

Proficiency Testing in the U.S. Context: An Introduction



Susan M. Gass and Paula Winke

Abstract The introductory chapter introduces the readers to The Language Flagship Proficiency Initiative in which institutions were charged with institutionalizing proficiency assessment practices that align student placement with course goals, document ways in which assessment results are integrated into foreign language programs, and share practices within the broader foreign language community. The chapter provides background on this project and summarizes the contents of the 14 chapters in the book.

Keywords Proficiency Testing · The Language Flagship · Assessment · ACTFL · Foreign Language Curriculum · Instructors · Second Language Learners · Speaking · Listening · Reading

In 2014, Michigan State University, along with the Universities of Minnesota and Utah, were awarded grants through The Language Flagship Proficiency Initiative to conduct foreign language proficiency assessments on their college campuses. The initiative was funded by the National Security Education Program (NSEP), a part of the Department of Defense. (Note that in 2012, NSEP merged with the Defense Language Office to form the Defense Language and National Security Education Office, otherwise known as DLNSEO). The grant program is under the umbrella of the Language Flagship program, and is intended to “integrate Flagship proficiency assessment practices and processes within existing high quality academic language programs. The purpose of this initiative is to introduce the Flagship proficiency assessment process to established academic foreign language programs to measure teaching and learning, and to evaluate the impact of such testing practices on teaching and learning” (p. 1, Request for Proposals, The Language Proficiency Flagship Initiative).

The Language Flagship programs were established in 2000 with the express goal of creating programs that would move students to advanced language proficiency in

S. M. Gass (✉) · P. Winke

Second Language Studies Program, Michigan State University, East Lansing, MI, USA

e-mail: gass@msu.edu; winke@msu.edu

© Springer Nature Switzerland AG 2019

P. Winke, S. M. Gass (eds.), *Foreign Language Proficiency in Higher Education*,
Educational Linguistics 37, https://doi.org/10.1007/978-3-030-01006-5_1

a select number of critical languages. Initially, the program served the graduate student population, but in 2006 moved toward a model of creating global professionals with high levels of proficiency, that is, Advanced or higher on the American Councils on the Teaching of Foreign Language (ACTFL) proficiency scale (ACTFL, 2012), which is equivalent in many respects to a level 3 proficiency level on the Interagency Language Roundtable (ILR) proficiency scale (<http://www.govtilr.org>). From their website (www.thelanguageflagship.org, retrieved 9/18/18) comes the following: “The Language Flagship graduates students who will take their place among the next generation of global professionals, commanding a superior level of proficiency in one of ten languages critical to U.S. national security and economic competitiveness.” Assessment is, of course, an important part of any language program as a way of understanding curricular needs and of determining successes and shortcomings of language programs in meeting their goals (cf. Bernhardt, 2008, 2014).

It was against this backdrop that the Language Flagship Program issued a call for institutions of higher education to partner with the Defense Language and National Security Education Office (DLNSEO) “to create a viable process to assess proficiency learning in high quality, well-established academic language programs and to document the impact of introducing rigorous proficiency assessment on language pedagogy practice and outcomes” (p. 3, Request for Proposals, The Language Proficiency Flagship Initiative). For a broader understanding of the Language Flagship program, the interested reader is referred to Murphy and Evans-Romaine (2017), and for a more complete discussion of the history of the Flagship Program and its contextualization into issues related to foreign language instruction more generally, see Nugent and Slater (2017). Prior to 2014 the Language Flagship programs had already had significant involvement with assessment and archived robust proficiency data from overseas study (see, in particular, Davidson, Garas, & Lekic, 2017). The data from Michigan State University and the Universities of Minnesota and Utah add to the already existing data from Flagship programs.

Our mandates for this project were the following:

- Institutionalize proficiency assessment practices that align student placement with course goals;
- Document ways in which assessment results are integrated into foreign language programs (curriculum and teaching);
- Share practices with others in the foreign language community.

This book is an attempt to provide information about assessment practices and results from the three universities to whom funding was provided. We have expanded the scope to include experiences and reports from other institutions in order to provide as broad a range of efforts to document language proficiency experiences and practices as possible.

