 1	431	2.41	2	0.81	1	4
Year 2	209	2.59	3	0.85	1	4
Year 3	6	3.20	4	1.30	1	4
Year 4	45	3.84	4	0.78	3	6
Year 5	0					

district. Even within this subset of the district's elementary and high schools, data showed that teachers used a variety of textbook series and assessments. Furthermore, the wide variation in the amount of time that was allocated to foreign language programs at different schools resulted in extremely uneven foreign language experiences for students across the district. Because families were, to a certain extent, allowed to choose to which elementary and high schools send their children rather than follow the pre-established feeder system, instituting common curricula and assessments was an essential first step in unifying programs and establishing equivalent learning opportunities for all students.

Sequences of Study Within a Feeder Stream

The feeder school analysis, student demographic data, and STAMP scores provided strong evidence that students needed more opportunities to participate in extended and well-articulated sequences of study. Results suggested that most students did not complete four years of study in any foreign language. This is a particularly important finding, as data show that four years of study seem to be required to reach an Intermediate level of proficiency. At the time of the study, the city's mayor and the school board were considering an increase in the length of the school day, which might make additional time available for foreign language study. Data clearly showed that most schools needed to allocate more minutes of weekly instruction to foreign language learning and that extended sequences of instruction needed to be established in both elementary and high schools. These steps would allow the majority of students who followed the established feeder patterns access to the same languages(s) throughout their K–12 experience. Although the district's current policy allows school principals to choose languages that are offered in their building, the data strongly support the need for increased collaboration between feeder schools as well as a policy that requires feeder elementary and high schools to select a more limited set of languages in which longer sequences of study could be provided.

Responsive Placement Strategies

Based on STAMP scores, student demographic data, and the post-assessment questionnaire, it was recommended that schools develop more responsive foreign language placement strategies. In most schools, students advanced to the next level of foreign language instruction based on a passing letter grade in the course. Because curricula and methods of assessment varied so dramatically across the district, grades did not necessarily reflect comparable progress toward proficiency. The large variation in STAMP scores within a particular year of study also confirmed that years of previous instruction was not the most effective variable to use when placing students. This finding is supported by previous research (Glisan & Foltz, 1998; Huebner

& Jensen, 1992) showing that years of study is not a strong predictor of language proficiency and that learner characteristics have a profound impact on language proficiency (Scarino et al., 2011).

Proficiency Scores

Reading

STAMP scores revealed that students struggled to meet ACTFL standard 1.2, "Learners understand, interpret, and analyze what is heard, read, or viewed on a variety of topics" (ACTFL, 2014, p. 1) and confirmed the need for literacy instruction across languages. Across all languages, the majority of test completers in grades 7–12 scored at the Novice Low level in reading, and fewer than 15% of students reached the Intermediate Low level of reading across all three languages tested. Although existing literature has confirmed that scores in reading often lag behind those in speaking and writing (Avant Assessment, 2010) and that four years of sequential study are generally not sufficient for a student to reach Intermediate levels of proficiency, students' reading proficiency scores in the present study were of particular concern because the median student studying Chinese or French remained at the Novice Low level of proficiency even after three years of study. The median student studying Spanish did not advance past the Novice Low level of proficiency in reading until completing three years of study.

While these data were less surprising for Chinese, a character-based language, they suggest an important area for attention in Spanish and French programs. At the time of the study, the school district was preparing for the implementation of the Common Core State Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010), which require all content areas to include a focus on the literacy skills that are needed for college and career readiness. Following the evaluation, foreign language teachers were invited to attend professional development sessions on how to address the Common Core State Standards in their classrooms. The combination of STAMP scores, a knowledge of existing research, and the political climate of accountability shaped the recommendation that teachers incorporate more literacy instruction in foreign language classrooms.

