Educational Linguistics

Paula Winke
Susan M. Gass *Editors*

Foreign Language Proficiency in Higher Education

Springer

Educational Linguistics

Volume 37

Educational Linguistics is dedicated to innovative studies of language use and language learning. The series is based on the idea that there is a need for studies that break barriers. Accordingly, it provides a space for research that crosses traditional disciplinary, theoretical, and/or methodological boundaries in ways that advance knowledge about language (in) education. The series focuses on critical and contextualized work that offers alternatives to current approaches as well as practical, substantive ways forward. Contributions explore the dynamic and multi-layered nature of theory-practice relationships, creative applications of linguistic and symbolic resources, individual and societal considerations, and diverse social spaces related to language learning.

The series publishes in-depth studies of educational innovation in contexts throughout the world: issues of linguistic equity and diversity; educational language policy; revalorization of indigenous languages; socially responsible (additional) language teaching; language assessment; first- and additional language literacy; language teacher education; language development and socialization in non-traditional settings; the integration of language across academic subjects; language and technology; and other relevant topics.

The *Educational Linguistics* series invites authors to contact the general editor with suggestions and/or proposals for new monographs or edited volumes. For more information, please contact the publishing editor: Jolanda Voogd, Senior Publishing Editor, Springer, Van Godewijckstraat 30, 3300 AA Dordrecht, The Netherlands.

More information about this series at http://www.springer.com/series/5894

Paula Winke • Susan M. Gass
Editors

Foreign Language Proficiency in Higher Education

Springer

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

Susan M. Gass
Second Language Studies Program
Michigan State University
East Lansing, MI, USA

ISSN 1572-0292 ISSN 2215-1656 (electronic)
Educational Linguistics
ISBN 978-3-030-01005-8 ISBN 978-3-030-01006-5 (eBook)
https://doi.org/10.1007/978-3-030-01006-5

Library of Congress Control Number: 2018961737

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part I
Preliminaries

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 assessement 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

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**

[a]Tests 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 Manichaean 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 Soneson 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 tin 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*.

The Power of Performance-Based Assessment: Languages As a Model for the Liberal Arts Enterprise

Benjamin Rifkin

Abstract *The ACTFL Proficiency Guidelines* and the *World-Readiness Standards for Language Learning* have had an enormous impact on the design and delivery of instruction in the second language and foreign language fields in the United States; this, in turn, has had an impact on the nature of learning outcomes in second and foreign language education, as demonstrated in the essays in this volume. With these performance metrics and curricular foci in mind, experts in the second and foreign language fields have engaged in the purposeful design and delivery of curricula and the assessment of learning outcomes that meet the aspirations of the Association of American Colleges and University's *Liberal Education and America's Promise* program (LEAP) as well as the expectations of the regional accrediting agencies for higher education in the United States (Higher Learning Commission, Middle States Commission on Higher Education, New England Association Schools and Colleges Commission on Institutions of Higher Education, South Association of Colleges and Schools Commission on Colleges, and the Western Association of Schools and Colleges Senior College and University Commission). In this context, the second and foreign language fields can serve as a model for many other liberal arts disciplines in the development and use of their own discipline-specific performance benchmarks and the development and implementation of curricula that foster the attainment of measurable student learning outcomes.

Keywords Proficiency · Assessment · Learning outcomes · Second language · Liberal arts · Accreditation · Foreign language · Metrics

The contributions to this volume stand as compelling evidence of the leading position of the world languages field among liberal arts disciplines in post-secondary education, despite marginalization – whether deliberate or not – by governmental

B. Rifkin (✉)
Hofstra College of Liberal Arts and Science, Hofstra University, Hempstead, NY, USA
e-mail: benjamin.rifkin@hofstra.edu

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_2

bodies (local, state, or federal) as well as by administrative leadership in local institutions. When governmental bodies and institutional administrations eliminate or reduce funding for world language education, leading to the elimination of languages of instruction in post-secondary institutions despite the importance of language for economic, geopolitical, and cultural purposes, world language instruction is devalued and marginalized, and instructors demoralized. However, despite these funding practices, the world language field leads other liberal arts disciplines in post-secondary education in two critically important ways, instantiated by these two documents:

(a) The development and adoption of internationally recognized performance benchmarks: *The ACTFL Proficiency Guidelines* (2012)
(b) The development and widespread implementation of curricular guidelines: *The World-Readiness Standards for Learning Languages* (2015)

These two documents, about which more below, lie at the core of each and every essay in this volume, demonstrating their importance for our field and distinguishing our field from the other liberal arts disciplines in the post-secondary educational enterprise.

The ACTFL Proficiency Guidelines constitute a series of performance benchmarks enabling the evaluation of language learner performances in all four modalities of speaking, listening, reading, and writing in accordance with an internationally recognized rubric. The benchmarks, for those who are not familiar with them, break down performances into four major levels (named "novice," "intermediate," "advanced," and "superior") with sublevels at all but the superior level (low, mid, and high). Descriptions of performance at each major level and sublevel are provided for each of the four modalities, with a focus on length, complexity, register, and genre of discourse, and, for the productive modalities of speaking and writing, expectations for grammatical and syntactical accuracy. Given the articulation of performance benchmarks, *The ACTFL Proficiency Guidelines* have subsequently become the source text for the development of proficiency-based assessments, including the best-known Oral Proficiency Interview, the Writing Proficiency Test, and other instruments designed to assess listening and reading proficiency in correlation with the proficiency guidelines for those modalities.

The World-Readiness Standards for Language Learning constitute a series of curricular foci to help those engaged in the design and/or delivery of curricula in world languages: communication (ability to interpret spoken and written texts, interact and present in speech and writing in the target language), culture (understanding products and perspectives), connections (using the language to connect to information not generally available in the native language), comparisons (comparing native and target languages and cultures), and communities (engaging with communities in which the target language is spoken, whether in person or through social media, becoming a life-long learner in doing so). The standards help us as a field move away from an exclusive focus on the teaching of grammar, while provid-

ing instructors with a framework in which to purposefully construct lessons focused on using the target language to learn about the target culture's history, for example, or lessons focused on using the target language to engage with native speakers.[1]

These two documents are closely correlated with the *Liberal Education and America's Promise* (LEAP) program of the Association of American Colleges and Universities (AAC&U). *The Essential Learning Outcomes of LEAP* (2005) call for students to develop:

> Knowledge of Human Cultures and the Physical and Natural World through study in the sciences and mathematics, social sciences, humanities, histories, languages, and the arts, focused by engagement with big questions, both contemporary and enduring.

The Essential Learning Outcomes document also calls for students to develop

> intellectual and practical skills, including inquiry and analysis, critical and creative thinking, written and oral communication, quantitative literacy, information literacy, and teamwork and problem solving practiced extensively across the curriculum in the context of progressively more challenging problems, projects, and standards for performance.

The document then calls for students to develop

> personal and social responsibility, including civic knowledge and engagement, both local and global, intercultural knowledge and competence, ethical reasoning and action, and foundations and skills for lifelong learning, anchored through active involvement with diverse communities and real-world challenges.

Finally, the *Essential Learning Outcomes* calls for students to engage in

> integrative and applied learning, including synthesis and advanced accomplishment across generalized and specialized studies, demonstrated through the application of knowledge, skills, and responsibilities to new settings and complex problems.

As I have written elsewhere (Rifkin, 2012), the *Essential Learning Outcome*s map quite beautifully onto the *Proficiency Guideline*s and the *World-Readiness Stan*dards. Students of world languages, in a proficiency-oriented and standards-based curriculum, have the opportunity to grow in all of these areas in the context of their foreign language learning experiences, as illustrated in Table 1.

In addition to the correlation of the LEAP *Essential Learning Outcomes* to the *Proficiency Guidelines* and the *World-Readiness Standards*, the systematic assessment of learning outcomes is critical for each of the post-secondary regional accreditation agencies: The Higher Learning Commission, Middle States Commission on Higher Education, New England Association Schools and Colleges Commission on Institutions of Higher Education, South Association of Colleges and Schools Commission on Colleges, and the Western Association of Schools and Colleges Senior College and University Commission. Each of these accrediting agencies requires the demonstration of direct evidence of learning for reaccreditation. Faculty

[1] I believe that the two documents are unfortunately misnamed. *The Proficiency Guidelines* are actually standards, in that they describe benchmarks of performance, while the *World-Readiness Standards* are actually guidelines, in that they describe points of focus for a coherent curriculum.

Table 1 Map of essential learning outcomes/Proficiency guidelines/World-readiness standards

Essential learning outcomes	Proficiency guidelines	World-readiness standards
Knowledge of Human Cultures		Culture Standard
Engagement in Big Questions, Contemporary and Enduring	Advanced and Superior-Level Proficiency	Culture Standard
		Connections Standard
		Comparisons Standard
Inquiry and Analysis	Intermediate, Advanced, and Superior-Level Proficiency Functions	Connections Standard
		Comparisons Standard
Critical and Creative Thinking	Intermediate- through Superior-Level Proficiency	Communication Standard
		Connections Standard
		Comparisons Standard
Written and Oral Communication	All levels of Proficiency Guidelines	Communication Standard
Information Literacy		Culture Standard
		Connections Standard
Progressively More Challenging Problems	All levels of Proficiency Guidelines	All Standards
Civic Knowledge and Engagement	Advanced and Superior-Level Proficiency	Connections Standard
		Communities Standard
Intercultural Knowledge and Competence	Advanced and Superior-Level Proficiency	Culture Standard
		Comparisons Standard
		Connections Standard
		Communities Standard
Lifelong Learning	Intermediate- through Superior-Level Proficiency	All 5 Standards
Involvement with diverse communities and real-world challenges	Intermediate- through Superior Level Proficiency	All 5 Standards
Integrative and Applied Learning	Intermediate through Superior-Level Proficiency	All 5 Standards

in the world languages field at accredited institutions have nationally recognized frameworks, the *Proficiency Guidelines* and the *World-Readiness Standards*, to demonstrate the quality of curricula (*World-Readiness Standards*), the "input," as it were, and the nature of the learning outcomes (*Proficiency Guidelines*), the "output," as it were.

And here is where the world languages field distinguishes itself among the liberal arts disciplines of post-secondary education. Some of the fields, such as Chemistry (which has its own professional accrediting association, the American Chemical Society), have documents describing the areas of study within the field (ACS, 2015), but do not provide descriptions of performance benchmarks. What does it mean to demonstrate historical thinking skills at the intermediate level? The American Historical Association does not give us any guidance in the latest version of their undergraduate "tuning project" (2016). What are the characteristics of a successful psychology senior capstone research paper? The American Psychological Association is silent, although its *Psychology Major Guidelines* (2013) does list topics for inclusion in the curriculum. For example, the *Psychology Major Guidelines* do stipulate that learners should demonstrate effective writing in the psychology major, but give no indications as to what that would look like developmentally over the course of a 4-year curriculum. The Mathematical Association of America, similarly, lists topics to be taught in its 117-page *Curriculum Guide* (2016), but does not offer any benchmark performance descriptions. After conversations with scholars representing the breadth of the liberal arts disciplines in post-secondary education – including disciplines in the humanities, social sciences, and natural sciences – and after reviewing websites for disciplinary associations in many liberal arts disciplines across the disciplinary spectrum, I concluded that many disciplines offer no curricular frameworks whatsoever (akin to our field's *World Readiness Standards for Language Learning*) and that **none**, not a single other field, even Chemistry, offers benchmark performance descriptions. To be clear, I commend all of the organizations for their work thus far on the consideration of curriculum and learning outcomes; all of the projects I've mentioned here are serious and meaningful. Nonetheless, they fall short of the achievements of the world languages field.[2]

The impact of the achievements of the world languages field, as evidenced by the *Proficiency Guidelines* and the *World-Readiness Standards*, is enormous, as evidenced by the chapters in this volume. We begin with an overview of proficiency testing in the US and an analysis of the impact of that testing on language teaching and Learning (Winke and Gass). Moving on to consider curricular issues, we take up the questions of: vocabulary, reading proficiency and curricular design (Hacking, Rubio, and Tschirner); how proficiency assessment can enhance curricular design

[2]The professional fields typically do have learning outcomes metrics and the liberal arts disciplines, including world languages, could benefit from a closer investigation of the success of those metrics in their respective fields, e.g., AACSB for Business, ABET for Engineering, and so forth. Moreover, it could be very useful for the liberal arts fields to examine learning outcomes metrics in the public K-12 setting.

(Sonenson and Tarone); how assessment impacts curricular design for heritage learners (Kagan and Kudyma); and how assessment informs the design of the foreign language major (Winke and Gass). In the second section of the volume, we learn about: the distinction between proficiency and performance (Hacking and Rubio); the value of self-assessment in two essays (Tigchelaar and Sweet, Mack and Olivero-Agney); and the impact of proficiency assessment on the learning of Arabic (Vanpee and Soneson). In Sect. 3, we take up the question of instructors and learners and consider: an approach to Japanese teacher development (Dillard); digital literacy practices and world language learning (Maloney); and proficiency test scores and classroom practices for Chinese (Polio). In Sect. 4 we focus on skill development, considering: the connection between developing listening skills and speaking skills (Tschirner, Gass, Hacking, Rubio, Soneson, Winke) and the language of the question in measuring reading proficiency (Cow and Bown). Finally, the volume concludes with an afterword by Malone.

With nationally recognized guidelines for curricular design, world language educators share curricular parameters. This allows us to discuss, with a common language: the appropriateness of content for one or another pedagogical level; and the effectiveness of any given content for working with different learner populations (e.g., foreign language learners, heritage learners, first-generation learners, and so forth), among other topics. With nationally recognized benchmark performance indicators, we can discuss, with a common language: the effectiveness of one or another pedagogical approach (based on learning outcomes); the number of contact hours needed to help learners reach a certain learning outcome; the impact of changes in curricular design and/or delivery on learning outcomes. In turn, these discussions, such as those that will inevitably arise in response to the chapters in this volume, will shape the subsequent design and delivery of world languages instruction. Thus, our field is poised to make steady progress toward improving learner outcomes because we share a common language to post meaningful hypotheses, collect and analyze data, and share findings. Clearly, not all world language instructors subscribe to *The Proficiency Guidelines* and *The World-Readiness Standards*, but in time, I predict, the success of those programs in which faculty design and deliver their curricula and measure learning outcomes in accordance with these documents will ultimately attract more and more practitioners to adopt proficiency-oriented and standards-based pedagogical models for their programs.

Thus, the impact of work, such as that showcased in this volume, is reflected in the growing number of programs that embrace this evidence-based approach to the learning and teaching dynamic in the world languages curricula at the post-secondary level. Moreover, work in the world languages field that focuses on the power of performance assessment has the potential to have a significant impact on the other liberal arts disciplines of the post-secondary educational enterprise more generally. While the instructional objectives for faculty in anthropology, history, philosophy, sociology and all the other liberal arts disciplines are certainly different from those in the foreign languages, the potential for curricular improvement drawn from the assessment of actual student performance in the given field remains enormous. The development of performance benchmarks, i.e., the articulation of what

performance or products are expected of a novice-level student of anthropology, an intermediate-level student of history, an advanced-level student of philosophy, or a superior-level student of sociology, would have transformative impact on the design and delivery of instruction in those fields. As the dean of a liberal arts college within a larger research university, I invite scholars from other disciplines to consider the rich experience of the foreign language field in the design and implementation of performance-based assessment and the impact that assessment has had, continues to have, and will have on the design and delivery of instruction in the foreign language field.

References

American Chemical Society. (2015). *Undergraduate professional education in chemistry: ACS guidelines and evaluation procedures for bachelor's degree programs.* Retrieved from https://www.acs.org/content/dam/acsorg/about/governance/committees/training/2015-acs-guidelines-for-bachelors-degree-programs.pdf

American Council on the Teaching of Foreign Languages. (2012). *The ACTFL proficiency guidelines 2012.* Retrieved from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

American Historical Association. (2016). *Tuning the history discipline*. Retrieved from https://www.historians.org/teaching-and-learning/tuning-the-history-discipline

American Psychological Association. (2013). *APA guidelines for the undergraduate psychology major, version 2.0.* Retrieved from http://www.apa.org/ed/precollege/about/psymajor-guidelines.pdf

Association of American Colleges and Universities. (2005). *Essential learning outcomes of liberal education & America's promise.* Retrieved from http://aacu.org/leap/essential-learning-outcomes

Mathematical Association of America Committee on Undergraduate Majors and Programs. (2015). http://maa.org/sites/default/files/CUPM%20Guide.pdf

National Standards for Foreign Language Education Project. (2015). *World-readiness standards for language learning*. Retrieved from https://www.actfl.org/publications/all/world-readiness-standards-learning-languages.

Rifkin, B. (2012, Summer). The world language curriculum at the center of the post-secondary curriculum. *Liberal Education, 98.3*, 54–57. Also on line at http://www.aacu.org/liberaleducation/le-su12/rifkin.cfm

Benjamin Rifkin is Professor of Russian and Dean of the College of Liberal Arts and Sciences at Hofstra University, USA. His research interests include Russian language and culture, foreign language education, and higher education policy. He has served as an ACTFL Oral Proficiency Interviewer and OPI Trainer and as director of the Middlebury School of Russian. His work has appeared in journals such as Foreign Language Annals, Modern Language Journal, and Slavic and East European Journal.

Part II
Curriculum

Vocabulary Size, Reading Proficiency and Curricular Design: The Case of College Chinese, Russian and Spanish

Jane F. Hacking, Fernando Rubio, and Erwin Tschirner

Abstract A key goal of college foreign language study is L2 literacy development and literary texts from the target culture form the backbone of upper division curricula. Much of the empirical research to date on vocabulary size and reading proficiency has focused on learners of English. This article presents data on the reading proficiency level of 155 college students of Chinese (N = 46), Russian (N = 48) and Spanish (N = 61) and considers these results in terms of these same students' receptive vocabulary knowledge in the language. The study shows very high correlations between reading proficiency and receptive vocabulary size and that, in general, the vocabulary sizes of the college students participating in the study were not sufficient to read at the Advanced level. We suggest that programs and instructors consider a more intentional approach to vocabulary learning across the curriculum.

Keywords Reading · Proficiency · Vocabulary · Postsecondary · Assessment · Chinese · Russian · Spanish

1 Introduction

Undergraduate foreign language programs aim to develop students' L2 proficiency in speaking, reading, listening and writing. Reading proficiency assumes particular importance as it gives students access to literary texts from the target culture that form the backbone of upper division curricula. There is extensive research on

J. F. Hacking (✉)
World Languages and Cultures, University of Utah, Salt Lake City, UT, USA
e-mail: j.hacking@utah.edu

F. Rubio
Second Language Teaching and Research Center, University of Utah, Salt Lake City, UT, USA
e-mail: Fernando.Rubio@utah.edu

E. Tschirner
Herder Institute, University of Leipzig, Leipzig, Germany
e-mail: tschirner@uni-leipzig.de

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_3

second language reading that explores the role of various factors (e.g., L1 literacy, grammatical and/or vocabulary knowledge) in the development of L2 reading proficiency. Bernhardt (2011) argues for a model of L2 reading proficiency that captures the contribution and interaction of these many variables. A deeper understanding of the role of specific variables, such as vocabulary knowledge, will contribute to such a comprehensive model. This article presents data on the reading proficiency level of 155 college students of Chinese (N = 46), Russian (N = 48) and Spanish (N = 61) and considers these results in terms of these same students' receptive vocabulary knowledge in the language. Much of the empirical research to date on vocabulary size and reading proficiency has focused on learners of English, but a recent study of L2 Russian learners (Hacking & Tschirner, 2017) reported identifiable lexical minima associated with particular levels of L2 Russian reading proficiency. This chapter builds on the Russian data with the addition of data from learners of Chinese and Spanish.