The three universities that are part of the original grant project have approached their assessments in different ways and with different languages. Despite the individual directions and research reports, they have worked collaboratively to create common questions on a background questionnaire (given to all test takers at all three universities) and to combine results from their testing of speaking, listening, and reading into a large anonymized database so as to begin the process of creating

a broad picture of the proficiency levels of undergraduate students. The database will be sufficiently rich to allow researchers from around the world to address numerous research questions involving years of language study, the impact of study abroad, and other factors that might predict gains in proficiency (e.g., Winke & Gass, *in press*). Following a five-year embargo (to further protect the anonymity of the participants), this database will be available to researchers with specific research questions.

To give a sense of the scope of inquiry, we present data (Tables 1, 2, and 3) from the three universities involved in this project. The first two tables (Michigan State University and the University of Minnesota) show data per academic year; the third displays data in calendar years.

Table 1 Number of students who took proficiency tests administered at Michigan State University from 2014–2017 in four languages

	2014–2015	2015–2016	2016–2017 ^a
Chinese	250	272	162
French	526	510	301
Russian	116	115	57
Spanish	1155	962	936
Totals	2047	1859	1456

^aTests were not administered in Fall 2016, so the 2016–2017 numbers only include Spring 2017

Table 2 Number of students who took proficiency tests administered at University of Minnesota from 2014–2017 in seven languages

	2014–2015	2015–2016	2016–2017
Arabic	105	121	122
French	148	167	129
German	62	94	105
Korean	70	63	55
Portuguese	31	46	59
Russian	43	83	73
Spanish	239	254	267
Totals	698	828	810

Table 3 Number of students who took proficiency tests administered at University of Utah and Salt Lake Community College from 2014–2017 in five languages

	2014	2015	2016	2017
Arabic	0	35	38	45
Chinese	23	116	115	79
Korean	10	69	54	41
Portuguese	21	75	46	23
Russian	0	92	84	52
Totals	54	387	337	240

Michigan State University selected four languages to investigate: Chinese, French, Russian, and Spanish. These language programs represented different student-population sizes and provided a broad view of language proficiency levels across multiple (and diverse) programs. In Table 1 are data that show the distribution of the over 5300 students tested over the three-year period (2014–2017) in which proficiency tests were administered. Students took up to three proficiency tests each (speaking, reading, and listening), but due to some students not taking all three tests, only the number of students who took tests are displayed, not the total number of tests administered.

The University of Minnesota worked with 7 languages: Arabic, French, German, Korean, Portuguese, Russian, and Spanish. As can be seen in Table 2, over the three years of the grant, tests to 2336 students were administered over the course of three academic years.

University of Utah's assessments were of five languages: Arabic, Chinese, Korean, Portuguese, and Russian. Their testing program, unlike those of Michigan State University and the University of Minnesota, included students at a community college (Salt Lake Community College). Their student numbers are reported in calendar year as opposed to academic year.

As can be seen, the dataset that we are working with is large with nearly 9000 individuals tested across seven foreign languages.

Liberal arts, in general, and foreign language instruction, in particular, have been the subject of much debate. The Association of American Colleges and Universities (<https://www.aacu.org/leap>, retrieved 10/5/17), a public advocacy group launched in 2005, "champions the importance of a liberal education—for individual students and for a nation dependent on economic creativity and democratic vitality." Even though the specific context is the United States (the context for the chapters in this volume), this statement is not limited to the specific context in which it is written. As they noted on their website, there is a greater demand "for more college-educated workers and more engaged and informed citizens." Language is key to the enterprise of liberal education.