Writing

Scores on the writing portion of STAMP were generally the highest across all three domains for all three languages tested. In writing, the correlation of STAMP scores and student demographic data showed that four years of foreign language study were required to reach a mean score of Intermediate Low across all three languages. These data suggested that the student population tested in this district was developing writing proficiency more quickly than the national averages reported in Avant Assessment's aggregated data from 2008–2009 (Avant Assessment, 2010). As a result, no additional district-wide recommendations centered upon writing proficiency. However, this result highlighted the need for a district-wide philosophy, outlining the balance of desired outcomes for student proficiency across all communicative modes (interpersonal, interpretive, and presentational) and questioning what data suggest is a primary emphasis on interpersonal or presentational writing.

Speaking

Supporting existing research, students who had studied the target language for two years achieved a mean score between Novice Mid and Novice High (Glisan & Foltz, 1998; Huebner & Jensen, 1992). Students who had studied the target language for three years scored lower in speaking than the averages reported by Huebner and Jensen (1992), but their scores were similar to those reported by Glisan and Foltz (1998). Results showed that the median student studying Chinese, French, or Spanish reached the Intermediate Low level of proficiency after three years of study. However, across all three languages tested, unlike the national

averages compiled by Avant Assessment (2010), the mean did not reach the level of Intermediate Low even after four years of study. In addition, supporting existing research, scores typically ranged across four proficiency sublevels for all years of study, suggesting wide variation in student proficiency in speaking (Fall et al., 2007; Glisan & Foltz, 1998; Huebner & Jensen, 1992). This finding provides further support for the need for extended, well-articulated sequences of study across the district and suggests that increased amounts of time should be allocated to interpersonal communication (ACTFL standard 1.1) at all levels and in all languages across the district.

Overall, data show that foreign language programs were not meeting the stated district-wide goal of graduating students at the Intermediate Low level of proficiency in Chinese, French, or Spanish. Few students reached the Intermediate Low level of proficiency across the three languages in any domain. Because this district only required two years of foreign language study for graduation, many of those students who indicated studying a language for three or four years likely did so by choice, which may have positively impacted scores.

Conclusions

Multiple sources of data were required in the present evaluation to understand the characteristics of existing programs and the progress toward proficiency of a subset of the district's foreign language learners. Confirming existing research from the broader field (Honig & Coburn, 2007; Kowalski & Lasley, 2009; Little, 2012), the process of converting raw data into useable information required more evidence than just the distribution of proficiency scores provided by Avant Assessment. Each of the recommendations for program development that resulted from this program evaluation was developed using a variety of evidence, supporting Norris's (2009) statement that "Multiple methodologies of data collection are essential for capturing not only the measurable outcomes of language teaching but also the value of those outcomes from distinct perspectives and the variety of factors that contribute to or constrain their achievement" (p. 11). Clearly, evidence must be drawn from a variety of sources when conducting a program evaluation (Elder, 2009; Norris, 2009; Norris & Watanabe, 2007; Patton, 1997), and, as illustrated in this study, considering proficiency data alone (Figures 2, 3, and 4) without the added information on learner variables, enrollment patterns, curriculum, assessment, and allocated time would have resulted in less robust, and substantially different, findings.

While the generalizations that can be drawn from these data may be limited due to the sample size, the study has important implications for school districts wishing to implement a large-scale proficiency assessment. First, assessing a foreign language program, and interpreting the resulting data in meaningful ways, particularly in a large school district in which building-level control is valued, is quite complex and requires substantial time, funding, knowledge of existing research, and data use skills. It is not enough to simply administer a foreign language proficiency assessment; districts must ensure that a team is in place to oversee the collection, organization, and analyses of the data. If an external person is hired to oversee the process, he or she must be able to work closely with a member of the district staff who has intimate knowledge of the district and its resources. Second, the process of converting raw data to useable information requires substantial time and resources. In this case study, the cost of the foreign language proficiency assessment itself, the required technology updates, and the evaluator to assist was significant. However, none of these expenses could have been omitted without significantly compromising the process and the results. Third, foreign language programs and coordinators must have the support of higher-level administrators within the district, including the superintendent, school board members, and program coordinators, who

together have both a high level of commitment to the project as well as the power to act on the recommendations (Donato & Tucker, 2010).