2 Background

2.1 Vocabulary Size and L2 Development

The relationship between a second language (L2) learner's vocabulary size and his or her overall L2 proficiency has been well-established. Milton (2009) provides a comprehensive overview of this general line of research. There are also a number of studies that focus on the relationship between vocabulary knowledge and reading ability in particular (e.g., Nation, 2006; Tschirner, 2004). Staehr (2008) reports research showing that vocabulary size correlated strongly with listening, reading and writing scores, but that, of these, the strongest correlation was with reading scores. An important finding in the reading proficiency and vocabulary knowledge research is that there are lexical thresholds associated with the achievement of specific language goals. Studies have found that, for example, a particular exam score, a certain proficiency rating, or a reading comprehension score requires specific levels of L2 vocabulary knowledge.

Overall, the research suggests that lexical thresholds are tied to expected text coverage, that is, the percentage of lexical items in a text the reader understands. Researchers concur that readers need to understand between 95% and 98% of the tokens (running words) of a text in order to comprehend the text (Carver, 1994; Hirsh & Nation, 1992; Hu & Nation, 2000; Nation, 2006; Schmitt, 2008; Schmitt, Jiang, & Grabe, 2011). Nation (2006) proposes that for English the most frequent 9000 word families provide coverage of 98% of words in a wide range of texts. Similarly, Laufer and Ravenhorst-Kalovski (2010) assert that for English learners, an optimal threshold for reading purposes is knowledge of 8000 words and a minimal one is 4000–5000 words. Clearly then, adequate comprehension of unsimplified texts is not readily accessible for L2 learners until they have acquired substantial

vocabulary knowledge. As Nation (2007, p. 9) noted, "…most text beyond the 3,000-word level of graded readers series is very difficult for foreign language learners. This is because in most novels a very large number of different words occur beyond the learners' current vocabulary knowledge." While Nation focuses on narrative texts, similar figures have been cited for newspaper and other kinds of writing (Schmitt et al., 2011). And while reading itself can be a route to vocabulary acquisition, research shows approximately twelve repetitions of a word in different contexts may be needed for the word to be acquired (Nation, 2014), and that, to achieve twelve repetitions of, e.g., the fourth most frequent 1000 words in English, i.e. the most frequent 3000–4000 words, approximately half a million running words need to be read. We explore the implications of this research in light of results of this study in the discussion section where we consider approaches to vocabulary learning.

The idea of lexical thresholds is found also in descriptions and discussions of various proficiency scales. Descriptors can be broadly worded, for example, "limited range of vocabulary" or "uses an adequate range of vocabulary for the task" (IELTS, n.d.). Or, they can appeal to concrete numbers. Milton (2009), in discussing the CEFR, notes that at earlier stages of development, CEFR descriptors included vocabulary lists. He references lists tied to level B1, observing that different languages had different thresholds according to these lists (German 2400 words, English 2200 words, Italian and French 1800 words, and Spanish 800 words.) For the languages considered here, the national testing instruments of those languages establish several levels of proficiency and in the case of Russian and Chinese, establish lexical minima for each level.

The Russian Ministry of Education and Science has developed the Test of Russian as a Foreign Language (TORFL)/Тест по русскому языку как иностранный (ТРКИ) as part of the CEFR. Some official test specifications can be found on the website of Moscow State University's Training and Testing Language Center for Foreigners: http://russian-test.com/tests/torfl/. Learners may achieve one of six levels: Elementary, Basic, Level 1, Level 2, Level 3, Level 4. Each level outlines a set of competencies and is accompanied by a description of what language at that level enables the learner to do. For example, a learner who scores at the Basic Level, the level required to become a naturalized Russian citizen, is described as being able to "satisfy the most basic communicative needs…in a limited number of predictable situations" (http://russian-test.com/assets/docs/Trebovaniya_-_basic.pdf; translated from the Russian). The official test documentation also specifies a minimum number of vocabulary words required at a given level as shown in Table 1. After Level 1, vocabulary knowledge targets are split into receptive and productive categories. Productive vocabulary knowledge targets are smaller than those for the receptive lexicon. Test documentation states that a learner must have achieved Level 1 to begin a course of study at a Russian institution of higher education, and that Level 2 proficiency is necessary to receive a degree taught in Russian (with the exception of degrees in, for example, philology for which Level 3 proficiency is the stated requirement). Table 1 shows TORFL proficiency levels with minimum vocabulary for each as well as established equivalencies between TORFL levels and those

Table 1 ACTFL, CEFR, correspondences to TORFL and lexical minimums per level as established by the TORFL

ACTFL	CEFR	TORFL	Minimum vocabulary
N	A1	Elementary	780
IM	A2	Basic	1300
IH	B1	Level 1	2300
AM	B2	Level 2	10,000 (6000 active)
AH	C1	Level 3	12,000 (7000 in active)
S	C2	Level 4	20,000 (8000 in active)

Table 2 ACTFL, CEFR, and HSK correspondences as claimed by Mandarin House (ACTFL) and FaCH (CEFR) and their lexical minimums as established by Hanban

ACTFL	CEFR	HSK	Vocabulary
NL		Level 1	150
NM	A1.1	Level 2	300
IL	A1	Level 3	600
IM	A2	Level 4	1200
IH	B1	Level 5	2500
Advanced	B2	Level 6	5000

for ACTFL and CEFR. It is unclear if the suggested correspondences between ACTFL/CEFR and TORFL were based on empirical studies. The correspondences between the CEFR and ACTFL however, seem to be fairly consistent with the official ACTFL CEFR crosswalk (ACTFL, 2016).

The Chinese proficiency test Hanyu Shuiping Kaoshi (HSK) developed by the Office of Chinese Language Council International (Hanban), an organization associated with the Chinese Ministry of Education, specifies six levels of proficiency and provides both correlations to the CEFR scale and brief descriptions of what a learner can be expected to be able to do at each level. For example, "[T]est takers who are able to pass the HSK (Level III) can communicate in Chinese at a basic level in their daily, academic and professional lives. They can manage most communication in Chinese when travelling in China." (http://english.hanban.org/node_8002.htm#no1). While Hanban equates Level III with B1 in the CEFR, the German Association of Teachers of Chinese (FaCH) has resolutely questioned the correspondences established by Hanban and has instead proposed to equate HSK 3 with A1, HSK 4 with A2, and so on (http://www.fachverband-chinesisch.de/sites/default/files/FaCh2010_ErklaerungHSK_en.pdf). Evidence to support this equation comes from the crosswalk published by Mandarin House that uses the Hanban correspondences for the CEFR but very different ones for ACTFL (http://www.mandarinhouse.cn/images/general_chinese_program.pdf). Using the ACTFL CEFR crosswalk established by ACTFL (2016), Table 2 presents the vocabulary size requirements for each HSK level as proposed by Hanban for reading purposes and their putative corresponding ACTFL levels.

Table 3 General and specific notions identified for each level by the Instituto Cervantes's *Plan Curricular*

CEFR	General	Specific
A1	463	1146
A2	608	1584
SUBTOTAL A	1071	2730
B1	1567	3336
B2	2730	5541
SUBTOTAL B	4297	9100
C1	3976	5810
C2	4243	5676
SUBTOTAL C	8219	11,513
TOTAL	13.587	23.343

As Table 2 shows ACTFL IM appears to be associated with the 1000 band, IH with the 2000 band, and *Advanced* with the 5000 band.

The Instituto Cervantes has also developed a set of language proficiency tests for Spanish developed around the CEFR. Although there are no explicit lexical minima associated with particular levels of proficiency, the exams are designed to take into account the vocabulary lists specified for each level in the *Plan Curricular* published by the Instituto Cervantes. Following a notional-functional approach, the *Plan* specifies two lists of *nociones* (notions) per level, one labelled as 'general' and one as 'specific'. These *nociones* are lexical units that include individual words as well as collocations, idiomatic expressions and other phrases. Table 3 provides the overall count that was provided to us by Mr. García-Santa Cecilia, from the Instituto Cervantes (personal communication, June 5th, 2017), expected to be included in the curriculum at each CEFR level of instruction.

The vocabulary sizes expected for each level by the largest national cultural organizations representing the three languages in question vary considerably, and – as far as Russian and Spanish are concerned – are substantially larger than the ones established for English, especially for receptive (reading) purposes, namely, Russian 10,000 and Spanish 9100 (specific vocabulary) for B2 and Russian 20,000 and Spanish 23,000 for C2.

2.2 Vocabulary Learning in L2 Methods Textbooks

It is risky to make generalizations about what goes on in language classrooms without conducting systematic observations, which are beyond the scope of this study. It is the case however, that there is often a gap between empirical research and the implementation of its findings by practitioners. For example, it is probably reasonable to assume that most language instructors do not access primary research to inform their teaching. Therefore, in this section we examine textbooks on L2 teaching methodology because they reflect the profession's favored pedagogies and provide a good indication of the training that prospective teachers receive. Grabe (2009)

asserts that in order to learn vocabulary, students need a combination of "vocabulary instruction, vocabulary-learning strategies, extensive reading and word learning from context, heightened student awareness of new words, and motivation to use and collect words" (p. 283). While L2 teaching pedagogy has emphasized the importance of learning new words in context and the importance of extensive reading for vocabulary acquisition, the specific word-level learning strategies that Grabe mentions are conspicuously absent from the L2 methods textbooks used in most teacher training programs. In fact, many explicitly or implicitly avoid direct vocabulary instruction techniques.

All of the most commonly used textbooks on second language teaching methodology dedicate attention to the development of reading proficiency, sometimes combined with listening as the development of interpretive skills, but the emphasis is typically on the process of reading and a description of the tasks involved in that process (pre-reading, scaffolding, guided practice, etc.). A good example is Omaggio Hadley's (2001) *Teaching Language in Context*, perhaps the most influential and widely used methods textbook in the US for the past two decades. The author dedicates a chapter to developing proficiency in listening and reading and provides a rationale for focusing on the receptive skills as well as a number of techniques and ideas for activities. However, there is no mention of the role of vocabulary knowledge in reading comprehension and no specific suggestions of ways to help learners build their vocabulary. A similar focus can be found in Lee and VanPatten's (2003) *Making Communicative Language Teaching Happen*, another widely used textbook on language teaching methodology. The authors emphasize the role of comprehensible input in the acquisition of new vocabulary and encourage techniques that facilitate making form-meaning connections following Terrell's (1986) notion of binding. They also warn that memorization is "no substitute for meaning-bearing comprehensible input in learning vocabulary" (p. 37). In the same vein, Shrum and Glisan (2010) emphasize the importance of placing new lexical items within a meaningful context and engaging learners in collaborative work with peers to facilitate vocabulary acquisition. They also discuss the limitations of incidental acquisition through reading and acknowledge the need to provide students with opportunities for focused work on vocabulary, but they do not offer any suggestions as to what that focused work should look like.

Brandl (2008) places more emphasis on the importance of lexical acquisition by devoting an entire chapter specifically to the teaching and learning of vocabulary (which is separate from a chapter dedicated to the development of reading skills). In the introduction to the chapter, the author states that, "the learning and acquisition of vocabulary plays one of the most vital roles in becoming proficient in the target language" (p. 75). Brandl provides a variety of suggestions on how to use instructional resources and techniques that can be useful when presenting new vocabulary, with special emphasis on the advantages of using visual support such as realia and multimedia. However, as is the case with the other textbooks reviewed, his emphasis is on the presentation of new lexical items through meaningful, contextualized input.

Research Questions

To provide further evidence of the relationship between reading proficiency and vocabulary size and to make this relationship meaningful within the context of college foreign language study in the U.S., the following research questions were addressed.

1. What reading proficiency levels are attained by the study participants after how many years of college language study for these three languages?
2. How well does vocabulary size – measured as the receptive knowledge of various bands of the most frequent four to five thousand words of a language – predict reading proficiency levels as defined by ACTFL?
3. What ACTFL reading proficiency levels are predicted by what vocabulary sizes?
4. What are the differences, if any, between Chinese, Russian, and Spanish with respect to the relationship between vocabulary size and level of reading proficiency?

3 Methods

3.1 Participants

Participants in this study were college students of Chinese, Russian, and Spanish at a large Western US state university. The ACTFL Reading Proficiency Test (RPT) and the Vocabulary Levels Tests (VLT) were administered to a total of 155 students (Chinese: 46; Russian: 48; Spanish: 61) from the fall semester 2015 to the spring semester 2017. 61 students were female and 94 students were male. 15 students of Russian were enrolled in a second-semester course, 50 students were enrolled in a fourth-semester course (Chinese: 17; Russian: 7; Spanish: 26), 11 students were enrolled in a third-year course (Chinese: 9; Russian: 2), and 79 students were returned missionaries, who had spent between 18 and 24 months in a country where the target language was spoken and who were enrolled in an advanced language course. There was no random selection of students, and no attempt was made to match student characteristics across the three languages. Moreover, the extended immersion of the returned missionary students makes their language learning experience qualitatively and quantitatively different from other participants in the study. We present their data separately below.

3.2 Instruments

The ACTFL Reading Proficiency Test (RPT) is a standardized test for the global assessment of reading ability in a language (ACTFL 2013). The test measures how well a person spontaneously reads texts when presented with texts and tasks as described in the 2012 ACTFL Proficiency Guidelines. The test formats used in this

study consisted of 10–25 texts depending on a participant's proficiency level. There are five sublevels: Intermediate Low (IL), Intermediate Mid (IM), Advanced Low (AL), Advanced Mid (AM), and Superior (S). Each sublevel consists of five texts accompanied by three tasks (items) with four multiple-choice responses, only one of which is correct. Test specifications include genre, content area, rhetorical organization, reader purpose, and vocabulary (cf. ACTFL, 2013). Texts and tasks align at each level, for example, an *Intermediate* task requires understanding information that is contained in one sentence, whereas *Advanced* tasks require the ability to understand information that is spread out over several sentences or paragraphs. Tasks and multiple-choice responses are in the target language.

The RPT is a timed test with a total test time of 25 min per sublevel. Two sublevels are scored together, either the two levels taken or, if more than two levels were taken, the two highest levels that can be scored according to the specific algorithm of the test. Because there are no Novice texts or tasks, the Novice levels are determined according to how close the test-taker is to the Intermediate level. Test takers whose scores are below 33.33% of the maximum Intermediate score possible are rated NL, test takers whose score is between 33.33% and 50% are rated NM, and test takers whose scores are between 50% and 66.66% are rated NH. The test is Internet-administered and computer-scored (ACTFL, 2013).

The Vocabulary Levels Test (VLT) consists of a receptive and a productive test (Institute for Test Research and Test Development, 2013). It is modeled after the English Vocabulary Levels Test pioneered by Paul Nation (Nation, 1990). The VLT measures how many of the most frequent 4000 words of Chinese and 5000 words of Russian and Spanish are known. It consists of four to five bands: the most frequent 1000, 1001–2000, 2001–3000, 3001–4000, and 4001–5000 words. The receptive test, which was used in the present study, consists of ten clusters of six words each for each of these four or five bands. Each band is thus represented by 60 words. These words consist of 30 nouns, 18 verbs, and 12 adjectives and are chosen at random from the 1000 words of a band. Each cluster focuses on one part of speech. Three words of a cluster are targets, which need to be defined by choosing from a list of synonyms and paraphrases. The other three words are distractors.

The definition of receptive mastery of a particular band varies slightly in the literature. The two most common percentages used are 80% (e.g. Xing & Fulcher, 2007) and 85% (e.g., Schmitt, Schmitt, & Clapham 2001).

3.3 Data Coding

The RPTs used in this study were scored using either the Interagency Round Table (IRL) or the ACTFL scale.[1] ILR ratings were recoded into ACTFL ratings on the basis of raw scores according to the ACTFL algorithm. Following Rifkin (2005) and

[1] The RPT can be scored using the ILR or the ACTFL scale according to two different algorithms. The Chinese and Russian RPTs were originally scored using the ILR algorithm because the agency supporting the research requested ILR ratings. The Spanish RPTs were scored using the ACTFL rating.

others, ACTFL RPT results were coded numerically as follows: NL = 1, NM = 2, NH = 3, IL = 4, and so on, up to S = 10. VLT results were analyzed to determine if the words of a particular band (e.g., the most frequent 1000 words) were known. Three mastery criteria were investigated per language: 75% correct, 80% correct, and 85% correct (see below). The highest band at which students attained 75%, 80%, or 85% correct was considered their vocabulary level.

4 Results

4.1 Reading Proficiency

The results of the ACTFL Reading Proficiency Test (RPT) ranged from NL to S. Table 4 shows the distribution of the results by language and class level.

Table 4 shows that reading proficiency levels varied considerably across languages. While the median for fourth-semester Chinese students was 2 (NM), it was 3 (NH) for Russian and 5 (IM) for Spanish. The top 25% of students were at least 3 (NH) in Chinese, 4 (IL) in Russian, and 5 (IM) in Spanish. Returned missionaries had very high reading proficiencies in Russian and Spanish, while they scored below third-year students in Chinese. The median for Russian was 7 (AL) and for Spanish, it was 9 (AH). For Chinese, it was 3 (NH), lower than the 4.5 (IL to IM) for regular third-year students. The top 25% students were 10 (S) in both Russian and Spanish, while they started at 4.75 (IM) in Chinese, which was slightly lower than the third quartile of 5 (also IM) for third-year students. Of the 20 students returning from a mission in a Chinese-speaking country, only three were close to or at the *Advanced* level of reading proficiency (1 each at IH, AL, and AM) and many were NL ($N = 7$). While only one third-year student was AL, only one was NL, and most were IL or IM ($N = 4$). The median for second-semester Russian students was 1 (NL) and the top 25% were 3 (NH) or higher.

Table 4 Reading proficiency by language and class levels

	N	Min	Quartile 1	Median	Quartile 3	Max
Chinese	45					
Fourth semester	17	1	1	2	3	4
Third year	8	1	2.25	4.50	5	7
Returned missionaries	20	1	1	3	4.75	8
Russian	48					
Second semester	15	1	1	1	3	5
Fourth semester	7	1	2	3	4	6
Third year	2	5				6
Returned missionaries	24	1	5	7	10	10
Spanish	59					
Fourth semester	24	1	4	5	5	6
Returned missionaries	35	7	8	9	10	10

4.2 Vocabulary Scores

The maximum time allowed for the VLT was 25 min, approximately 5 min per band. Test takers, however, could spend as much time on any single band as they wanted. To determine the internal consistency of the various VLTs and to provide an overall reliability estimate, Cronbach's alpha was computed with the individual band scores as input. Cronbach's alpha is a measure of internal consistency, and it provides an estimate of the relationship between items. In this case, each band is considered an item. Cronbach's alpha examines how closely related these bands are, and if they can be considered to measure the same construct. In this sense, it can be considered to be a measure of scale reliability. Cronbach's alpha levels above 0.70 are considered to be acceptable levels. Table 5 provides Cronbach's alpha for the three languages. Table 5 also provides the correlations between each mastery criteria: 75%, 80%, and 85% correct and a composite score consisting of the summed individual band scores to determine the mastery criteria that correlates most strongly with the composite score. The maximum composite score for the five bands was 150 for Russian and Spanish and 120 for Chinese (4 bands only).