As we will see in Rifkin's chapter, more than most disciplines, language professionals have given considerable thought to and dedicated research efforts towards (1) the understanding of benchmarks in foreign language education, (2) the development of curricula that are geared toward helping students achieve those benchmarks, and (3) an understanding of ways to measure learning outcomes. As we will see, the foreign language community is in a strong position to serve as a model for other disciplines given the experience of articulated curricula and an understanding of ways to document and measure progress. In asking what higher education will look like in 2015 (10 years from the writing of his article), Yankelovich ("Ferment and Change: Higher Education in 2015." *Chronicle of Higher Education*, 25 Nov. 2005: 14) stated that "Our whole culture must become less ethnocentric, less patronizing, less ignorant of others, less Manichaeian in judging other cultures, and more at home with the rest of the world. Higher education can do a lot to meet that important challenge." He identified "the need to understand other cultures and languages" as significant to the future relevance of higher education. In fact, he stated that it is

one of five imperatives that must be foremost in thinking about higher education in the following 10 years. Clearly, language is central to this imperative and assessment of language proficiency is the only way to clearly understand the extent to which we can meet the goals as outlined in this statement.

In a report from the Commission on Language Learning established by the American Academy of Arts and Sciences (“America’s languages: Investing in language education for the 21st century,” 2017, <https://www.amacad.org/content/publications/publication.aspx?d=22474>, retrieved 9/18/18), the Commission on Language Learning points out the lack of emphasis over the years on language education and “recommends a national strategy to improve access to as many languages as possible for people of every region, ethnicity, and socioeconomic background—that is, to value language education as a persistent national need similar to education in Math or English, and to ensure that a useful level of proficiency is within every student’s reach” (p. viii). They suggest that there be a national strategy to “broaden access” (p. 27) to international study including cultural immersion and a general emphasis on “building a strong world language capability alongside English” (p. 31). To accomplish the goal of ensuring “a useful level of proficiency,” robust assessment measures are needed and an understanding of how to use those assessments to understand failures and successes in language programs.

The chapters in this volume address a range of issues that relate to policy as well as to specific practices. Data come from actual proficiency testing as well as from focus groups, surveys, and classroom observations. As will be seen, the issues are complex and include discussions of types of learners (e.g., heritage speakers, language majors) and specific uses of test results (e.g., self-assessments, proficiency/performance).

Rifkin, in his chapter sets the scene by discussing the role of performance-based assessment on the world stage. He highlights the work done in the field of foreign language instruction through ACTFL Proficiency Guidelines (ACTFL, 2012) and the World-Readiness Standards for Language Learning (NSFLEP, 2015) and the impact that these have had on curricular and pedagogical issues in foreign language teaching and, with particular relevance to the current volume, on issues of assessment. He discusses how foreign language education can be a leader in the liberal arts by modeling how disciplines can develop their own performance goals and align curricula with those goals to document the extent to which student learning outcomes match the pre-established performance benchmarks.

The remainder of the book is organized into three sections: (1) curricular issues, (2) assessment, and (3) instructors and learners. In the first section are four chapters dealing with proficiency goals within the programs of Arabic, Chinese, French, German, Korean, Portuguese, Russian, and Spanish in the United States, from different perspectives, including issues related to heritage language learners and language majors as well as more general issues in which curricular implications result from assessment results. In the first chapter in this section, Hacking, Rubio, and Tschirner report on data from the University of Utah. Their concern is with vocabulary size and reading proficiency for college level students of Chinese, German, Russian and Spanish. Using a database from approximately 200 students who had

taken both receptive vocabulary tests and reading proficiency tests, the authors show a strong correlation between receptive vocabulary knowledge and level of reading proficiency for all four languages. Noteworthy are the surprising low vocabulary sizes of the students tested. Relying on previous research which suggests that text comprehension requires vocabulary knowledge of 95%–98%, it is not surprising that it is difficult to reach advanced levels of proficiency without an emphasis on vocabulary. With regard to language program curricula, they suggest a rethinking of the emphasis on vocabulary. They note the paradox between the desire to focus on original literary texts and the low level of vocabulary knowledge of undergraduate language students.

Soneson and Tarone's chapter picks up on the issue of curriculum and assessment. The authors make the argument that foreign language programs can be greatly enhanced by three factors: regular assessments, student involvement in self-assessment, and professional development which includes the important aspect of community that comes from working across languages. In their chapter, they describe an ongoing project at the University of Minnesota that incorporates both of these dimensions and report on proficiency results after 2, 3, and 4 years of study. It is clear from reading their chapter that language programs can be significantly and positively impacted by incorporating all three dimensions. Their chapter is closely linked to the one by Sweet, Mack, & Olivero-Agney (Chapter 10) in which self-assessments are described, and the one by Dillard (Chapter 13) in which issues of professional development are detailed.