In this era of assessment and accountability, it will become increasingly important for foreign language programs to gather systematic documentation of student achievement. Proficiency assessments, when conducted along with other forms of data collection, are a powerful tool that can shed light on the strengths and needs of foreign language programs. Through program evaluation and processes of data-based decision making, teachers, administrators, foreign language coordinators, and other stakeholders can convert data to knowledge to improve foreign language programs.

Acknowledgments

We wish to thank Tavis Jules, Sabina Neugebauer, and Lara Smetana for their support and numerous insights on this manuscript. We also acknowledge Anne Nerenz, Elvira Swender, and the anonymous reviewers for their feedback and guidance.

Notes

1. Conducting an evaluation with limited resources—a common hurdle in foreign language programs in the United States—prevented certain forms of data collection such as observations, interviews, and focus groups common to other, more funded evaluations. Due to limited resources for foreign language within the district, other qualitative methods of data collection did not meet the evaluation criteria of feasibility (Joint Committee on Standards for Educational Evaluation, 1994).
2. See Wiley online appendix for more information on STAMP.
3. Of the 10 elementary schools included in the sample, two were K–8 schools, one was grades 6–8, and seven were pre-K–8. Of the 10 high schools included in the sample, six were grades 9–12 and four were grades 7–12.
4. The STAMP literature states that STAMP levels are "related to" (according to the STAMP Web site, http://www.avantassessment.com/sites/default/files/STAMP4S%20Reporting%20Guide.pdf) and "defined by" (according to the STAMP Student Guide, at http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012) the proficiency levels and sublevels that are described in the ACTFL Proficiency Guidelines (e.g., Novice High, Intermediate Mid). Some STAMP results are reported here using those designations; however, they are not equivalent to official ACTFL ratings.

References

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*, 3–9.

ACTFL. (2014). *World-readiness standards for learning languages*. Yonkers, NY: Author.

Avant Assessment. (2010). *STAMP results: National averages 2008-2009* (pp. 1–10). Retrieved March 18, 2014, from http://www.avantassessment.com

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing*, *18*, 393–407. doi:10.1177/026553220101800405.

Burch, P., & Spillane, J. P. (2005). How subjects matter in district office practice: Instructionally relevant policy in urban school district redesign. *Journal of Educational Change*, *6*, 51–76. doi:10.1007/s10833-004-7781-5.

Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education*, *112*, 469–496.

Coburn, C. E., Toure, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision making at the district central office. *Teachers College Record*, *111*, 1115–1161.

Datnow, A., Hubbard, L., & Mehan, H. (2002). *Extending educational reform: From one school to many*. London: Routledge Falmer.

Donato, R., & Tucker, G. R. (2010). *A tale of two schools: Developing sustainable early*

foreign language programs. Clevedon, UK: Multilingual Matters.

Elder, C. (2009). Reconciling accountability and development needs in heritage language education: A communication challenge for the evaluation consultant. *Language Teaching Research*, *13*, 15–33. doi:10.1177/1362168808095521.

Elmore, R., & Burney, D. (1997). *School variation and systemic instructional improvement in Community School District #2, New York City*. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Fall, T., Adair-Hauck, B., & Glisan, E. (2007). Assessing students' oral proficiency: A case for online testing. *Foreign Language Annals*, *40*, 377–406.

Glisan, E., & Foltz, D. (1998). Assessing students' oral proficiency in an outcome-based curriculum: Student performance and teacher intuitions. *Modern Language Journal*, *82*, 1–18.

Harris, J. (2009). Late-stage refocusing of Irish-language programme evaluation: Maximizing the potential for productive debate and remediation. *Language Teaching Research*, *13*, 55–76. doi:10.1177/1362168808095523.