Table 5 shows that Cronbach's alpha was statistically significant and strong for all three languages, and particularly strong for Russian and Spanish, indicating high internal consistency and reliability of the three vocabulary tests. Table 5 also shows that the mastery criterion correlating best with the composite score was 80% for Chinese and Spanish. Because the mastery criteria 75% and 80% were almost identical in Russian, the criterion 80% was used for all three languages.

Table 6 shows the distribution of vocabulary levels using the 80% mastery criterion by language and class level.

Table 6 shows that second and fourth semester students generally did not yet have receptive mastery of the most frequent 1000 words of their respective languages. There were 2 out of 7 fourth-semester Russian students who had receptive mastery of the most frequent 1000 words, one out of 19 fourth-semester Spanish student who had a vocabulary level of 2000 words, and one fourth-semester Spanish student who had mastered the most frequent 4000 words.[2] Of the two Russian third-year students, only one had mastered the most frequent 1000 words.

Table 5 Cronbach's alpha computed between bands ($p < 0.05$) and Pearson's r correlations between composite vocabulary score and three mastery criteria: 75%, 80%, and 85%

	N	*alpha*	75%	80%	85%
Chinese	46	0.867	0.871*	0.892*	0.886*
Russian	48	0.951	0.960*	0.959*	0.923*
Spanish	54	0.950	0.957*	0.958*	0.934*

*Correlations were significant at $p < 0.01$.

[2] Students with prior exposure to the language are required to take a placement test, but there is no mechanism to exclude a student from enrolling in a level below their placement level. The instructor for this Spanish course confirmed that these two students seemed more advanced than fourth semester.

Table 6 Vocabulary levels by language and class level

	N	Min	Quartile 1	Median	Quartile 3	Max
Chinese	45					
Fourth semester	17	0	0	0	0	0
Third year	9	0	0	0	2000	4000
Returned missionaries	20	0	0	0	0	4000
Russian	48					
Second semester	15	0	0	0	0	0
Fourth semester	7	0	0	0	1000	1000
Third year	2	0				1000
Returned missionaries	24	0	1000	3000	5000	5000
Spanish	59					
Fourth semester	19	0	0	0	0	4000
Returned missionaries	35	3000	4000	4000	5000	5000

Both Russian and Spanish returned missionaries, however, returned with impressive vocabulary levels. The median for Russian was the 3000 band and for Spanish, the 4000 band. The top 25% of students in both Russian and Spanish had mastered the 5000 band. For Chinese, it was different. Sixteen out of 20 did not have receptive mastery of the most frequent 1000 words, 1 had mastered the most frequent 3000 words and 3 the most frequent 4000 words. This is most likely due to the fact that the VLT is a written test, and in the case of Chinese, it requires the ability to read Chinese characters. It is likely that the returned missionaries from China and Taiwan had much higher proficiency levels in speaking than in reading and that the VLT did not capture their oral vocabulary level, but only their written one.

4.3 Reading Proficiency and Vocabulary Levels

In the following, we present crosstabulations of vocabulary levels and reading proficiency and linear regression analyses to predict reading proficiency levels on the basis of vocabulary levels by language.

Table 7 shows that vocabulary levels of less than 1000 were associated with Novice and Intermediate levels and that the 4000 level appeared to be associated with the *Advanced* level (IH readers are almost at the *Advanced* level but inconsistently so).

A linear regression analysis was conducted to predict ACTFL ratings from the vocabulary score. Readers with proficiency levels below NH comprehend words and lists of words only and are unable to comprehend sentences or texts. Because we consider reading ability to involve textual understanding, these levels were excluded. Pearson's correlation between vocabulary score and reading proficiency was 0.843 with $p < .001$ ($N = 23$). The model explained 71.1% of the reading results ($R^2 = 0.711$). The linear regression analysis with reading proficiency as the dependent variable thus yielded a significant and large predictive effect of the

Table 7 Crosstabulation of Vocabulary and Reading Proficiency Levels – Chinese

		Reading proficiency								
		NL	NM	NH	IL	IM	IH	AL	AM	Total
Vocabulary level	Less than 1000	11	7	6	8	3				35
	1000					1				1
	3000					1		1		2
	4000						1	1	1	3
Total		11	7	6	8	5	1	1	1	41

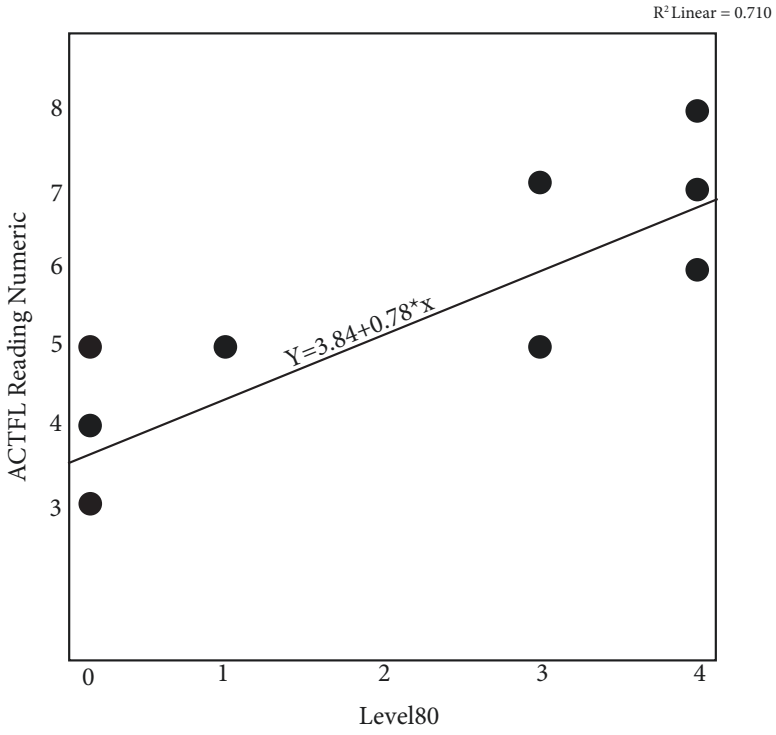


Fig. 1 Vocabulary size predicting reading proficiency levels – Chinese

vocabulary score on the reading proficiency rating: $p < .001$, Intercept (α): 3.84, Slope (β): 0.78. Figure 1 plots vocabulary and reading proficiency levels and includes the results of the regression analysis.

Table 8 (regression analysis) shows that vocabulary sizes of 1000 and 2000 predict the IM level in reading proficiency, while a vocabulary size of 3000 predicts IH and a vocabulary size of 4000 predicts *Advanced* levels of proficiency.

Table 9 shows that vocabulary levels of less than 1000 were associated with *Novice* and *Intermediate* levels. The 1000 level was associated with *Intermediate*,

Table 8 Predicting ACTFL reading proficiency levels on the basis of vocabulary size – Chinese

Vocabulary size	1000	2000	3000	4000
Reading proficiency numeric	4.62	5.40	6.18	6.98
ACTFL reading proficiency	IM	IM	IH	AL

Table 9 Crosstabulation of vocabulary and reading proficiency levels – Russian

		Reading proficiency									
		NL	NM	NH	IL	IM	IH	AL	AM	S	Total
Vocabulary level	Less than 1000	13	2	4	3	3					25
	1000				3		2				5
	2000					2			1		3
	3000					1		2			3
	4000						1	2		2	5
	5000							1	1	5	7
Total		13	2	4	6	6	3	5	2	7	48

the 2000 and 3000 levels with *Intermediate* and *Advanced*, the 4000 level with *Advanced* and *Superior*, and the 5000 level mostly with *Superior*.

A linear regression analysis was conducted to predict ACTFL ratings from the vocabulary score. Reading proficiency levels below NH were excluded. Pearson's correlation between vocabulary score and reading proficiency was .872 with $p < .001$ ($N = 33$). The model explained 76% of the reading results ($R^2 = 0.760$). The linear regression analysis with reading proficiency as the dependent variable thus yielded a significant and large predictive effect of the vocabulary score on the reading proficiency rating: $p < .001$, Intercept (α): 3.81, Slope (β): 1.06. Figure 2 plots vocabulary and reading proficiency levels and includes the results of the regression analysis.

Table 10 shows the results of the regression analysis.

Table 10 shows that vocabulary sizes of 1000 and 2000 predict the IM and IH levels, respectively, in reading proficiency, while vocabulary sizes of 3000, 4000, and 5000 predict AL, AM, and AH, respectively. The regression analysis thus supports the assumptions derived from the crosstabulation of vocabulary and reading levels.

Table 11 shows that that there were no students with reading proficiency levels below Intermediate.[3] Sixteen students who were Intermediate had a vocabulary level below 1000 words. A closer look at the data revealed that many of them had composite vocabulary scores that were similar to composite scores of the 1000 band in Russian (Spanish: Min = 34; Max = 81; Mean = 55; SD = 15.62; Russian: Min = 42; Max = 75; Mean = 59; SD = 12.98).[4] The student who had a vocabulary size of 2000 was one point short of receiving an IM rating. The 3000 and 4000 levels were mostly associated with *Advanced* and the 5000 level mostly with *Superior*.

[3] Two students who took the RPT were rated NL and NH, but neither of them took the VLT.

[4] Cf. Russian statistics for less than 1000: Min = 8; Max = 53; Mean = 33.84; SD = 15.42.

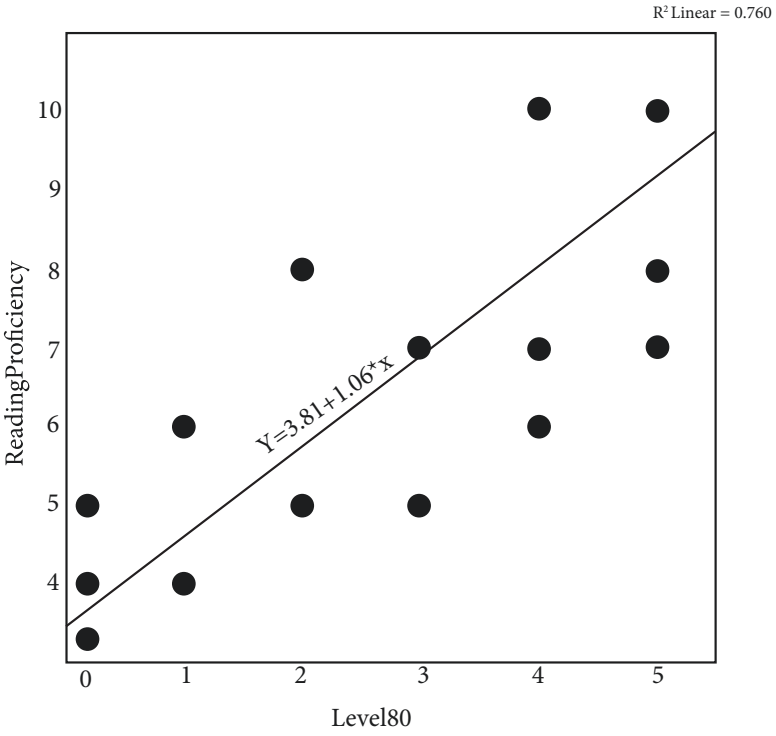


Fig. 2 Vocabulary size predicting reading proficiency levels – Russian

Table 10 Predicting ACTFL reading proficiency levels on the basis of vocabulary size – Russian

Vocabulary level	1000	2000	3000	4000	5000
Reading proficiency numeric	4.87	5.93	6.99	8.05	9.11
ACTFL reading proficiency	IM	IH	AL	AM	AH

Table 11 Crosstabulation of vocabulary and reading proficiency levels – Spanish

		Reading proficiency										
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S	Total
Vocabulary level	Less than 1000				6	8	2					16
	2000				1							1
	3000							2		1	1	4
	4000							5	5	6	3	19
	5000							1	1	5	5	12
Total		0	0	0	7	8	2	8	6	12	9	52

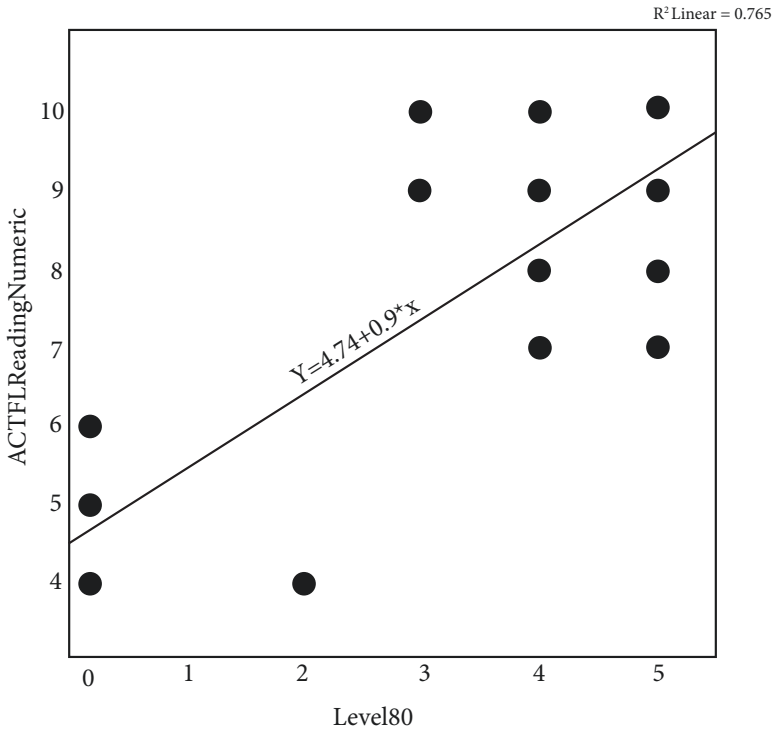


Fig. 3 Vocabulary size predicting reading proficiency levels – Spanish

Table 12 Predicting ACTFL reading proficiency levels on the basis of vocabulary size – Spanish

Vocabulary level	1000	2000	3000	4000	5000
Reading proficiency numeric	5.64	6.54	7.44	8.34	9.24
ACTFL reading proficiency	IH	AL	AL	AM	AH

A linear regression analysis was conducted to predict ACTFL ratings from the vocabulary score. There were no reading proficiency levels below NH. Pearson's correlation between vocabulary score and reading proficiency was .875 with $p < .001$ ($N = 52$). The model explained 76.5% of the reading results ($R^2 = 0.765$). The linear regression analysis with reading proficiency as the dependent variable thus yielded a significant and large predictive effect of the vocabulary score on the reading proficiency rating: $p < .001$, Intercept (α): 4.74, Slope (β): 0.90. Figure 3 plots vocabulary and reading proficiency levels and includes the results of the regression analysis.

Table 12 shows the results of the regression analysis.

Table 12 shows that vocabulary sizes of 1000 and 2000 predict the IH and AL levels, respectively, in reading proficiency, while vocabulary sizes of 3000, 4000, and 5000 predict AL, AM, and AH, respectively. While the reading proficiency levels

of the 1000 and 2000 bands appear to be inflated, probably due to the large number of below 1000 associated with intermediate reading proficiency, the results of the 3000–5000 bands are similar to Russian, and only slightly higher than Chinese. Note also that the 1000 and 2000 bands barely predict IH and AL reading proficiencies, being very close to the midpoint between IM and IH, and IH and AL, respectively.

5 Discussion

The present study showed very high correlations between reading proficiency and receptive vocabulary size (.843-Chinese; .872-Russian; .875-Spanish). Vocabulary size thus accounted for 71% (Chinese), 76% (Russian), and 76.5% (Spanish) of reading proficiency, and vice-versa, reading proficiency explained the same percentages of vocabulary size. These are very large effect sizes. Furthermore, crosstabulations and linear regression analyses showed that vocabulary sizes of the most frequent 1000 and 2000 words were generally associated with the ACTFL *Intermediate* level, while vocabulary sizes of 3000 and 4000 were associated with the ACTFL *Advanced* level. A vocabulary size of 5000 was associated with the ACTFL *Superior* level. Because student vocabulary sizes and reading proficiency levels were unevenly distributed between and within languages, it is even more remarkable that such a clear pattern emerged across three very different languages.

In general, the vocabulary sizes of the college students participating in the study were not very impressive. Even after 2 years of foreign language study, the vocabulary sizes of the students did not include even the most frequent 1000 words of their respective second language, particularly for Chinese and Russian. These low vocabulary sizes seem to be directly related to low reading proficiency ratings. The median after four semesters was NM in Chinese and NH in Russian with the top 25% of students reaching NH and higher in Chinese and IL and higher in Russian. While most Spanish students were solidly *Intermediate* at the end of 2 years, they had not quite mastered the first 1000 band, which may have precluded them from reaching higher proficiency levels.

Immersion learners (returned missionaries) who spent between 18 and 24 months in a country where the target language was spoken fell into two groups. Russian and Spanish students returned with impressive vocabulary sizes and high levels of reading proficiency. The Russian median was 3000 words, while the Spanish median was 4000 words. The top 25% of students of both languages had mastered at least 5000 words. This correlated with high reading proficiency levels. The median for Russian immersion learners was AL, for Spanish immersion learners, it was AH. The top 25% of students in both languages scored at the *Superior* level. For Chinese, it was different. 16 out of 20 of the immersion learners had not mastered the 1000 most frequent words of Chinese. As mentioned previously, this was most likely due to the fact that VLT was a written test, capturing only written vocabulary knowl-

edge. The other four had levels of 3000 and 4000, and possibly higher than 4000.[5] Overall, the median reading proficiency for this group was NH, lower than the median reading proficiency of regular third-year Chinese students, which was 4.50 (IL to IM). Thus, in general, third-year Chinese students were better in both reading proficiency and written vocabulary knowledge. The top 25% of third-year students had a reading vocabulary size of at least 2000 words.

The difference in reading level and vocabulary knowledge between the Russian and Spanish students on the one hand, and the Chinese students on the other, is striking. While the Chinese students had acquired fairly high levels of speaking proficiency while abroad, their literacy did not show similar development. Why this should be the case is beyond the scope of this study, but the explanation most likely lies in the difference in writing systems. The Russian and Spanish learners accessed their L2 through an alphabetic system as in their native language. The Russian learners did have the additional challenge of learning the Cyrillic alphabet, nonetheless, the essential principle of grapheme to sound correspondence remains consistent across the two alphabets. By contrast, the Chinese learners had to shift from their familiar alphabetic orthography to the character based Chinese writing system, the mastery of which requires extensive time and memorization. It is likely that the Chinese learners have greater vocabulary knowledge than was measured. The VLT is a written test. Had they been tested auditorially, they may well have scored better. The traditional third-year students had *Intermediate* levels of reading proficiency as well as higher vocabulary scores, results consistent with classroom learning that focuses on developing knowledge of Chinese characters.