Kagan and Kudyma focus their chapter on heritage speakers of Russian. This chapter makes an important link between the Language Flagship Proficiency Initiative and two centers at the University of California, Los Angeles (Russian Language Flagship and another federally-funded center, the National Heritage Language Center, funded by the Department of Education, as part of their Title VI Language Resource Center programs) in that the data presented originated from students participating in these centers. Their database comes from questionnaires and online placement tests administered to heritage speakers of Russian. They question the placement test itself (concluding that the use of a multiple skills test is important in that skills differ from student to student), ask about strength and weaknesses of heritage speakers (listening is typically strong, but amount of schooling is a significant variable), and address the relationship between the placement test and the curriculum. In particular, they argue for the need to have the curricula address the specialized needs of heritage learners to allow the learners greater opportunity to reach high levels of proficiency.

Winke, Gass, and Heidrich consider data from language majors to determine the proficiency levels of French, Russian, and Spanish majors in listening, speaking, and reading. They compare their data with the data of Carroll (1967), an important study that took a broad view of proficiency levels of foreign language majors. Fifty years later, a similar picture emerges with speaking and listening skills falling behind. What is different, however, is the general picture of what it means to be a major. In 1967, the typical profile of a language major was a specialization in the

language and literature of that culture as a sole major. Today, most language majors have another major alongside language study (e.g., business, engineering). A second area of investigation concerned the possible predictors of success amongst language majors. They found that heritage status, study abroad and intrinsic motivation were important predictors, but amongst those three, it was intrinsic motivation that stands out. Similar to the findings of Carroll, a factor that stands out is when language learning begins, with greater progress being made in college-level courses when language learning begins early. They make suggestions that relate to general issues of curriculum and emphasize the important role of foreign language study in secondary education.

In the second section of this book six chapters deal with assessment with many of the same languages dealt with in Part 1, in particular, Arabic, Chinese, French, German, Japanese, Korean, Portuguese, Russian, and Spanish. Self-assessment is the topic of two of the papers as a way of helping students learn to help themselves understand and increase language proficiency.

Cox, Bown, and Bell question the assessment measure itself. The specific focus is on reading proficiency assessments and the format of the test. What should the language of the question be? Should it be in the first or target language? Cox, Bown, and Bell investigate the common wisdom showing that when the question is in the L2, scores are lower. However, the issue of why this should be the case has not been explored. Their database comes from reading tests taken by advanced adult L2 learners and incorporates affective characteristics. Russian learners responded to short reading passages that were followed by a single multiple-choice question, half of which were in Russian and half in English. Measures of confidence and anxiety were collected after each question. The language of the question did have an impact to score differences (responses to English questions were higher than responses to Russian questions). Cox, Bown, and Bell explored reasons why some preferred L2 questions and others preferred English questions. Further discussion in their chapter deals with alignment of the language of the question and the criteria that are being assessed.

In the second paper, Hacking and Rubio look at the vexed question of proficiency (using language in situations that reflect a real-world context) and performance (using language that has been learned in an instructional setting). They question if the construct of proficiency is appropriate for students at low levels of instruction. In other words, is it realistic to expect students to take language learned in a classroom setting and extend it to novel situations? They point to the contradiction in most testing programs at the lower levels, namely, that they are designed to measure global proficiency, but definitions (e.g., ACTFL Standards) of low levels entails a lack of functional proficiency and focus on memorized speech. In other words, there is no proficiency at low levels.

In one of the two chapters on self-assessment in this section, Tigchelaar compares self-assessment data from French language students with their actual speaking test scores. She includes in her discussion the notion of rating scales which have been used to convert ACTFL Oral Proficiency Interview (computer delivered; OPIc)

scores to a numeric scale. How well self-assessment scores predicted actual OPIc scores depended on the actual numeric scale used. There are numerous important implications that stem from this chapter for using self-assessments for placement and instruction and for converting ACTFL scores to numeric scores when conducting research.