Hightower, A., Knapp, M. S., Marsh, J. A., & McLaughlin, M. W. (2002). *School districts and institutional renewal*. New York: Teachers College Press.

Honig, M. I., & Coburn, C. (2007). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational Policy*, *22*, 578–608. doi:10.1177/0895904807307067.

Huebner, T., & Jensen, A. (1992). A study of foreign language proficiency-based testing in secondary schools. *Foreign Language Annals*, *25*, 105–115.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2nd ed.). Thousand Oaks, CA: Sage.

Knapp, M., Copland, M., & Talbert, J. (2003). *Leading for learning: Reflective tools for school and district leaders*. Seattle: University of Washington, Center for the Study of Teaching Policy.

Kowalski, T. J., & Lasley, T. J. (2009). *Handbook of data-based decision making in education*. New York: Routledge.

Little, J. W. (2012). Understanding data use practice among teachers: The contribution of micro-process studies. *American Journal of Education*, *118*, 143–166.

Loewensen, F., & Gómez, R. (2009). Coming to our senses: The realities of program evaluation. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 83–96). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Magnan, S. S. (1986). Assessing speaking proficiency in the undergraduate curriculum: Data from French. *Foreign Language Annals*, *19*, 429–438.

Mandinach, E. B., Honey, M., Light, D. (2006 , April). *A theoretical framework for data-driven decision making* Paper presented at the annual meeting of the American Educational Research Association (pp. 1-18), San Francisco, CA.

Mandinach, E. B., Honey, M., Light, D., & Brunner, C. (2008). A conceptual framework for data-driven decision making. In E. B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 13–31). New York: Teachers College Press.

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, *114*, 1–48.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research*. Santa Monica, CA: RAND.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Author.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education*, *14*, 149–170. doi:10.1080/09695940701478321.

No Child Left Behind (NCLB). (2002). Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425.

Norris, J. M. (2009). Understanding and improving language education through program evaluation: Introduction to the special issue. *Language Teaching Research*, *13*, 7–13. doi:10.1177/1362168808095520.

Norris, J. M., Davis, J. M., Sinicrope, C., & Watanabe, Y. (2009). *Toward useful program*

evaluation in college foreign language education. Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Norris, J. M., & Watanabe, Y. (2007). *Roles and responsibilities for evaluation in foreign language programs*. Honolulu: University of Hawai'i at Manoa.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.) Thousand Oaks, CA: Sage.

Pfeiffer, P. C., & Byrnes, H. (2009). Curriculum, learning, and the identity of majors: A case study of program outcomes evaluation. In J. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 183–208). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Ramsay, V. (2009). Study abroad and evaluation: Critical changes to enhance linguistic and cultural growth. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 163–182). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Scarino, A., Elder, C., Iwashita, N., Hee, S., Kim, O., Kohler, M., & Scrimgeour, A. (2011). *Student achievement in Asian languages education; Part 1: Project report*. Retrieved March 18, 2014, from http://www.saale.unisa.edu.au/doclib/Part_I_all.pdf

Spillane, J. P. (2012). Data in practice: Conceptualizing the data-based decision-making phenomena. *American Journal of Education*, *118*, 113–141.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research*, *72*, 387–431. doi:10.3102/00346543072003387.

Walther, I. C. (2009). Developing and implementing an evaluation of the foreign language requirement at Duke University. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 117–138). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Watanabe, Y., Norris, J. M., & González-Lloret, M. (2009). Identifying and responding to evaluation needs in college foreign language programs. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 5–56). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Zannirato, A., & Sánchez-Serrano, L. (2009). Using evaluation to design foreign language teacher training in a literature program. In J. M. Norris, J. M. Davis, C. Sinicrope, & Y. Watanabe (Eds.), *Toward useful program evaluation in college foreign language education* (pp. 97–116). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Submitted January 18, 2014

Accepted March 2, 2014