6 Curricular and Pedagogical Implications

Upper division language curricula aim to introduce students to the literature of the target culture. As noted earlier, research has shown that a reader must know 95–98% of a text's vocabulary in order to understand the text. If the participants in this study are typical, and we focus in particular on students who did not have an extended immersion experience, we see that authentic literary texts are beyond the reach of many. Students do not have the necessary vocabulary knowledge to read at the *Advanced* or *Superior* level. The challenge for programs and instructors is how to promote vocabulary learning so that students achieve higher levels of reading proficiency.

Grabe (2009) makes a number of suggestions for vocabulary instruction based on findings from research on vocabulary acquisition. These include, for example, reading aloud to students and drawing their attention to keywords while reading; teaching a limited set of key words for depth, precision and multiple encounters; focusing on word relationships (parts-of-speech variations, word families, synonyms,

[5] The three students who had mastered the most frequent 4000 words had the three highest reading proficiency ratings (IH, AL, and AM), while the student who had 3000 words was rated IM.

antonyms, graded relations); working with dictionary definitions to rewrite more accessible definitions (pp 283–284). A number of other studies have pointed out the benefits of explicit vocabulary instruction, for example the effects of helping students recall and produce newly-learned words (Lee & Muncie, 2006; Lin & Hirsh, 2012; Webb, 2009). Strategies such as these appear to be at odds with the recommendations of the most frequently used L2 methods textbooks. As discussed above, the communicatively oriented ethos that underpins L2 methods textbooks privileges contextualized presentation of vocabulary with a focus on developing reading ability. For example, Lee and VanPatten (2003) caution against memorization in place of meaning-based comprehensible input, and yet, there is recent research to suggest that memorization may deserve reconsideration. A study comparing vocabulary learning via rote memorization on the one hand, and semantic mapping on the other, showed "no significant difference … between the vocabulary mean scores of the two groups on the post-test at the end of the four-month treatment period." (Khoii & Sharififar, 2013, 206). More empirical research on the efficacy of a variety of vocabulary learning strategies is needed.

7 Conclusion

The current study demonstrates that there is a strong correlation between vocabulary size and reading proficiency and it also shows that the level of L2 vocabulary that students acquire during a typical undergraduate program is not substantial. In its 2009 report to the Teagle Foundation, the Modern Language Association (MLA) acknowledged the crucial role of the study of texts for an undergraduate degree in languages and insisted that "the most beneficial among these are literary works, which offer their readers a rich and challenging—and therefore rewarding—object of study" (p. 4). As discussed in the introduction, a reader needs to know between 95% and 98% of the words in a text in order to comprehend it. This implies that access to original literary texts in the target language requires a vocabulary size that is probably beyond the reach of most undergraduate majors. The curricula of undergraduate degrees in languages typically include a number of required upper-level courses built around literary texts. The assumption is that advanced undergraduates should be able to engage in the critical reading of these texts and that exposure to them will improve their vocabulary and their command of the language. However, language learners face a disheartening conundrum: while incidental acquisition of a new lexical item requires multiple encounters with it, words beyond the most common 3000 are so infrequent that ordinary reading is not enough to learn them. Yet, explicit attention to vocabulary acquisition is often only a tangential focus of the introductory curriculum and may be altogether absent from advanced literary courses. If the ability to engage with literary texts is to remain a central goal of the language major, a renewed and revised attention to vocabulary building is necessary.

References

ACTFL. (2012). *ACTFL proficiency guidelines*. Alexandria, VA: ACTFL. Available from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

ACTFL. (2013). *ACTFL reading proficiency test* (*RPT*). *Familiarization manual and ACTFL proficiency guidelines 2012—Reading*. Retrieved June 8, 2016, from http://www.languagetesting.com/wp-content/uploads/2015/02/ACTFL_FamManual_Reading_2015.pdf

ACTFL. (2016). *Assigning CEFR ratings to ACTFL assessments*. Retrieved July 12, 2017, from https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

Bernhardt, E. (2011). *Understanding advanced second-language reading*. New York, NY: Routledge.

Brandl, K. (2008). *Communicative language teaching in action: Putting principles to work*. Upper Saddle River, NJ: Pearson.

Carver, R. P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior, 26*(4), 413–437. https://doi.org/10.1080/10862969409547861

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, NY: Cambridge University Press.

Hacking, J., & Tschirner, E. (2017). Reading proficiency, vocabulary development and curricular design: The case of college Russian. *Foreign Language Annals, 50*(3), 1–19. https://doi.org/10.1111/flan.12282

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689–696. Available at: http://nflrc.lll.hawaii.edu/rfl/PastIssues/rfl82hirsh.pdf

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

IELTS. (n.d.). *IELTS writing mark schemes*. Retrieved April 20, 2018, from https://www.examenglish.com/IELTS/IELTS_Writing_MarkSchemes.html

Institute for Test Research and Test Development. (2013). *Assessing evidence of validity of the ACTFL reading proficiency test* (*RPT*). Retrieved June 8, 2016, from http://www.languagetesting.com/wp-content/uploads/2013/10/Technical-Report-ACTFL-RPT-for-publication.pdf

Khoii, R., & Sharififar, S. (2013). Memorization versus semantic mapping in L2 vocabulary acquisition. *English Language Teaching Journal, 67*(2), 199–209. https://doi.org/10.1093/elt/ccs101

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30.

Lee, J. F., & Van Patten, B. (2003). *Making communicative language teaching happen*. Boston, MA: McGraw-Hill.

Lee, S. H., & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *TESOL Quarterly, 40*(2), 295–320. https://doi.org/10.2307/40264524

Lin, C.-C., & Hirsh, D. (2012). Manipulating instructional method: The effect on productive vocabulary use. In D. Hirsh (Ed.), *Current perspectives in second language vocabulary research* (pp. 117–148). New York, NY: Peter Lang.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York, NY: Heinle and Heinle.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening. *Canadian Modern Language Review, 63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 35–43). Cambridge: Cambridge University Press.

Nation, I. S. P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language, 26*(2), 1–16.
Omaggio-Hadley, A. (2001). *Teaching language in context*. Boston, MA: Heinle.
Rifkin, B. (2005). A ceiling effect in traditional classroom foreign language instruction: Data from Russian. *Modern Language Journal, 89*(1), 3–18. https://doi.org/10.1111/j.0026-7902.2005.00262.x
Schmitt, N. (2008). Review article. Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329–363. https://doi.org/10.1177/1362168808089921
Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing, 18*(1), 55–88. https://doi.org/10.1191/026553201668475857
Shrum, J. L., & Glisan, E. W. (2010). *Teacher's handbook: Contextualized language instruction*. Boston, MA: Heinle.
Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139–152. https://doi.org/10.1080/09571730802389975
Terrell, T. (1986). Acquisition in the natural approach: The binding/access framework. *The Modern Language Journal, 70*(3), 213–227. https://doi.org/10.1111/j.1540-4781.1986.tb05266.x
Tschirner, E. (2004). Breadth of vocabulary and advanced English study: An empirical investigation. *Electronic Journal of Foreign Language Teaching, 1*, 26–38. Available at: http://www.itt-leipzig.de/static/literatur/Tschirner_2004_Breadth_of_Vocabulary.pdf
Webb, S. (2009). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review, 65*(3), 441–470. https://doi.org/10.3138/cmlr.65.3.441
Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of vocabulary levels tests. *System, 35*(2), 181–191. https://doi.org/10.1016/j.system.2006.12.009

Jane F. Hacking is Associate Professor of Russian and Linguistics at the University of Utah, where she Co-Directs the Second Language Teaching and Research Center (L2TReC). Her research focuses on L2 phonology and the overall development of L2 proficiency. She received the 2017 award for Outstanding Contribution to the Profession from the American Association of Teachers of Slavic and East European Languages.

Fernando Rubio is Professor of Spanish Linguistics at the University of Utah, where he also serves as the Co-Director of the Second Language Teaching and Research Center. His research interests include second language acquisition, language teaching methodology, and proficiency assessment. From 2014 to 2018 he served as the PI in the Language Flagship Proficiency Initiative grant at the University of Utah.

Erwin Tschirner is Gerhard Helbig Professor of German as a Foreign Language at the University of Leipzig, Germany. His main research areas are language testing and second language acquisition, primarily the acquisition of vocabulary and grammar and reading and listening proficiency.

Picking Up the PACE: Proficiency Assessment for Curricular Enhancement

Dan Soneson and Elaine E. Tarone

Abstract This chapter describes a project at the University of Minnesota designed to improve the quality of language learning and teaching in seven different language programs through articulated coordination of three ongoing activities: annual administration of ACTFL proficiency tests of reading, speaking and listening to students at several points in the curriculum; a professional development (PD) program conducted by and for instructors of all languages, engaging them in exploratory practice to help students achieve higher proficiency outcomes; and a systematic program of proficiency-based self-assessment for undergraduate language students.

In this chapter, we provide details on each of these three initiatives all working together to raise the level of students' language proficiency. First, we analyze the proficiency assessment results of 1477 students enrolled in years 1–4 in seven language programs over the course of 2 years. Second, we describe a program of professional development for instructors that is conducted by and for instructors of all languages, the goal of which is to fine-tune instructors' delivery of the curriculum – to adjust their pedagogy and use of curricular materials to enable students to achieve higher proficiency outcomes. Finally, we provide an overview of a proficiency-based self-assessment measure for students to engage them in their own proficiency development.

Keywords Assessment · Professional development · Self-assessment · Proficiency · Flagship · Curriculum enhancement

D. Soneson (✉)
University of Minnesota Language Center, Minneapolis, MN, USA
e-mail: soneson@umn.edu

E. E. Tarone
University of Minnesota, Minneapolis, MN, USA
e-mail: etarone@umn.edu

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_4

As part of a national effort over the last 15 years to provide more U.S. citizens with high levels of proficiency in foreign languages that are critical for the nation's security, the Language Flagship Program has funded innovative post-secondary programs that integrate periodic language proficiency assessment into curriculum and instruction to produce graduates with superior levels of language proficiency (The Language Flagship, 2017). In this chapter, we focus on how the University of Minnesota has implemented the Flagship model of language proficiency development.

In 2014 the University of Minnesota was awarded a Language Flagship Proficiency Initiative grant "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" (The Language Flagship, 2014, p.1). At Minnesota, this grant would be used to implement language proficiency testing of its students at various stages of the curriculum in seven language programs. Testing in five languages (French, Portuguese, Russian, Spanish, and Arabic) would be funded by the grant, and parallel testing of students in German and Korean would be funded by the University's College of Liberal Arts (CLA).

The present chapter details the implementation of this initiative at the University of Minnesota in PACE: Proficiency Assessment for Curricular Enhancement, which was administered through the CLA Language Center. PACE was designed to improve the quality of language learning and teaching in seven language programs through articulated coordination of three components:

- Annual administration of ACTFL proficiency tests (as offered through Language Testing International, https://www.languagetesting.com) of reading, speaking, and listening at specific points in the curriculum;
- A professional development (PD) program conducted by and for instructors of all languages in the College of Liberal Arts;
- A systematic program of proficiency-based self-assessment for language learners.

PACE was implemented to establish at all levels a culture of proficiency assessment: professional development for instructional staff, collaboration among language programs, and self-assessment for students to enhance learning. Proficiency assessment was instituted to identify areas of strength as well as areas for improvement in language curricula; professional development provided means to support instructors' work on curriculum improvement and implementation; and establishment of systematic self-assessment throughout the language curriculum was meant to raise students' awareness of proficiency and increase their agency and responsibility for their own learning by developing and addressing realistic goals.

This chapter focuses on the first two and one half years of the three-year grant project, from the onset of the PACE project in August 2014 through December 2016. It begins with a short history of the role of language proficiency goal-setting and assessment in Minnesota. Next, we present and analyze 2 years of proficiency assessment results in reading, listening, and speaking for 1549 students at specific

stages in the language curriculum. We then describe a program of professional development for instructors that was aimed at improving the quality of their curriculum development and implementation and instilling a sense of community and collaboration. Finally, we provide an overview of a proficiency-based self-assessment instrument developed and implemented at the university as one means of sustaining the focus on proficiency beyond the grant period.

1 Review of Research

A considerable amount of scholarship has been devoted to the construct of language proficiency and the assessment of the proficiency levels attained by learners of different ages and in different social contexts. Previous research in university-level language programs has demonstrated the value of systematic ACTFL proficiency assessment through the use of ACTFL tests provided by Language Testing International, as referenced above (see also ACTFL, 2012) in order to better understand the degree to which university programs achieve their stated instructional goals. For example, Rifkin (2005) assessed listening, speaking, reading and writing skills both before and after more than 350 university-level students took part in the Middlebury Russian School's 9-week 'summer immersion' program. Students' proficiency levels upon entry to the summer program had been attained in traditional post-secondary classrooms. Those with a year of instruction had attained Novice High (NH) proficiency on the ACTFL (2012) proficiency scale in all four skills, with listening proficiency being the lowest, reading and speaking higher, and writing proficiency being the highest. The more years of classroom instruction the students had had, the higher their proficiency levels were upon entry to the summer immersion program. However, consistently their assessed receptive skills (listening and reading) were lower than their productive skills (speaking and writing). For example, after 4 years of college classroom learning, mean measured receptive skills were Intermediate Mid (IM), lower than measured productive skills at Intermediate High (IH). Interestingly, 9 weeks of immersion study at Middlebury often moved the students' receptive skills higher than their productive skills.

In another study, Davidson (2010) recorded ACTFL proficiency gains in Russian listening and reading during study abroad, also using the ACTFL (2012) proficiency scale. His database consisted of over 1200 reading assessments and 390 listening assessments. For those entering study abroad with two or 3 years of college study the mean listening score – IM – was lower than the group's mean reading score of IH. During study abroad, the students' reading scores all moved to higher levels (AL, AM, AH depending on length of study). However, their listening scores went up only to IH, with only a few 9-month study abroad students reaching an advanced level.

Tschirner (2016) gathered ACTFL listening and reading proficiency scores from over 3000 students studying seven languages at 21 postsecondary institutions in the U.S. Just as in the Rifkin and Davidson studies, Tschirner found a consistent pattern

in which listening proficiency seemed to develop more slowly in all these languages and contexts than reading proficiency.

There is need for more such studies on proficiency-oriented instruction and assessment, integrating the regular assessment of student proficiency levels with the ongoing process of language program curriculum development. To date, most studies have focused only on proficiency scores, without exploring the ongoing development of the language programs to address problems identified by assessment, or focusing on the role of instructor professional development or of student training in self-assessment in documenting the impact of assessment on curriculum and instruction in the form of washback or backward design (Wiggins & McTighe, 2005).

2 Language Proficiency in Minnesota

The PACE project was introduced at an important time in the history of language instruction at the University of Minnesota, which has long been a proponent of proficiency-oriented instruction and assessment. Its College of Liberal Arts (CLA) fully implemented a proficiency-based foreign language requirement relatively early, in Fall 1988, after 4 years of development and pilot testing. Language proficiency testing played a central role in the curriculum. Entering students who passed proficiency tests in French, German or Spanish at ACTFL levels of Intermediate Low for listening and reading, and Novice High for writing were placed into the second year of language instruction. To fulfill the language graduation requirement all CLA students were required to pass a second battery of four tests (including an oral interview) at ACTFL Intermediate High for listening and reading, and Intermediate Mid for speaking and writing (Eden, 1998; Lange, Prior, & Sims, 1992). This ambitious project established the University of Minnesota as a national leader in implementing a program of proficiency-oriented language instruction in which a set proficiency level was required for graduation, and high-stakes assessment played a pivotal role (Arendt, Lange, & Wakefield, 1986; Lange, 1988; Lange et al., 1992).

In 1993, this program was extended to a network of other institutions of higher education throughout the state through the Minnesota Articulation Project (MNAP), which was launched by a group of language professionals from the University of Minnesota, the state university system, private colleges, the state department of education, and the Minnesota Council on Teaching of Languages and Cultures. The goal of MNAP was to offer Minnesota students articulated programs of study, across elementary, middle and high school and college, in the most commonly-taught languages of French, German and Spanish, with the aim of fostering the attainment of higher levels of language proficiency (Metcalf, 1995; Tedick, 2002). Agreed-upon proficiency benchmarks at two levels were similar to those established in 1988 in CLA at the University of Minnesota. Led by Profs. Michael Metcalf and Dale Lange at the University of Minnesota and Suzanne Jebe in the state department of education, MNAP secured grant funding from the Fund for the Improvement of

Postsecondary Education, the National Endowment for the Humanities, and two USDE Title VI programs, IRSP (International Research and Studies Program) and LRC (Language Resource Center), administered at the Center for Advanced Research on Language Acquisition (CARLA) at the University of Minnesota. Over 50 language professionals from 23 post-secondary institutions, junior high and high schools participated in MNAP, divided into an assessment team, a curriculum team, and a political action team.

The assessment team, headed by Micheline Chalhoub-Deville at CARLA, began work in 1994, using the CLA entrance and graduation tests as a basis to develop new items and fully validated large-scale proficiency-oriented assessments called the Minnesota Language Proficiency Assessments, or MLPAs, for French, German and Spanish (see Chalhoub-Deville, 1997, for a review). MLPA Entrance Proficiency Tests (EPTs) were designed for graduating high school seniors intending to go into postsecondary institutions; if they passed the EPT at benchmark levels of Intermediate-Mid for receptive skills and Intermediate-Low for productive skills, they could be admitted into the second year of college-level language study. At the end of the second year of college-level language study, they could fulfill the CLA graduation requirement by passing MLPA Graduation Proficiency Tests (GPTs), designed with benchmarks of Intermediate High for listening and reading and Intermediate Mid for speaking and writing. The curriculum team, headed by Diane Tedick, was funded by the CARLA LRC to develop a curriculum handbook for language professionals, including exemplary lesson plans and units, to show teachers how to prepare their students to achieve the designated and assessed levels of proficiency (Tedick, 1998). By 2000, fully validated MLPAs were assessing all four skills in French, German and Spanish, at both the EPT and GPT benchmark levels. In 2001–2002 these MLPAs were computerized with funding from CARLA's LRC grant as well as substantial contributions from CLA language program instructors and the Testing Program in the CLA Language Center (CARLA, 2002, 2017).

However, the GPT was increasingly viewed by many students as stressful, time-consuming, and punitive, particularly when it delayed or outright prevented graduation. CLA began receiving growing numbers of complaints from those who failed the GPT, particularly when they had previously passed final exams in all their second-year language courses. In 2003, the dean of CLA took action in response to student and parent complaints to remove the GPT as a graduation requirement, and in Spring 2004 that test, renamed the Language Proficiency Exam (LPE), was relegated to become one of two options for fulfilling a CLA graduation requirement. The second option was completion of the fourth semester of a language sequence. Some language programs integrated the LPE into their fourth semester course, factoring test scores in with other course requirements, but many did not. Interestingly, from the very first semester following the change in graduation requirement, student pass rates on first takes of that same test plummeted, from 77–81% when it had been a graduation requirement to 37–55% when it was optional. At the same time LPE test completion rates dropped precipitously; incompletes (meaning, failure to complete all components) on the Spanish test, for example, went from 7% to 35.6% (Tarone & Lentz, 2008; Tarone, Lentz, & Eden-Frahm, 2009).