In the second chapter dealing with self-assessment, Sweet, Mack, and Olivero-Agney acquaint us with the self-assessment tool Basic Outcomes Student Self-Assessment (BOSSA). They build on the assumption that self-assessment increases learners' involvement in the learning process and, as a consequence of engagement, increases success. They describe the components of BOSSA and report on its use at the University of Minnesota. Students, through the use of the BOSSA tool, learn how to track their progress, understand how their learning progresses, and set their learning goals. When learning goals are established, students are involved in determining how to reach those goals. Their data show that students demonstrate increased awareness over time of their own learning processes and their own abilities. In this chapter, Sweet et al. report on the degree of accuracy (more so in speaking and less so in reading and listening) in self-assessment as related to actual scores, and chart the future for using self-assessment particularly in light of the fact that not all students perceive the value of BOSSA to their own learning.

The next chapter in this section by Vanpee and Sonesson specifically describes the implementation in the Arabic language program of a project at the University of Minnesota, Proficiency Assessment for Curricular Enhancement (PACE). The particular focus is on how regular assessments can result in actual proficiency improvement. They show how the triangulation of efforts of proficiency assessment, self-assessment (i.e., student involvement), and professional development can ultimately result in improved proficiency. Their focus is on speaking and reading results and they report on improvements over a two-year period. An important point made is the 'culture of assessment' that is the result of the Flagship Proficiency Initiative and the need to supplement these external assessments with student involvement and a regular program of professional development of individual instructors and the collaborative work of all instructors. They recognize the difficulty in implementing these programs and discuss these limitations as a way of guiding others in institutionalizing some of the best practices they outline.

Davidson and Shaw, in the final chapter of the assessment section, present a detailed analysis of L2 outcomes of students who studied abroad in a year-long (academic year) program. They report substantial proficiency gains from pre-post program. Their results are particularly powerful in that the gains are not limited to one language but rather hold across all languages tested (Arabic, Chinese, Russian) and across all modalities. There are a number of important correlations found in their study. Of particular note are the pre-program listening scores which positively correlate with speaking skill growth. Reading abilities are related to gains in speaking and listening. In general, what can be seen from their paper is the importance of structured immersion programs even when there is little prior L2 knowledge.

The third section focuses on individuals, in particular, learners and instructors. The chapters in this section individuals teaching and learning three languages: Japanese, Spanish, and Chinese.

In the first chapter, Dillard builds on the work discussed by Vanpee and Soneson (Chapter 11) regarding PACE, this time focusing on two Japanese instructors who participated in an inquiry group following changes to the Japanese language program curriculum and to actual instructional practices needed to address those changes and the problems associated with those changes. The basis for the inquiry group discussions was the exploratory practice model and lesson study. A guiding question was: *How do elements of a multilingual language instructor inquiry group serve to mediate language teacher conceptual development within the broader sociocultural context?* Numerous tools were used to address this question including classroom observations and video recordings both with the goal of understanding student learning by identifying moments of teacher learning through transcripts of group conversations. The chapter serves to illustrate the development of teacher cognition through the group inquiry system; it also makes a methodological point by examining the usefulness of the inquiry group model itself. Rich with examples, this chapter shows how teacher growth comes from contradictions and tensions resulting in changes in teacher awareness and acceptance of different ways of thinking.

The chapter by Maloney investigates the important topic of digital literacy practices of students and the resultant connection to proficiency. His study is based on a survey administered to students studying Spanish in which information was requested on the use of technology in Spanish for language learning and for entertainment. The survey was followed by proficiency assessment in speaking, listening, and reading. In addition, interviews were conducted with students in which attitudes toward technology digital literacy use were probed. To complete the study, instructors' views on incorporating technology in the classroom were collected in order to get a more complete view of technology use and attitudes. Maloney found a relationship between proficiency and different practices of technology use addressed in the survey. One of the difficulties uncovered from interviews is the lack of knowledge of potential L2 resources as well as limited proficiency (particularly their perception of their proficiency), making it difficult to use the full range of L2 materials available. Not surprisingly, those who had studied abroad reported greater use.