Despite CLA's elimination of the GPT as a graduation requirement in Spring 2004, both the EPT and the computerized LPE continued to be used in CLA for placement. In fact, by 2014, LPE use for placement testing for the original three languages had been expanded to other language programs, with un-validated versions being used for 13 other languages in CLA. At the inception of the PACE project in 2014, the CLA Language Center Testing Program employed three full-time staff who administered and coordinated placement and proficiency testing of more than 6000 students per year. Students could fulfill the graduation requirement either by passing the LPE, or by passing a fourth semester language course (several of which incorporated the LPE as a component factored into the final course grade). Also, students wishing to place into the third year of programs such as Spanish had to first pass the LPE.

In its efforts to promote language proficiency after 2004, CLA had moved from the stick (passage of a proficiency test as a graduation requirement) to the carrot (use of a test as a component in a program of supports and inducements to encourage a culture of proficiency). As part of this move, CARLA, the CLA Language Center, and the language teacher education program in the College of Education and Human Development had continued to provide high levels of professional development for language teachers focused on proficiency-oriented instruction. In 2013, the Spanish language program established a Certificate in Advanced-Level Proficiency in Spanish to encourage students to aim for higher tested proficiency levels. The certificate in Spanish served subsequently as a model for similar certificates in Chinese, French, and German. Requirements for all these certificates include proficiency self-assessment and ratings of Advanced Low or above on ACTFL proficiency tests in all four modalities. In 2013, the UMN Chinese language program was successful in securing funding to develop a Chinese Language Flagship program, which began operations in Spring 2014. It was in this historical context of CLA emphasis on language proficiency that Minnesota's Language Flagship Proficiency Initiative PACE project was established in 2014.

3 PACE Project

Prior to establishing the Language Flagship Proficiency Initiative in 2014, the federal Language Flagship Program had focused funding on a small set of universities in the U.S., each university using Flagship principles to develop and deliver one or more programs for one "critical" language – that is, a less-commonly-taught language deemed to be important for national security and defense. Flagship principles in these programs involved concentrated attention to the development of a very high level of language proficiency and cultural awareness, with the aim to bring students to the level of professional proficiency in that language (level 3 on the ILR scale, or Superior on the ACTFL scale). The Language Flagship Proficiency Initiative was designed to implement Flagship principles more broadly throughout the nation by funding a transformation of existing language programs at three already-strong

postsecondary institutions (Michigan State University, the University of Minnesota, and the University of Utah); the mission at each institution was to carry out large scale assessment of listening, speaking and reading proficiency levels of students in a selection of critical language programs as well as in two languages not identified as "critical," French and Spanish. At the University of Minnesota, the inclusion of French and Spanish, as well as German added locally, brought larger numbers of instructors into the project and greater visibility of the program throughout CLA. As mentioned above, the PACE Project housed in the CLA Language Center had three components: the administration of proficiency tests in seven language programs (Arabic, French, German, Korean, Portuguese, Russian, Spanish), a professional development program on proficiency-oriented instruction for language instructors, and a student self-assessment project.

This chapter reports on research during the first two and a half years of the PACE project on ACTFL assessment of speaking, listening and reading levels of students in the seven language programs; on the impact of professional development activities; and on the establishment of systematic student self-assessment, specifically:

1. What levels of proficiency do students achieve at which course levels in which languages?
2. Do students in higher level courses demonstrate higher levels of proficiency than students in lower levels?
3. Does a systematic program of professional development for all language instructors in the college:
   (a) establish a sense of community among language instructors from diverse programs?
   (b) contribute to a culture of assessment?
   (c) impact language instruction?
4. Does systematic self-assessment contribute to the development and maintenance of a culture of proficiency and proficiency assessment?

4 Proficiency Assessment of Seven Languages

In assessing language proficiency levels achieved by students at designated points in the seven language programs, the PACE project was able to include all students enrolled in all levels of the critical language curricula of Arabic and Russian, and second year and beyond in Korean and Portuguese. However, due to logistical, organizational and funding constraints, the project was not able to test all students in the larger language programs of French, German, and Spanish. (The fourth-semester Spanish course, for example, has over 20 sections with 25 students each.) For such larger language programs, where there were multiple sections of a course, the project tested at least two sections at each level of first- and second-year courses, as well as at least two, if not all sections of identified third-year courses and above.

In addition to testing all students in selected sections of these language courses, the PACE project included students returning from study abroad in France, Spain, and Ecuador, students in Spanish seeking a Certificate of Advanced Proficiency, and up to 75 students per year who volunteered to be tested. For the purposes of the project, any first- or second-semester course was considered "first year", any third- or fourth-semester course was considered "second year", and any fifth- or sixth-semester course was considered "third year". Only Korean offered a true first- through fourth-year sequence, and so the fourth year in the other languages was broadly interpreted to be any course beyond the sixth semester, including capstone courses. Students completing tests for the certificate were required to have taken at least two content courses taught in the language, so they were also counted among fourth-year students as well.

4.1 Method

Almost all the proficiency assessment instruments used in this study were developed by ACTFL and available through Language Testing International (LTI). For speaking, the project used ACTFL's Oral Proficiency Interview by computer (OPIc); for listening, its Listening Proficiency Test (LPT); and for reading, its Reading Proficiency Test (RPT). During the first year of the grant, LPT and RPT tests were not available for Arabic, so that year, that program used a computer adaptive proficiency test for listening and for reading developed at Brigham Young University (BYU), similar to the tests described by Clifford and Cox (2013) and Cox and Clifford (2014). Because that instrument was not available for Arabic the second year of the grant, the program used a newly available ACTFL RPT for reading, but opted not to test listening. For reading and listening the Korean program used the Test of Proficiency in Korean (TOPIK), which functions on a different scale than the ACTFL instruments, so TOPIK results will not be presented in this chapter. Table 1 summarizes the proficiency tests administered.

In addition to these proficiency tests, students also completed a self-assessment for speaking, reading and listening, and a survey providing information on their language background, their exposure to the target language, study abroad, motivation

Table 1 Instruments used to assess language proficiency in 7 languages in PACE

	Speaking	**Reading**	**Listening**
Arabic	ACTFL OPIc	ACTFL BYU/RPT	ACTFL BYU/--
French	ACTFL OPIc	ACTFL RPT	ACTFL LPT
German	ACTFL OPIc	ACTFL RPT	ACTFL LPT
Korean	ACTFL OPIc	TOPIK	TOPIK
Portuguese	ACTFL OPIc	ACTFL RPT	ACTFL LPT
Russian	ACTFL OPIc	ACTFL RPT	ACTFL LPT
Spanish	ACTFL OPIc	ACTFL RPT	ACTFL LPT

for learning the language, as well as how they currently used the target language. Institutional data for each student, including gender, age, year in school, and previous or current courses in any language taken at the University of Minnesota was also collected. The relationship between self-assessment and ACTFL ratings is discussed in detail in Chapter "Where am I? Where am I going, and how do I get there?: Increasing learner agency through large-scale self-assessment in language learning" in this volume.

During the first year (2014–2015) the PACE project divided the testing between the fall and spring semesters, and during the second year the majority of testing was completed in the spring semester. Table 2 shows that, in all, the PACE project tested a total of 1549 students at specific levels of the curriculum in the seven languages:

ACTFL tests for reading and listening consist of different versions that are valid only for specific level ranges. For example, Test B in any given language is targeted at Novice High to Advanced Low. This means that a learner must clearly demonstrate reading proficiency at Novice High or above in order to receive a valid rating on Test B. Should a student's proficiency exceed Advanced Low, Test B would not be able to provide a valid rating; a different test version would then be required. All students who were unable to clearly demonstrate a rating at or above the floor of the test version they took received a rating of "BR", standing for "Below Rating." This rating means that the student's proficiency could not be rated accurately and that it probably falls below the range of the test (how far below cannot be accurately determined). The OPIc is also targeted at specific level ranges of proficiency, where selection of a given range of testing is determined by the student in a short self-assessment built into the instrument. If students assess themselves too high or too low, the instrument is not able to provide a valid rating and raters return either a "BR" or "AR" for "above rating." If the speech sample is too short to be rated, a student could receive a "UR" for "unratable." In the results presented below, we indicate how many received BR, AR, or UR and do not include their results in averaging the ratings for the course level. The Center for Applied Linguistics (CAL) used this system to complete the analysis of the first year's data for the PACE project, and the CLA Language Center followed suit in the next 2 years of the project.

Table 2 Total number of PACE students tested with ACTFL Fall 2014–Fall 2016

	1st year	2nd year	3rd year	Abroad	4th year	Ind.	Totals
Arabic	114	99	12				225
French	82	100	86	6	29	15	315
German		104	23		31	11	152
Korean		87	24		12		123
Portuguese		61	16				77
Russian	39	68	17				124
Spanish	87	140	95	47	73	71	480
Total	**322**	**659**	**273**	**53**	**145**	**97**	**1549**

4.2 Proficiency Results

Research question 1 asks: What levels of proficiency do students achieve at which course levels in which languages? Average scores for each modality at each year in the curriculum were calculated on a weighted scale shown in Table 3. First developed by Liskin-Gasparro, and reported in Lange and Lowe (1987) and Liskin-Gasparro, Wunnava, and Henry (1991), this scale was used in all 3 years of the project.

Using the numeric rating scale in Table 3, the following tables present mean ratings for all students tested at each level in the curriculum for each tested modality from 2014–2016. Table 4 presents aggregate proficiency results for all included languages at each level in the curriculum, including mean proficiency ratings for listening, reading and speaking across the board. This includes both Roman-character and non-Roman-character languages. Tables 4, 5, 6, and 7 present mean proficiency ratings for each modality for each language program from the first through the fourth years of the curriculum.

In response to research questions 1 and 2, the above tables indicate that **in speaking**, students in the first year of the curriculum on average reach Intermediate Low, those in second year reach Intermediate Mid, those in third year reach Intermediate High, and those in fourth year score Intermediate High, approaching Advanced Low. **In reading**, students in first year on average reach Novice High, in second year Intermediate Mid, and in third year and above are in the Advanced Low range.

Table 3 Weighted scale showing numeric conversions of ACTFL proficiency ratings

ACTFL rating	NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Numeric equivalent	0.1	0.3	0.8	1.1	1.3	1.8	2.1	2.3	2.8	3.0

Table 4 Aggregate mean proficiency results for all PACE language programs 2014–2016

Students	Listening	Reading	Speaking	N
1st Year	NH (0.67)	NH (0.88)	IL (1.08)	321
2nd Year	IL (1.07)	IM (1.36)	IM (1.28)	654
3rd Year	IH (1.80)	AL (1.98)	IH (1.60)	246
4th Year	AL (2.04)	AL (2.15)	IH/AL (1.97)	148

Table 5 Mean **speaking** proficiency ratings by language & curriculum year in PACE 2014–2016

Language	1st year	2nd year	3rd year	4th year
Arabic	1.16 (IL)	1.14 (IL)	1.93 (IH)	
French	1.05 (IL)	1.35 (IM)	1.64 (IH)	1.97 (IH)
German		1.19 (IL)	1.37 (IM)	1.97 (IH)
Korean		1.19 (IL)	1.08 (IL)	1.82 (IH)
Portuguese		1.54 (IM)	1.95 (IH/AL)	
Russian	1.05 (IL)	1.29 (IM)	1.84 (IH)	
Spanish	1.00 (IL)	1.33 (IM)	1.59 (IH)	1.89 (IH)

Table 6 Mean **reading** proficiency ratings by language & curriculum year in PACE 2014–2016

Language	1st year	2nd year	3rd year	4th year
Arabic	0.82 (NH)	0.97 (NH/IL)	1.64 (IH)	
French	0.99 (NH/IL)	1.41 (IM)	2.05 (AL)	2.08 (AL)
German		1.13 (IL)	1.68 (IH)	1.80 (IH)
Portuguese		1.89 (IH)	2.12 (AL)	
Russian	0.55 (NM/NH)	1.02 (IL)	1.79 (IH)	
Spanish	1.01 (IL)	1.62 (IH)	2.01 (AL)	2.29 (AM)

Table 7 Mean **listening** proficiency ratings by language & curriculum year in PACE 2014–2016

Language	1st year	2nd year	3rd year	4th year
Arabic	0.80 (NH)	0.90 (NH)	1.17 (IL)	
French	0.83 (NH)	1.11 (IL)	1.88 (IH)	1.91 (IH)
German		1.05 (IL)	1.75 (IH)	2.02 (AL)
Portuguese		1.22 (IL/IM)	1.71 (IH)	
Russian	0.44 (NM)	0.92 (NH)	1.75 (IH)	
Spanish	0.59 (NH)	1.15 (IL)	1.82 (IH)	2.16 (AL)

Finally, **in listening**, first-year students average Novice High; those in second year, Intermediate Low; in third year, Intermediate High, and in fourth year and beyond, Advanced Low.

Tables 5, 6, and 7 can shed light on proficiency development in each of the language programs shown. For example, a reference point for the University of Minnesota is the stated outcome expectation for the end of the second year: Intermediate Mid in speaking and writing and Intermediate High in reading and listening for Roman script languages, and one sublevel lower for non-Roman script languages. For the skill of speaking, PACE data show those expectations for second-year speaking were largely met: Table 4 shows that second-year students on average are rated at Intermediate Mid in speaking, and Table 5 breaks the aggregate down into language programs, showing that most language programs except for Arabic, German and Korean (IL) were rated IM as expected for speaking.

In another example, the ACTFL test results shown in Tables 5, 6, and 7 reveal that Spanish students (the last line of Tables 5, 6, and 7), on average, reach the Spanish program's established learning goal for speaking of Intermediate Mid by the end of the second year, and that the mean rating in reading approaches Intermediate High. However, the data in these tables also show that listening proficiency lags behind both reading and speaking during the first 2 years of the curriculum. Specifically, students in first-year Spanish reach a proficiency on average of Novice High (.59) in listening, Intermediate Low (1.01) in reading, and Intermediate Low (1.00) in speaking. Likewise, in second-year Spanish, Tables 5, 6, and 7 show the mean listening proficiency rating is Intermediate Low (1.15), reading proficiency Intermediate High (1.62), and speaking is Intermediate Mid (1.33). Another point of interest in the tables is that, at Spanish students' graduating major level, in fourth year, both listening and reading are stronger than speaking profi-

ciency; they reach the Advanced Low level in listening (2.16) and Advanced Mid in reading (2.29), but their mean speaking proficiency is lower, at Intermediate High.

Interestingly, in all the languages we tested, during the first 2 years of language study, listening ratings are consistently lower than speaking ratings. This might indicate that listening proficiency develops more slowly than speaking proficiency, especially in the early stages. It might also reflect an emphasis on speaking proficiency in communicative language teaching, so that students might not receive as much exposure to listening texts and not have as much experience or many opportunities to engage with listening activities as they do with speaking or reading. Students in the third and fourth year of the curriculum post stronger listening ratings. One explanation may be that a large number of students do not continue beyond the second-year language requirement, so the population of students at this level is different than in the first 2 years. The level of motivation or interest among those in third year may be stronger than those who are primarily fulfilling a requirement, which may account for more time and effort spent experiencing the spoken language and thus a higher level of listening proficiency. Another explanation may be that third and fourth year students have had more opportunities for listening, either through immersion or study abroad opportunities, or through participation in courses with extensive listening opportunities, such as lectures or instructor-led discussions.

Of particular interest are the proficiency ratings of students learning German, which tend to mirror those in Arabic, Korean, and Russian in all three modalities in both second and third year. These ratings might indicate that German ranks closer to those languages in terms of time required to reach specific proficiency levels, than to the Romance languages (French, Portuguese, and Spanish).

Expanding the scope to include all language programs in the project (Arabic, French, German, Korean, Portuguese, Russian, and Spanish), and zeroing in on the end of the second year, or fourth semester, the relationship among the modalities is similar across all the programs. Figure 1 below shows the number of students rated at each proficiency level in the 3 skills of speaking, reading and listening, among all seven language programs at the end of the fourth semester**.** (The discussion below Fig. 1 will not include students who received a BR rating in listening and reading, nor those receiving UR and AR in speaking, since there is no accurate rating for them.)

Figure 1 illustrates that after 2 years, more than half of PACE students reach Intermediate Mid or higher in their proficiency in speaking and reading (418 of 656 speaking, 303 of 508 reading) (though there appears to be considerably more variation in reading proficiency scores than in speaking). However, in listening, at that same two-year mark, the bulk of student listening scores fall below, and even well below, that Intermediate Mid mark (293 of 440), with very few scoring higher. Thus, across all languages included in the project, at the end of 2 years in a language curriculum, mean student listening proficiency lags well behind mean proficiency in speaking and reading. These results are similar to those reported by Rifkin (2005).

In the more difficult non-Roman script languages (Arabic, Korean, Russian), Tables 5, 6, and 7 display results at the end of the fourth semester of study that are a bit lower than in the Roman script languages, as we might expect. Speaking tends to be strong, with just under half of the students' oral proficiency rated Intermediate Mid (IM) or above after four semesters. However, in these languages, both reading and listening tend to lag, with reading proficiency consistently skewing to Novice High – Intermediate

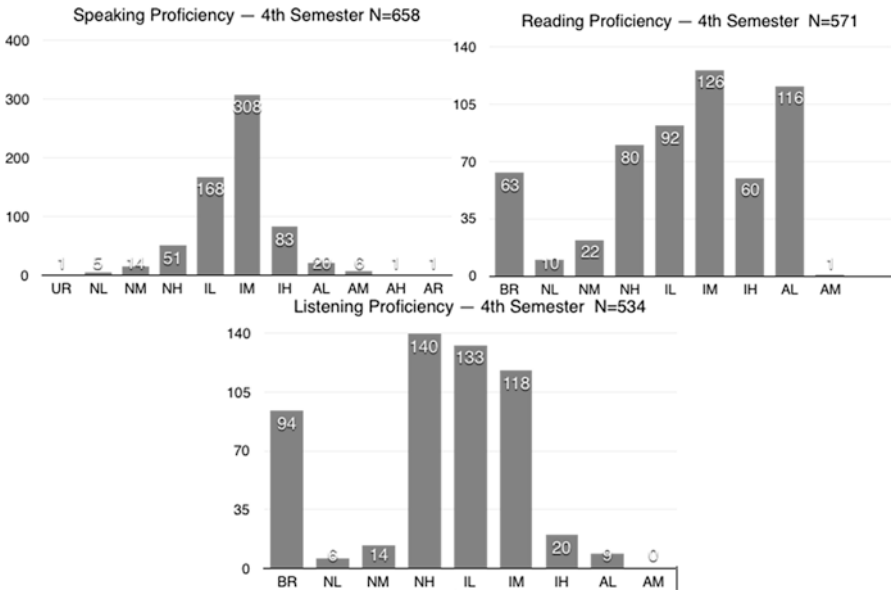


Fig. 1 Proficiency in PACE after four semesters in three skills for seven languages combined

Low after four semesters. Roughly 42% of the students in these languages were rated either Novice High or Intermediate Low, with only 26% rated Intermediate Mid or above after four semesters. About 11% had un-ratable tests (BR). Listening proficiency also skews to Novice High, with 40% rated Intermediate Low or above. Figure 2 below illustrates the number of students of languages with non-Roman scripts who, at the end of the fourth semester, were rated at each level for the 3 skills.

In the case of the non-Roman script languages, all students in fourth-semester Arabic, Korean, and Russian took the OPIc to assess their speaking skills. (As noted above, Korean students took the TOPIK for listening and reading, and are not included here.) Arabic students took the ACTFL-BYU listening test only in the first year of the grant. They were not tested in listening in 2016. (See Chapter "Arabic Proficiency Improvement through a Culture of Assessment" in this volume for detailed information about the Arabic program included in PACE.)

By the end of the PACE project, only French, German, Korean, and Spanish had language courses that extended beyond the third year. The PACE project tested students enrolled in advanced and capstone courses in these languages. In addition, students applying for the Advanced Proficiency Certificate were included in this group. Figure 3 illustrates the proficiency ratings of students of French, German, Korean and Spanish in the fourth year and beyond, identified in Fig. 3 as 8th semester.

Students at this level in the curriculum tend to exhibit reading and listening proficiency at the advanced level and above. Fully 79% are rated at Advanced Low or above in reading, and 69% have listening proficiency of Advanced Low or above. On the other hand, only 42% of students at this level taking the OPIc were rated at Advanced Low or above in speaking.

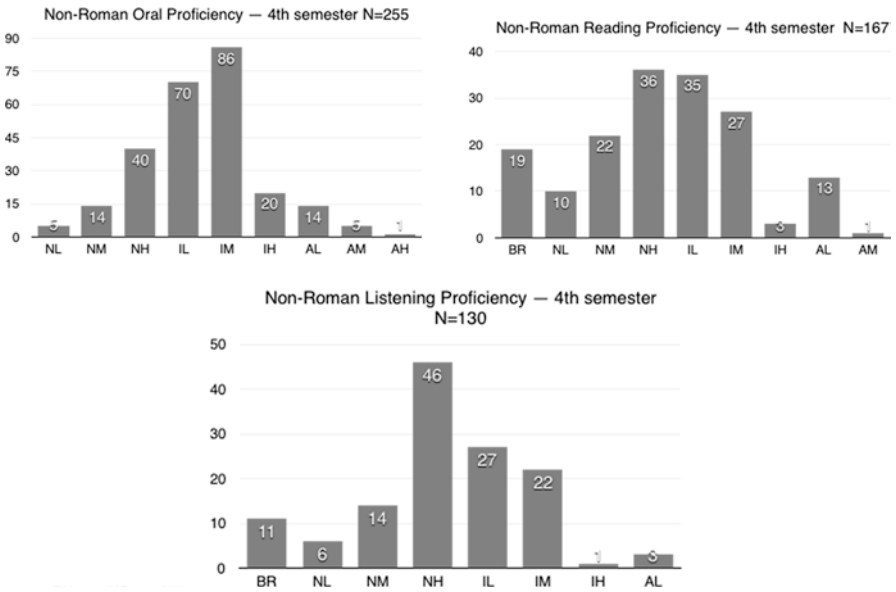


Fig. 2 Proficiency in three skills after four semesters for Arabic, Korean, and Russian in PACE

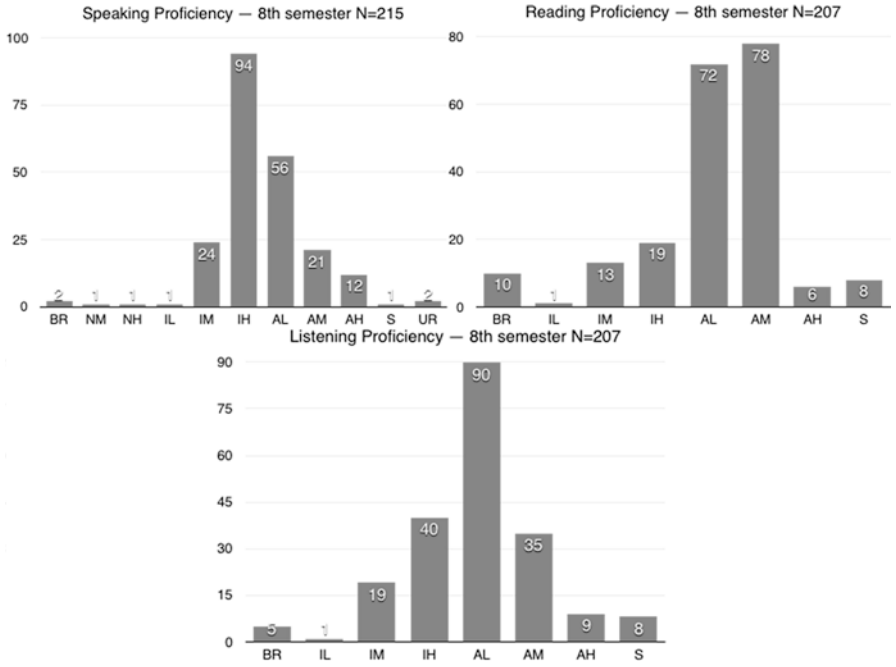


Fig. 3 PACE proficiency levels at fourth year and beyond, for French, German, Korean, and Spanish combined

5 Proficiency Growth Through the Curriculum

Research question 2 asks: "Do students in higher level courses demonstrate higher levels of proficiency than students in lower levels?" Close inspection of Tables 4, 5, and 6 shows that proficiency, on average, does indeed increase at higher levels in the curriculum. The visual display of these same results in Figs. 4 and 5 below makes it easier to see a relationship between program level and achieved proficiency.

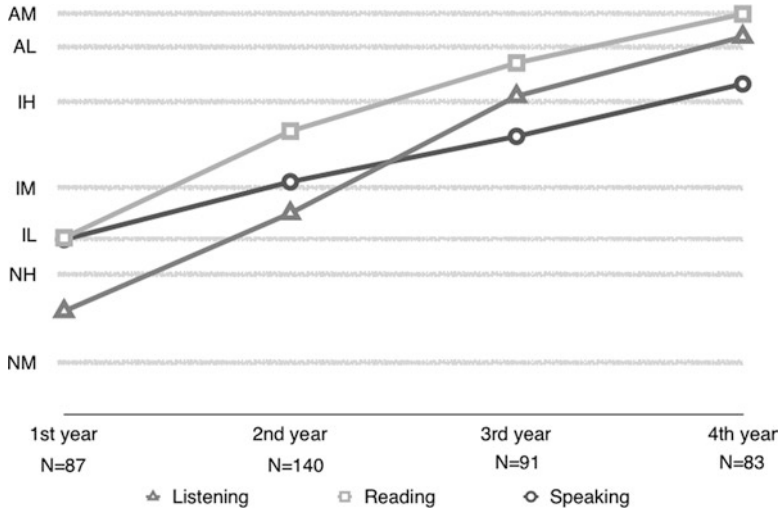


Fig. 4 PACE proficiency scores in three skills for four levels in the Spanish curriculum

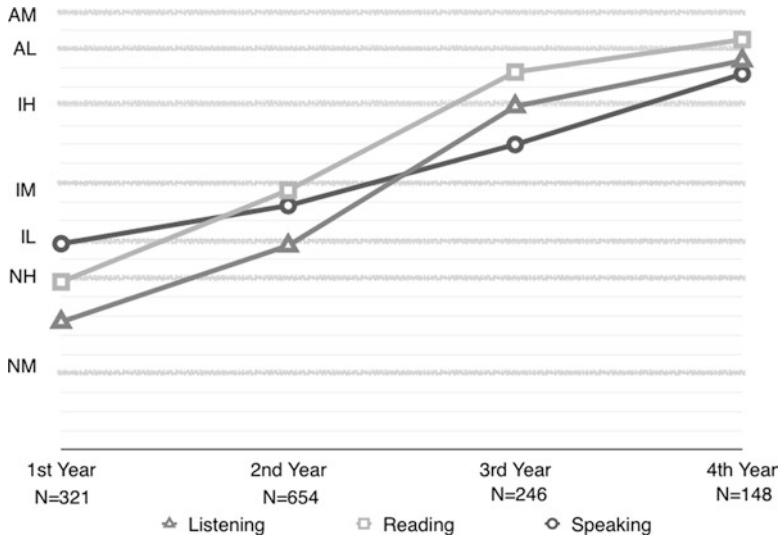


Fig. 5 Average proficiency ratings for all seven language programs 2014–2016

For example, Fig. 4 below illustrates the proficiency levels in the three skills in Spanish from lower to higher levels, year 1 through year 4. (In both the tables above and figures below, study abroad returnees are not included, because these students are at different levels in their Spanish learning experience and as such are difficult to place on a time line.)

Students in higher level Spanish courses do indeed demonstrate higher levels of proficiency than students in lower levels. The most dramatic increase is in listening. While listening lags behind reading and speaking in the first 2 years of the curriculum, mean listening proficiency is higher than mean speaking proficiency by the end of the third year. However, the increase in proficiency in all three modalities tends to level off after the third year.

A similar trajectory occurs in other language programs. Figure 5 shows the combined averages of proficiency in the three skills in each year of the language curriculum for all seven language programs between 2014 and 2016. Listening proficiency tends to be lower than speaking during the first 2 years of the curriculum, and slightly higher in the third year and above. Reading proficiency is lower in the first year of the curriculum and becomes higher than speaking proficiency during the second year, staying higher through the fourth year. All three are in the Intermediate High range on average in the fourth year.

To sum up this section, the regular administration of proficiency tests at key points in the language program did indeed tell us what proficiency levels the students in these seven language programs were reaching in the three skills in the different years of instruction; for example, we learned that proficiency in listening generally was lagging behind the skills of speaking and reading. Despite this disappointing result, we were reassured that overall, students in higher levels were reaching higher levels of proficiency. Unlike Minnesota's earlier initiative in the 1980s–1990s, which assessed only at the second-year level and only in the three languages commonly taught in high school, the PACE project's proficiency assessment component tested seven diverse languages at a wide range of levels from first-year to fourth-year students. It produced clear data across languages, skills and levels, and a shared terminology. Individual language programs can use this data as a basis for targeted curriculum revision.

6 Professional Development for Language Instructors

The third research question asks, "Does a systematic program of professional development for all language instructors in the college: a. establish a sense of community among language instructors from diverse programs? b. contribute to a culture of assessment? c. impact language instruction?" A major component of the PACE project was the establishment of a systematic professional development program for language instructors to help them make curricular changes in response to PACE assessment findings. Because change in curriculum centrally depends on the knowledge, skill, and actions of the instructors who design and deliver that curriculum in

response to assessment achievement and assessment goals, high quality professional development (PD) activities were offered to instructors throughout the PACE Project. The research questions for this component of the project involve the extent to which this component established a sense of community among language instructors from diverse programs, how it contributed to a culture of assessment, and how it impacted language instruction.

Increasing attention has been devoted to the central importance of integrating language proficiency assessment into program curricula and instructional practice. Indeed, research (e.g. McNamara, 2001; Wiggins & McTighe, 2005) suggests that in successful language programs, assessment should be tightly interwoven with all processes of teaching and learning throughout the curriculum. Based on these findings, the American Council of Teachers of Foreign Languages (ACTFL) has actively promoted the use of Integrated Performance Assessment which is based on "backward design" (Wiggins & McTighe, 2005) as one way to tightly integrate language assessment, teaching, and learning to improve student learning outcomes (Adair-Hauck, Glisan & Troyen, 2013).

Instructors typically need considerable professional development to learn to implement backward design in that it is a cyclical approach in which the instructor first "thinks like an assessor," identifying the desired end goal of instruction and the evidence to be used to indicate that those goals have been achieved, and only later "thinks like a teacher," developing classroom activities to enable learners to produce the end results. Such an assessment-first approach is very different from the traditional one that language instructors are used to, where they select learning activities first and design assessments later (Adair-Hauck et al., 2013). PACE's professional development program built on a strong foundation; as noted above, even after language proficiency assessment was dropped as a graduation requirement in CLA in 2004, a strong focus on proficiency-oriented professional development continued at CARLA, the College of Education and Human Development, and the CLA Language Center during 2004–2014. At the outset, the PACE project created a PD Peer Team consisting of a group of leading language instructors, representing each of the seven included language programs. The PD Peer Team was asked to help set objectives for and implement professional development, and provide feedback from the instructional staff to the PACE project leadership throughout the project.

Professional development has followed an arc that was initiated by the CLA Language Center in the year prior to the instantiation of the PACE project. Major workshops for all language instructors in CLA formed the basis of this professional development program, accompanied by frequent PD events in which instructors shared activities, approaches, and pedagogical ideas, including thoughts guiding revision of the curriculum. The centrality of curricular enhancement is represented in the very name of PACE, and the PD project was designed to function synergistically with the assessments and self-assessments to address needs and improve the learning experience by enhancing the curriculum. As such, the PD program was an integral part of the curriculum development cycle, which was: identify learning goals, design a curriculum to address the goals, assess the effect of the curriculum in reaching the goals, revise accordingly, and assess again (for a well-known presen-

tation of this cycle, see Graves, 2000). Since the program instructors were the ones charged to implement this curricular cycle, professional development was designed to help them learn to carry out this cycle effectively.

Prior to the onset of PACE, the CLA Language Center held two major workshops in 2014, one on designing student learning outcomes and one on using principles of backward design in curriculum development (Wiggins & McTighe, 2005). In the first year of the PACE project, major workshops dealt with language teachers' use of exploratory practice or action research (Allwright, 2005; Tarone & Allwright, 2005) in projects such as implementing the curriculum development cycle above, developing critical thinking through language courses, and the curricular design and implementation of ACTFL Integrated Performance Assessment (IPA) (Adair-Hauck et al., 2013). The second year brought workshops in developing advanced proficiency and using images to develop cultural awareness and critical thinking (Barnes-Karol & Broner, 2010). As a result of the 2 years of ACTFL testing, through which it became clear that listening proficiency in all language programs was not as strong as expected, particularly in the first 2 years, the second year of the grant closed with a day-long workshop on developing a principled approach to teaching listening skills. Instructors turned in a detailed evaluation of each workshop they attended, ranking its helpfulness and indicating what they had learned.

In addition to these major workshops, smaller events took place over the course of the grant project, with input from the PD Peer Team, including instructor presentations of activities, best practices, and implementation of technology in the language curriculum. Grant funds were also used to finance language instructor participation in one of several week-long CARLA summer institutes each summer. The PACE project also organized one four-day ACTFL Oral Proficiency Interview tester training workshop each summer. In 2015 ten instructors participated, and in 2016 two concurrent sessions took place, accommodating 20 instructors. In addition, small communities of practice formed to explore specific shared interests. For a detailed account of this type of teacher learning group, see Dillard, "Language Instructors Learning Together: Using Lesson Study in Higher Education" (this volume).

The evaluations of professional development activities that were submitted by the instructors show that these activities resulted in the establishment of a sense of community among language instructors from diverse programs, and a heightened awareness of proficiency and how to nurture it among instructors and students. Overall responses to the workshops organized by the PACE project on the most recent survey were positive, with 82% expressing that they were satisfied or very satisfied with the events during 2016–2017. For example, on the evaluation of the OPI tester training workshop we asked the question, "Has this workshop influenced you to consider changing in any way (e.g., objectives, methods, teaching practices, materials, etc.) your foreign language courses? Please elaborate." Responses include:

- "I will modify expectations to be more level appropriate and focus more on functions."

- "I learned you not only have to consolidate the lower level language skills, but also have to make them meaningful and spontaneous."
- "I'll try to modify my expectations, syllabi & grading grids according to the Guidelines."
- "The workshop helped me identify strategies for eliciting language that is appropriate to the level of students' proficiency."
- "I'm planning to change the curriculum of my courses to apply what I've learned. The courses I'm teaching are achievement-oriented courses but this workshop changed my mindset and I think I would change the direction towards to (sp) proficiency-oriented class as a degree that I can compromise between my ideal and my real situations at this point."

Instructors also said they experienced increased sharing of activities and approaches to language teaching, and a greater sense of community among language instructors across programs in CLA. Although most language programs are housed within one large building on campus, there had been very little interaction among instructors from different language programs prior to the PACE project. Instructors who completed the workshops commented that they were more comfortable interacting with each other in the building, including interacting more with instructors from other programs. Comments on workshop evaluations also indicated an appreciation for the developing sense of community among instructors, as the following comments illlustrate:

- "It was extremely interesting to me to hear about how other departments are doing things."
- "I really enjoyed the community building aspect of this day. I was happy to get to know instructors in other departments and really liked the opportunity to sit, talk, work with instructors in my own department."
- "I've gained a greater respect for colleagues in other language departments. I have been impressed with their curricular projects and feel I can turn to them for honest feedback and fruitful collaboration."

The workshops also fostered a culture of assessment among participant instructors, as evidenced by the subsequent development and incorporation of Integrated Performance Assessments (IPAs), particularly in the Spanish, German, Italian and French programs. Following a day-long workshop on the IPA, instructors expressed a positive reaction to the format and orientation of the IPA, pointing to positive impact on their students' learning: "It would be interesting for students," "make them better language users not just language learners," "get them more excited, Think more critically," "Try to make students more reflective about how they learn."

Finally, the professional development impacted language instruction, as evidenced by concerted efforts by language instructors to revise the curriculum in the light of proficiency test results. The PACE project funded language instructors to revise the curriculum and develop activities to address shortcomings evident in these results. For example, in 2016 instructors received summer funds and a course release to integrate principled listening activities into the Spanish curriculum at the

third- and fourth-semester level. Similarly, since listening proficiency ratings in French consistently fell below program expectations throughout the curriculum, the French program focused on listening in the fifth- and sixth-semester language sequence, and a French instructor received summer funds as well as a course release to develop listening activities for this level. Because ACTFL OPIc results revealed an issue with speaking proficiency in third-year Korean, an instructor of Korean was given funding to develop a proficiency-oriented approach to speaking in the Korean curriculum for this level. Finally, although students in the Arabic program received extraordinarily high ratings, particularly in the third year (see Chapter "Arabic Proficiency Improvement through a Culture of Assessment", this volume), attrition in the Arabic program leading to the sixth semester was so high that only six students completed the third year. Curriculum efforts on the part of the instructor thus went into developing an approach to differentiated instruction with the intention of retaining students through the sixth semester, with the result that in Spring 2017 twelve students had continued on to this level of instruction in Arabic – an increase of 100% over the previous spring. Students in all the redesigned courses are scheduled for testing at the end of Spring 2017. Instructors receiving funding for these curricular efforts have also folded their efforts back into the professional development program by presenting their projects publicly as a component of the PD program. A typical attendee response at the presentation of these projects is "It was helpful and interesting to see curricular development projects AND hear about the reasoning behind them."

To sum up, instructor responses to survey questions about the professional development opportunities in the PACE project show the growing sense of community among instructors from diverse language programs, a heightened awareness of proficiency and how to address it in the classroom, and a strong appreciation of assessment practices within the curriculum.

7 Student Self-Assessment Project

The fourth research question asks, "Does systematic self-assessment contribute to the development and maintenance of a culture of proficiency and proficiency assessment?"