In the final chapter in this book, Polio utilizes classroom observations as one data source for her chapter. She takes on the difficult task of relating proficiency scores to classroom practices, using Chinese language classes as the basis. Hers is a mixed-method study and combines data from the proficiency scores administered to Chinese students and qualitative data from classroom observations and focus group interviews. She used activity charts to document lesson foci, the type of interaction, and the amount of Chinese spoken. Tests were administered twice in an academic year and were scored using the ILR scale. She found that, indeed, there was improve-

ment in speaking scores over the year, but not in listening or in reading. The emphasis on oral skills was confirmed by the instructors and by the classroom traits that Polio focused on. Additionally, in this chapter Polio elucidates difficulties in conducting mixed-methods classroom research and suggests ‘ideal’ data for making the important link between classroom practice and learning outcomes.

And, finally, Malone summarizes the chapters and provides the assessment community with five recommendations for future research and action.

The chapters in this book all address issues of proficiency, albeit from different perspectives. Many, but not all, are based on data from the Proficiency Initiative, part of the Language Flagship Program funded by DLNSEO. Most use ACTFL testing as their assessment measure, although other measures are used as well. Languages represented are spoken in all corners of the world; some of the data come from large language programs; others come from relatively small programs. In all, they contribute to our understanding of foreign language education and include successes and failures in our efforts to increase language proficiency in undergraduate language programs.

References

- ACTFL (2012). *ACTFL proficiency guidelines*. Retrieved from <http://www.actfl.org>.
- Bernhardt, E. (2008). Systemic and systematic assessment as a keystone for language and literature programs. *ADFL Bulletin*, 40(1), 14–19. <https://doi.org/10.1632/adfl.40.1.14>
- Bernhardt, E. (2014). Assessment that supports teaching, learning, and program development. *ADFL Bulletin*, 43(1), 15–22. <https://doi.org/10.1632/adfl.43.1.15>
- Carroll, J. B. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals*, 1(2), 131–151. <https://doi.org/10.1111/j.1944-9720.1967.tb00127.x>
- Davidson, D., Garas, N., & Lekic, M. (2017). Assessing language proficiency and intercultural development in the overseas immersion context. In D. Murphy & K. Evans Romaine (Eds.), *Exploring the US language flagship program: Professional competence in a second language by graduation* (pp. 156–176). Bristol, UK: Multilingual Matters.
- Murphy, D., & Evans-Romaine, K. (Eds.). (2017). *Exploring the US language flagship program: Professional competence in a second language by graduation*. Bristol, UK: Multilingual Matters.
- NSFLEP. (2015). *World-readiness standards for learning languages* (4th ed.). Alexandria, VA: The American Council on the Teaching of Foreign Languages (ACTFL).
- Nugent, R., & Slater, R. (2017). The language flagship: Creating expectations and opportunities for professional-level language learning in undergraduate education. In D. Murphy & K. Evans Romaine (Eds.), *Exploring the US language flagship program: Professional competence in a second language by graduation* (pp. 9–28). Bristol, UK: Multilingual Matters.
- Winke, P., & Gass, S. (in press). Individual differences in advanced proficiency. In P. Malovrh & A. Benati (Eds.), *Handbook of advanced proficiency in second language acquisition*. New York, NY: Routledge.

Susan M. Gass is University Distinguished Professor at Michigan State University. She is co-PI with Paula Winke on the Proficiency Initiative Grant. She has published widely in the field of second language acquisition and is the winner of numerous local, national, and international awards for her research and contributions to the field. She has served as President of AAAL and AILA and is currently Editor of *Studies in Second Language Acquisition*.

Paula Winke received her Ph.D. from Georgetown University in 2005. She is Associate Professor at Michigan State University, USA. Her research interests are in language assessment for both summative and formative assessment purposes. She also researches proficiency and standards-based language assessments: She investigates the ethics of using scores from such tests to fulfill policies related to school and or career/position advancement. She is an incoming editor (starting in 2019) of the journal *Language Testing*.