The third major component of the PACE project focused on student self-assessment. Since the student self-assessment protocol is described in detail in the Chapter in this volume titled "Where am I? Where am I going, and how do I get there?: Increasing learner agency through large-scale self-assessment in language learning", here we will simply provide a broad overview of this component, the goal of which was to help students become informed agents in improving their own language learning processes and outcomes. Because a major aim of this project was to establish and maintain a culture of proficiency and assessment, the self-assessment component aimed to educate students about proficiency, foster realistic expectations for proficiency development, and promote their own agency in that development. To

what extent has the self-assessment component established a self-sustaining culture of proficiency and proficiency assessment? One way to answer this question is to track the number of participants engaged in self-assessment over the 2.5 years of the PACE project: Does the number of student participants and language program participation go up over time? The "Basic Outcomes Student Self-Asseessment" (BOSSA) protocol (detailed in "Where am I? Where am I going, and how do I get there?: Increasing learner agency through large-scale self-assessment in language learning)" formed an excellent foundation on which to build and sustain this culture and became one of the cornerstones of the PACE project. The BOSSA protocol was originally developed to help language learners in fourth-semester Spanish self-assess to develop realistic and achievable expectations of speaking proficiency as related to the fourth-semester course. A core development team, consisting of Sara Mack (Spanish & Portuguese), and Gabriela Sweet, Anna Olivero-Agney, Joanne Peltonen, and Diane Rackowski (CLA Language Center), developed a self-assessment protocol for this course that provided an opportunity for students to perform three speaking tasks and then reflect on and discuss their ability to do so, prior to completing an online self-assessment questionnaire. This protocol, first administered in Fall 2013, was extremely successful and was then adapted by the French, German, and Italian programs for implementation in fourth-semester classes in Spring 2014. In using this BOSSA, students assess themselves at the beginning of the semester to identify areas where they are strong and other areas that they can address throughout the semester. They then complete the same self-assessment protocol at the end of the semester as a reality check to see if they have improved an ability to perform specific types of oral tasks upon completion of the course. In the original protocol, the Spanish students also completed three follow-up reflections throughout the semester, to encourage them to monitor their learning.

The PACE project built on the success of the BOSSA protocol. Working with language instructors and students, CLA Language Center staff created similar self-assessment questionnaires targeting two wide proficiency levels in reading and listening, and created versions of the BOSSA protocol for speaking, targeting three additional proficiency levels: Novice High, Intermediate Low, and Intermediate High/Advanced Low. All PACE-tested students completed the three self-assessments: the BOSSA protocol for speaking, which includes an online self-assessment questionnaire, and the online reading and listening self-assessment questionnaires. In addition, they filled out a survey responding to questions about their previous experiences with the target language as well as with other languages, along with their motivations for learning the language and how they are currently using it.

The quantitative data gathered from these instruments was restricted to student responses on the questionnaires. Analysis of data led to revision of items on the questionnaires. Student responses were scored, producing an approximate rating, based on the ACTFL proficiency scale. Students received immediate feedback as to the ACTFL level corresponding to their level, along with a description of that proficiency level, both within the instrument and as an email message. They subsequently received follow-up emails, including suggestions for improving their proficiency.

(As mentioned above, Chapter "Where am I? Where am I going, and how do I get there?: Increasing learner agency through large-scale self-assessment in language learning" describes the BOSSA in more detail, and presents much of this data, along with qualitative student and instructor responses to the protocol.)

The PACE project has produced quantitative growth in the use of the BOSSA across language programs over time. Figure 6 shows the number of students who took the BOSSA self-assessment since its inception in Spring 2014 by program level. The BOSSA began with four programs at the fourth semester level. Through the PACE project, program involvement grew to encompass the first- through the eighth semester. Over all, between Spring 2014 and Spring 2017, a total of 14,354 students used the BOSSA. Figure 7 shows the growth in language programs employing the BOSSA each year, from four programs in Spring 2014 to ten programs in 2017. Figure 8 illustrates the growth in the number of students participating in the BOSSA protocol from its inception. Beginning with 628 students in Spring 2014, participation has grown to over 3136 in 2015–2016, with over 1800 completing the instrument in Spring 2017.

Although the reading and listening self-assessments have only been administered to students participating in PACE testing, the entire BOSSA protocol, including its speaking self-assessment questionnaire has received wide acceptance outside the PACE program, with ten language programs now using it regularly, making use of software developed with PACE funding at the CLA Language Center, and including over 50 instructors and 1400 students per semester. This wide acceptance of student self-assessment across language programs is a strong indication that a culture of proficiency and proficiency assessment is growing and making a difference, both among students and among instructors.

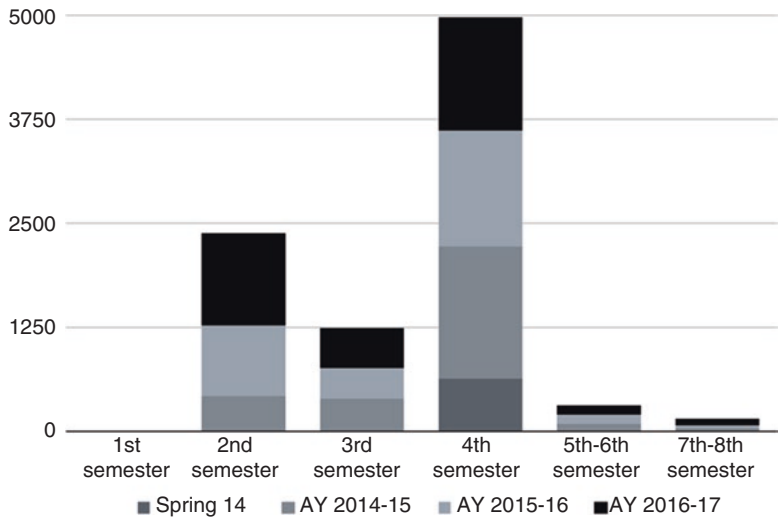


Fig. 6 Number of students per language program level per academic year using PACE BOSSA

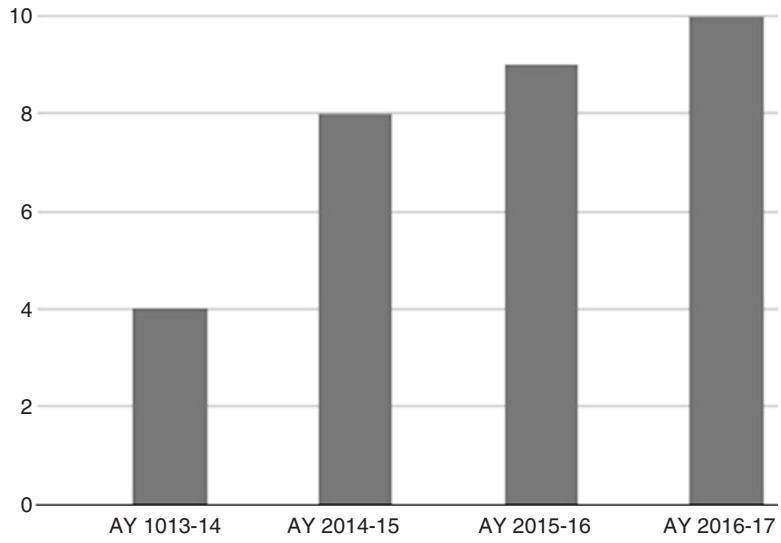


Fig. 7 Number of language programs using PACE BOSSA per academic year

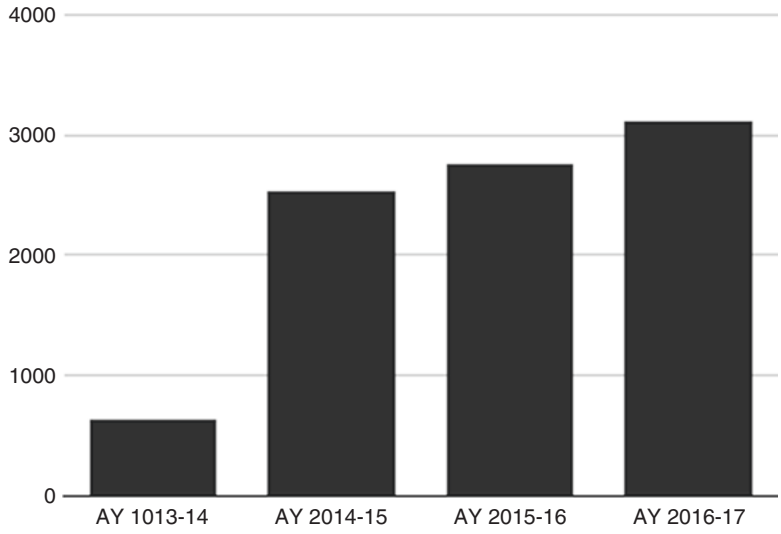


Fig. 8 Number of students self-assessing with PACE BOSSA per academic year

Student comments and instructor comments, detailed in "Where am I? Where am I going, and how do I get there?: Increasing learner agency through large-scale self-assessment in language learning", have been positive, viewing the experience as a "wake-up call," and as empowering for students. These self-assessment instruments are now integrated into the curriculum for many language programs. The speaking self-assessment will be used for 10 language programs in Fall semester 2017.

8 Conclusion

Three major components of the PACE program, administered by the CLA Language Center, have been implemented to reinforce the culture of proficiency and of assessment among the language programs at the University of Minnesota. Large-scale proficiency assessment has revealed increasing levels of student proficiency in speaking, listening, and reading at increasing levels of the curriculum in seven language programs. Consistencies among these results in relation to stated learning goals provide a common ground for language instructors engaged in exploratory practice to plan for and implement curricular revision in their programs. For example, testing revealed a lower level of listening proficiency than expected in the first 2 years of language curricula, and speaking proficiency among majors and those completing their undergraduate program was not as strong as desired. While some graduating students demonstrated speaking proficiency at Advanced Low or above, many did not, and the mean rating for these students was in the Intermediate High range. Based on these proficiency scores, efforts are now underway to address listening proficiency in the early stages of the language curriculum, and to begin to focus more clearly on speaking proficiency in the later stages.

Through professional development opportunities, instructors have a heightened awareness of proficiency characteristics and can identify proficiency levels and guide their students to a realistic expectation that supports their learning. Proficiency characteristics and goals are now discussed among students and instructors.

Student awareness of language proficiency has grown as a result of the testing process and self-assessment using BOSSA. Through the self-assessment process, students are more aware of how proficiency is defined, and have begun to understand their role in their language learning, gaining agency over the process and a sense of responsibility for their own learning. It is anticipated that the deepening focus of instructors on meaningful professional development, and the use of BOSSA to improve student awareness and agency will continue to support high levels of foreign language learning at the University of Minnesota long after the Flagship grant period has ended.

References

ACTFL. (2012). *ACTFL proficiency guidelines*. Alexandria, VA: ACTFL. Available from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Adair-Hauck, B., Glisan, E., & Troyan, F. (2013). *Implementing Integrated Performance Assessment*. Alexandria, VA: ACTFL.

Allwright, D. (2005). Developing principles for practitioner research: The case of exploratory practice. *Modern Language Journal, 89*(3), 353–366. https://doi.org/10.1111/j.1540-4781.2005.00310.x

Arendt, J., Lange, D., & Wakefield, R. (1986). Strengthening the language requirement at the University of Minnesota: An initial report. *Foreign Language Annals, 19*(2), 149–156. https://doi.org/10.1111/j.1944-9720.1986.tb03110.x

Barnes-Karol, G., & Broner, M. (2010). Using images as springboards to teach cultural perspectives in light of the ideals of the MLA Report. *Foreign Language Annals, 43*(3), 422–445. https://doi.org/10.1111/j.1944-9720.2010.01091.x

Center for Advanced Research on Language Acquisition (CARLA). (2002). *Manual for MLPA administration*. Minneapolis, MN: CARLA.

Center for Advanced Research on Language Acquisition (CARLA). (2017). *History of the MLPA [Minnesota Language Proficiency Assessments]*. Retrieved March 14, 2017 from http://carla.umn.edu/assessment/MLPA.html

Chalhoub-Deville, M. (1997). The Minnesota articulation project and its proficiency-based assessments. *Foreign Language Annals, 30*(4), 492–502. https://doi.org/10.1111/j.1944-9720.1997.tb00856.x

Clifford, R., & Cox, T. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals, 46*(1), 45–61. https://doi.org/10.1111/flan.12033

Cox, T., & Clifford, R. (2014). Empirical validation of listening proficiency guidelines. *Foreign Language Annals, 47*(3), 379–403. https://doi.org/10.1111/flan.12096

Davidson, D. (2010). Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals, 43*(1), 6–26. https://doi.org/10.1111/j.1944-9720.2010.01057.x

Eden, M. (1998). *German at the University of Minnesota: A case for articulation and accountability in a proficiency-based system*. Ph. D. dissertation, University of Minnesota.

Graves, K. (2000). *Developing language courses: A guide for teachers*. Boston, MA: Heinle & Heinle.

Lange, D. (1988). Some implications for curriculum and instruction for foreign language education as derived from the ACTFL proficiency guidelines. *Die Unterrichtspraxis/Teaching German, 21*(1), 41–50. https://doi.org/10.2307/3530743

Lange, D., Prior, P., & Sims, W. (1992). Prior instruction, equivalency formulas, and functional proficiency: Examining the problem of secondary school-college articulation. *The Modern Language Journal, 76*(3), 284–294. https://doi.org/10.1111/j.1540-4781.1992.tb06998.x

Lange, D. L., & Lowe, P. (1987). Grading reading passages according to the ACTFL/ETS/ILR reading proficiency standard: Can it be learned? *Selected Papers From the 1986 Language Testing Research Colloquium* (pp. 111–127). Monterey, CA: Defense Language Institute.

Liskin-Gasparro, J., Wunnava, P., & Henry, K. (1991). *The effect of intensive-immersive conditions on the acquisition and development of oral proficiency in Spanish and Russian.* (Grant No. P017A80033). Middlebury, VT: Middlebury College Language Schools.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333–349. https://doi.org/10.1191/026553201682430076

Metcalf, M. (1995). Articulating the teaching of foreign languages: The Minnesota Project. *ADFL Bulletin, 26*, 52–54. https://doi.org/10.1632/adfl.26.3.52

Rifkin, B. (2005). A ceiling effect in traditional foreign language instruction: Data from Russian. *The Modern Language Journal, 89*(1), 3–18. https://doi.org/10.1111/j.0026-7902.2005.00262.x

Tarone, E., & Allwright, D. (2005). Language teacher-learning and student language-learning: Shaping the knowledge base. In D. J. Tedick (Ed.), *Second language teacher education: International perspectives (5-23)*. Hillsdale, NJ: Lawrence Erlbaum Publishers.

Tarone, E., & Lentz, U. (2008). *Student performance: When the proficiency test becomes optional*. Presentation at the annual conference of the American Association of Teachers of Foreign Languages (ACTFL), Orlando, FL.

Tarone, E., Lentz, U., & Eden-Frahm, M. (2009). *Student performance: When the proficiency test becomes optional*. Presentation at the annual conference of the Minnesota Council of Teachers of Languages and Cultures (MCTLC), Minneapolis, MN.

Tedick, D. J. (Ed.). (1998). *Proficiency-oriented language instruction and assessment: A curriculum handbook for teachers.* Minneapolis, MN: Center for Advanced Research on Language Acquisition, University of Minnesota. Revised and updated, January 2002. Available online at: http://www.carla.umn.edu/articulation/handbook.html

Tedick, D. J. (2002). Background information. In *Proficiency language instruction and assessment: A curriculum handbook for teachers.* Minneapolis, MN: Center for Advanced Research on Language Acquisition, University of Minnesota. Available online at: http://carla.umn.edu/articulation/background.html

The Language Flagship (2014). *Request for proposal: The Language Flagship proficiency initiative application guidelines*. Retrieved from https://www.thelanguageflagship.org/sites/default/files/Flagship%20Proficiency%20Initiative%20_%202014_0.pdf

The Language Flagship (2017). *The Language Flagship: Creating global professionals*. Retrieved 10 June 2017 from https://www.thelanguageflagship.org/

Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals, 49*(2), 201–223. https://doi.org/10.1111/flan.12198

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Dan Soneson is Director of the Language Center in the College of Liberal Arts at the University of Minnesota Twin Cities and the Principal Investigator of the Proficiency Assessment for Curricular Enhancement (PACE) project, a four-year federal grant within the proficiency initiative sponsored by the Language Flagship. He has co-edited volumes on Language Teacher Education and Cultures and Languages Across the Curriculum, published as working papers by the Center for Advanced Research on Language Acquisition (CARLA) at the University of Minnesota, and has also published on language curriculum innovation and assessment. Other interests include self assessment and implementation of technology in second language education.

Elaine E. Tarone is Distinguished Teaching Professor Emerita, University of Minnesota-Twin Cities. She is widely published in the area of second language acquisition research, on topics like learner language analysis, interlanguage variation, the impact of literacy level on oral L2 processing, and language play. A recent interest is learners' enacted bilingual voices in oral narrative. She served as Director of the Center for Advanced Research on Language Acquisition (CARLA) 1993–2016.

Assessment and Curriculum for Heritage Language Learners: Exploring Russian Data

Olga Kagan and Anna Kudyma

Abstract This study analyzes data from the questionnaires and online tests administered to 94 heritage speakers of Russian who took placement tests at UCLA between 2013 and 2016.

The online test assesses all four skills, allowing an evaluation of learners' grammatical competence, knowledge of vocabulary, and ability to handle pragmatics. We make recommendations for curricular design based on the integrative nature of the test.

This study focuses on second-generation students, namely, those born in the United States to at least one Russian-speaking parent, as these students comprise over 50% of our test-takers. We also illustrate the findings by closely analyzing the background and performance of four second-generation students who are representative of the range of heritage language learners in our program.

While this study is based on Russian data, we anticipate that its conclusions will apply to other heritage languages and heritage language programs, in particular to less-commonly-taught languages.

Keywords Heritage students · Heritage language learners · Second-generation · Placement test · Assessment · Curriculum

1 Introduction

In the past 30 years, American secondary and postsecondary educational institutions have seen an increasing number of heritage language speakers, defined here as students who speak languages other than English at home and are to a degree bilingual in both English and the home language (Kagan & Dillon, 2006; Polinsky &

O. Kagan · A. Kudyma (✉)
Slavic, East European and Eurasian Languages and Cultures, University of California, Los Angeles, CA, USA
e-mail: akudyma@ucla.edu

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_5

Kagan, 2007; Valdés, 2000). Due to their initial and varied exposure to language at home, they typically exhibit a range of proficiencies in speaking and listening and may or may not be literate in their home language.

The UCLA Russian Program offers classes for heritage language (HL) learners. As a result, the program needs to place heritage language speakers accurately so that they will advance in their competency of Russian. Since 2013, the program has been offering a Russian online placement exam for all students, both L2 and HL learners, who seek to satisfy their language requirement or place into a class that is appropriate for their proficiency level. The majority of students who have taken the online placement test have been HL speakers. Based on the test results, HL learners without literacy or with low literacy abilities enroll in "Literacy in Russian," a beginning sequence of courses for Russian HL learners. Students with higher proficiencies take content-based courses, such as "Russian History" and "Russian Cinema." While the beginning sequence is strictly for heritage learners, content-based courses bring together HL and L2 learners at Intermediate High or higher proficiency levels. Our goal is thus for HL learners to be able to take content-based courses after 1 year of basic HL instruction.

This study focuses on an integrated test that assesses all four skills, allowing an evaluation of learners' grammatical competence, knowledge of vocabulary, and ability to handle pragmatics. Based on the integrative nature of the test, we make recommendations for curricular design. We focus on second-generation students, namely, those born in the United States to at least one Russian-speaking parent, as these students comprise over 50% of our test-takers. We also illustrate the findings by closely analyzing the background and performance of four second-generation students who are representative of the range of heritage language learners in our program.

2 Review of Literature

While much research on heritage language teaching has become available over the past decade (Schwartz Caballero, 2014), literature on heritage language assessment is not plentiful. Polinsky and Kagan (2007) suggested a three-component testing procedure consisting "of (1) an oral test loosely based upon the ACTFL oral proficiency interview; (2) a short essay (if the learner is literate in the heritage language); and (3) a biographic questionnaire" (p. 387).

The use of oral proficiency interviews (OPI) for HL speakers has been controversial. For example, Valdés (1989) suggested that using OPI may not be appropriate since OPI testers may unfairly penalize speakers of non-standard language varieties because "these levels were described for the foreign language learner, that is, for the traditional student of foreign languages in this country who begins his/her studies at point zero." (p. 395). Valdés's arguments may therefore have deterred many practitioners and researchers from using OPIs and ACTFL or ILR Guidelines to assess

oral proficiencies of HL learners; however, an investigation of oral proficiency of Russian HL learners for placement purposes (Kagan & Friedman, 2003) led to the conclusion that OPI use is justified, at least in the case of Russian. Kagan and Friedman (2003) also concluded that HL learners without literacy may display an intermediate or higher proficiency in speaking and listening. Sohn and Shin (2007) showed that Korean heritage speakers frequently demonstrate advanced proficiency on the ACTFL scale in speaking/listening, while they may have no functional proficiency in reading or writing.

So far, no studies have examined listening and reading proficiencies of HL speakers. We assume that listening is their strongest skill as this is typically how HL speakers assess their own abilities. Carreira and Kagan (2011) conducted a study of heritage language learners of 22 languages and collected data on their use of the heritage language, motivation for maintaining and advancing their knowledge of the language, and a self-assessment of their abilities in the four skills. Sixty-eight percent of all respondents ($N = 1732$) indicated that they were advanced or close to native speakers in listening comprehension, and 44% considered themselves advanced or native-like in speaking; in comparison, a much smaller number (27% and 19% respectively) answered that their reading and writing proficiencies were advanced or native-like.

While these evaluations are informative, reliance on self-assessment alone does not allow for robust measurements. To provide comprehensive information, Fairclough (2012) suggested that a placement test for HL speakers adhere to the following guidelines:

> Designed using a multifaceted approach that attempts to cumulatively measure the following areas: (a) receptive, such as knowledge of general vocabulary; (b) productive, with a focus on linguistic gaps, dialectal forms, and language transfer; and (c) creative, which includes speaking and writing abilities reflecting a range of functions and contexts. (p. 126)

One also needs to consider the institutional context. Discussing the placement test for Spanish heritage learners Fairclough elaborates: "[b]efore making decisions about the content and design" of a placement test, "test users must consider some very important issues, namely the relationships between language testing and (a) the mission of the program, (b) program/student characteristics, and (c) course content." (Fairclough, 2012, p. 124).

Based on our review of literature, we determined that an integrated and multifaceted testing instrument that leads to a thorough needs-analysis best serves HL learners. As we will show in this chapter, the results of our test confirmed that each HL learner's linguistic profile presents a complex picture of skill levels and competencies.

The research questions that we sought to answer in this study were as follows:

1. Why does the placement test for HL learners need to integrate all four skills and subskills such as grammar, vocabulary and pragmatics?
2. What are particular strength and weaknesses of second-generation learners?
3. How can the test determine the curriculum for HL learners?

3 Participants

Ninety-four Russian-speaking HL learners took the placement test over a 3-year period (2013–2016). Before taking the test, we instructed all students to fill out a background questionnaire, but only 81 students of the total 94 completed it. Out of the 81 students who filled out the background questionnaire from 2013–16, 35.8% ($n = 29$) belong to 1.5 generation. Fifty point six percent ($n = 41$) are second generation, and 13.6% ($n = 11$) came to the United States at age 15–18, and therefore belong to the first generation. We used the classification Rumbaut, Massey, and Bean, (2006) proposed, which involved the following distinctions between immigrant generations: "1.5 generation" if they came to the United States to live before the age of 15; "2nd generation" if they were born in the United States and had at least one parent who was foreign-born; and "3rd + generation" if both they and their parents were US-born but had one or more foreign-born grandparents." (p. 450)

3.1 Russian Heritage Learners

The most recent mass immigration to the United States from Russian-speaking countries[1] started in the mid-1970s. It peaked in the mid-1990s and has abated since then. The 2011–2015 Community Survey (U.S. Census Bureau, 2011–2015) listed 914,000 people who reported speaking Russian at home. For a detailed discussion of Russian immigration, see Andrews (1998), Isurin (2011), Kagan and Dillon (2010), and Zemskaya (2001).

Kagan and Dillon (2006) proposed that the level of competency in Russian of the children of the recent immigrants is directly tied to the amount of education they had received prior to immigration. However, for second-generation HL learners their level of competency in Russian depends on language socialization, which may include both home and classroom (He, 2016).

3.2 Sociolinguistic Background of the Test Takers

Eighty-one Russian HL test takers in the study completed the background questionnaire, which solicited responses about students' place of birth and age at immigration, use of Russian at home, and attendance of a Russian medium school or a community school or other instruction in the language. Other questions probed

[1] Russian-speaking immigrants originate not only in Russia but may come from the former Soviet republics. For many of them Russian is their native or dominant language whether they are ethnically Russian or not.

motivation for taking Russian in college and asked for self-assessment of linguistic competencies in the four skills.

3.2.1 Place of Birth

As previously mentioned, of the 81 students who filled out the questionnaires, 50.6% were born in the United States to immigrant parents, one or both of whom speak Russian as a native/first language. The families of our test takers came from Georgia, Ukraine, Armenia, Belarus, Kazakhstan, Lithuania, Uzbekistan and other countries that used to be republics of the former Soviet Union. The questions about languages spoken by parents revealed that in addition to Russian, families spoke other languages of the former Soviet Union, such as Armenian, Ukrainian, Belorussian, Lithuanian, as well as non-East European languages such as Italian and Spanish.

3.2.2 Self-Reported Language Use and Competence

Forty-nine percent of all test takers reported using Russian every day, typically with parents and grandparents, and 30% reported using it sometimes. Five percent went to a high school in Russia or another former Soviet Union country. Twelve percent went to a Russian-speaking preschool, 6% to church or Saturday school, 38% were taught to read and write by family members, and 4% had instruction from a private tutor. Of those who went to a Russian-medium school in the United States or abroad, 41% attended for less than a year.

To self-assess their proficiencies in the four skills areas, students used the scale from 0 to 5, with 5 being the highest. Forty-nine percent rated themselves at 5, and 35% at 4 in listening; in speaking, 25% rated their proficiency at 5, 31% at 4, and 32% at 3. The results for reading were as follows: 4% indicated they could not read, the majority range in reading was between 1 and 2 (55%) and between 3 and 4 (36%); 17% indicated they could not write and the majority self-assessment range in writing was 1–3 (57%). See Table 1 below for more information.

Table 1 Self-assessment: All students ($N = 81$)

Skills/Scale	0	1	2	3	4	5
Listening	0%	16%			35%	49%
Speaking	0%	12%		32%	31%	25%
Reading	4%	55%		36%		5%
Writing	17%	57%			26%	

3.2.3 Motivation

In response to the question about their motivation for studying Russian, students provided a gamut of responses, such as "self-interest," "for personal reasons," "language requirement," or provided some explanations that fall into two main categories: affect and career opportunities. Below are examples of affective reasons for studying Russian:

1. For someone who was born in Russia, I feel that it is embarrassing to forget the language.
2. I want to get close to mastering my first language!
3. Keep Russian language alive in my family (pass down to future generations).
4. My family has such rich Russian history that I would love to better understand and be part of that culture.
5. My speaking and listening is on par, but my goal is to learn to read, write, and understand the history and politics.

Other students mentioned professional and career goals, as the following examples show:

6. I plan to become an Ophthalmologist, so knowing fluent Russian would provide me with an opportunity to interact with my Russian-speaking patients without the need of an interpreter.
7. The graduate programs that interest me typically require professional level ability in French, Russian, or German.

One of the students explained it as follows:

8. For Language Requirement, Future Career Opportunities, and Speak Better Russian

Carreira and Kagan (2011) found similar motivations among respondents from all languages.

3.3 Second-Generation Students

Fifty point six (n = 41) of those who filled out the questionnaire were second-generation students, *born in the United States* to at least one Russian-speaking parent (see Rumbaut et al.'s (2006) definition of immigrant generations in the section on Participants above). Examining the backgrounds and proficiencies of these students allows us to come to some conclusions about curricular needs of this more homogeneous group that is beginning to dominate Russian HL classes in the U.S. (personal communication with teachers of Russian, February 28, 2017).

Table 2 Self-Assessment: Second-Generation Students ($N = 41$)

Skills	Mean
Listening	4.2
Speaking	3.5
Reading	2.4
Writing	1.8

3.3.1 Self-Reported Language Use and Competence

Twenty-two of our respondents speak Russian every day, 12 speak Russian sometimes, and seven reported speaking Russian only rarely. They mostly speak Russian with parents and grandparents. They studied Russian for 1.3 years on the average. On the scale from 0 to 5, they rated themselves the highest in listening (mean 4.2), followed by speaking (mean 3.5), reading (mean 2.4) and writing (mean 1.8.) See Table 2.

4 The Test Format

The test reported in this chapter consists of five subsections: reading, listening, writing, speaking, and grammar. The test format is similar to the standardized Russian Federation test of Russian as a Foreign Language (TRKI).[2] Students need to be literate to take all five subtests; those who cannot read and write can fill out the background questionnaire in English and take the listening subtest (the questions are both in Russian and in English), as well as the speaking subtest.

Reading Subtest In the reading subtest, an authentic descriptive text of about 500 words is followed by true/false statements that test Intermediate range comprehension. "Contextual clues" help facilitate reading (ACTFL, 2012, p. 22), such as dates, personal names, and place names. The prompt is an authentic reading that corresponds to the description of an Advanced level text in the ACTFL Guidelines: "connected discourse on a variety of general interest topics, such as news stories, explanations, instructions, anecdotes, or travelogue descriptions" (ACTFL, 2012, p. 17). The prompt is followed by multiple-choice responses at the Intermediate level given in English.

Listening Subtest The prompt is an authentic audio that corresponds to the description of an Advanced level text in the ACTFL Guidelines: "connected discourse on a variety of general interest topics, such as news stories, explanations,

[2] TRKI [тест по русскому языку как иностранному] is the Russian Federation language proficiency testing system for five areas of linguistic competence (aural comprehension, reading, writing, speaking, and grammar/lexicon) developed and administered by the Russian Ministry of Education and Science, and is the Russian component of the Common European Framework of Reference for Languages (CEFR) developed by the Council of Europe.

instructions, anecdotes, or travelogue descriptions" (ACTFL, 2012, p. 17). The prompt is followed by multiple-choice responses at the Intermediate level given both in English and in Russian. We used the same principle of adjusting the task as we did in the reading subtest. The responses in English are provided to ensure understanding and for those students who cannot read in Russian.

Writing Subtest The writing prompt instructs students to write a letter of approximately 15 sentences to their parents about their new friend. This is an Intermediate Mid task. Writing is graded holistically based on the ACTFL Proficiency Guidelines (ACTFL, 2012).

Speaking Subtest The speaking subtest offers two prompts, (1) giving advice of what clothes to pack for a trip to California (Intermediate Low/Mid task), and (2) recommending a book to read (Intermediate High/Advanced level task). Speaking is rated holistically based on the ACTFL Guidelines; however, since it is not a complete OPI interview, samples are assessed on a range (Intermediate or Advanced) that is sufficient for placement.

Grammar Subtest The grammar subtest has 81 multiple-choice questions that assess students' knowledge of basic grammar with special attention paid to the basic Russian grammar: nominal declensions, verbal conjugations, verbs of motion, and choice of verbal aspect.

Novice High to Intermediate Low performance on the reading, listening, speaking and writing subsections and the score of 75% on the grammar subtest allows students to place out of the language requirement as it roughly corresponds to the curriculum of first-year Russian (UCLA has a 1 year language requirement). A higher passing score on reading, listening, speaking and writing in combination with grammar results over 75% determines placement into a class.

5 Test Results

During the period of 2013–2016, we received 78 responses on the reading subtest; 83 students completed the listening subtest; and 83 recorded responses on the speaking subtest. All students ($N = 94$) completed the multiple choice grammar subtest. We also had 75 essays, some typed and submitted online and some written by hand and submitted separately. Many students are unfamiliar with typing in Russian, so we allowed students to submit handwritten essays. We are in the process of analyzing the results of speaking and writing, and in this chapter, we will only present these results for the four case studies (see *Case studies* below).

Table 3 Results on reading, listening, and grammar subtests

	Results			Mean	
Subtests	95–100%	75–95%	75% and below	All students	2d-gen students
Reading ($n = 78$)	$n = 25$ (32.1%)	$n = 20$ (25.6%)	$n = 33$ (42.3%)	71%	66%
Listening ($n = 83$)	$n = 20$ (24.1%)	$n = 22$ (26.5%)	$n = 41$ (49.4%)	69%	67%
Grammar ($n = 94$)	$n = 40$ (42.8%)	$n = 28$ (30.2%)	$n = 26$ (27%)	80%	69%

On the reading subtest ($N = 78$), 25 students scored 95–100%; 20 scored 75–95%; 33 scored 75% and below. The overall mean was 71%. The mean for second-generation students' was 66%. See Table 3.

On the listening subtest ($N = 83$), 20 students scored 95–100%; 22 scored 75–95%; 41 scored 75% and below. Overall mean is 69%. The second-generation group's mean score was 67%. See Table 3.

On the grammar subtest 40 students scored 95–100%, 28 students scored 75–95%, and 26 students scored 75% and below. The mean score on the grammar subtest, for all test takers, was 80%. For the second-generation cohort the grammar test mean was 69% (see Table 3). For specific difficulties on the grammar test, see Appendix 1. Table 3 shows the results and means for all test takers and second-generation test takers.

6 Case Studies

We have analyzed in detail proficiencies of four students (we will call them Anna, Julia, Daniel, and Maxine) from the second-generation cohort who, as experience shows, are representative of the range of students in our HL classes in the past 3 years. To select these four students, we first analyzed the questionnaire responses of all second-generation students ($N = 41$). We then analyzed all of these four students' test results, including writing and speaking subtests and created four representative profiles. On the questionnaires, students answered questions about language use (how often and with whom they speak Russian) and also self-assessed their ability in Listening (L), Reading (R), Speaking (S) and Writing (W) on the scale from 0 to 5, with 5 as the highest. The results are represented below as L4, R5 etc. Two certified ACTFL OPI testers rated each speaking sample. Since students did not undergo a complete OPI, we only rated results as "Intermediate range" or "Advanced range." See Appendix 2 for transcripts and English translation.

Anna Anna (test taken in 2015) speaks Russian with her parents every day, and she was taught to read and write at home for 2 or 3 years; she self-assessed her proficiencies as L5, S4, R3, and W3.[3] She received 81% on the reading test and 60% on the listening test. She successfully completed both speaking prompts at the Intermediate High as rated by two ACTFL-certified testers, even though she made some unusual stylistic choices reflecting the vocabulary typical of informal home language (for example, девчонка [devchonka (colloquial for girl)] and мужики [muzhiki (highly colloquial for guys)] that created an unintended humorous effect. She spoke "with ease and confidence" (ACTFL, 2012, p.6) and in some instances produced paragraph-length discourse. The meaning was mostly clear, and there were few grammar mistakes. Her essay, on the other hand, had a large number of grammatical mistakes. She wrote her essay on paper in block letters, not in cursive.[4] While she mostly chose correct forms of noun-verb agreement on the grammar subtest, she made a significant number of agreement mistakes in the essay. Her spelling reflects the spoken norm and thus contains multiple errors both of orthographic and morphological nature. Most notably, the text is very simple syntactically. "The writing style closely resembles oral discourse," and "writing is best defined as a collection of discrete sentences…loosely strung together" (ACTFL, 2012, p. 13). Her score was Intermediate Mid. Incorrect collocations (*играть в пиано* [igrat' v piano] to play the piano) were also documented. The grammar subtest revealed that she had a solid grasp on grammar: she scored 91% on the grammar test with the following problem areas: (a) reflexive verbs; (b) using *который* [which], (c) numeral-noun combination, and (d) use of participles.

Julia Julia (test taken in 2015) reported speaking Russian *sometimes* with her parents. She learned to read and write at home and attended a Russian pre-school for 1 year. Her self-assessment was L4, S3, R2, and W3. Julia received 86% on the reading test, and 80% on the listening test. Based on the two speaking prompts, her oral proficiency was rated Intermediate Mid. She used some incorrect vocabulary: *принести* instead of *привезти* [bring on foot instead of transport in a car]; *валенки* [valenki (Russian village felt boots]). She also made grammar mistakes, for example, using incorrect preposition/noun agreement: *для школе* [dlya shkole (for school)]. Her essay was similar to Anna's in that she produced "a collection of discrete sentences…loosely strung together" (ACTFL, 2012, p.13), but she did not make any grammar or orthographic mistakes in her essay. Her score was Intermediate Mid. Julia scored 93.83% on the grammar subtest, and some of the problem areas are the same as Anna's: (a) the use of '*который*' [which], (b) numeral-noun combination, and (c) participles.

[3] L – listening; S – speaking; R – reading; W – writing.

[4] In the Russian educational tradition, writing in block letters is tantamount to illiteracy. Preschoolers may use block letters, but children are taught penmanship as soon as they start school.

Daniel Daniel (test taken in 2014) reported studying Russian for less than a year, he did not indicate where. He can read Russian, but cannot write. He self-assessed as follows: L5, S4, R2, and W0. Daniel reported speaking Russian *sometimes* to his parents. Daniel is in the lowest percentile of all test takers: he scored 33% on the reading subtest and 20% on listening. His speaking sample was rated as Intermediate Mid. He did not quite respond to the prompts, instead of giving advice just talked about a friend who frequently visits and a book he started reading. His speech sample contained multiple grammar mistakes (*одеваться в свитеры* [to wear sweaters], *приезжать в Лос Анджелесе* [to come to Los Angeles], *прочитал такой книгу* [read such a book]), as well as lexical mistakes, wrong word usage, (*брать выбор,* instead of *делать выбор* [make a choice]; *не мог перестать читать*, instead of *не мог остановиться* [to stop reading]; *приезжать с рубашкой* instead of *брать с собой рубашку* [pack a shirt]). Daniel mostly speaks in simple sentences talking about his friend. When Daniel speaks about a book, he uses complex sentences using conjunctions *который, потому что, как будто* [which, because, as], but he omits some conjunctions where they cannot be omitted in Russian (*Я прочитал такой книгу [которая] называется* «The choice you make» [I read such a book called...]). Toward the end, his speech became partially incomprehensible, thus creating a communication breakdown. He scored 78% on the grammar subtest. The grammar subtest shows that he does not have a firm grasp on verbal agreement, and used the nominative case instead of oblique cases (see Polinsky 2006, 2008). His speech samples did contain some examples of using oblique cases. While he made more grammar mistakes in his grammar subtest than Anna and Julia, he also made the same mistakes as they did (*который* [which], the use of noun cases with numerals and use of participles).

Maxine Maxine (test taken in 2015) reported having 12 years of Russian instruction. She was taught literacy at home, attended Russian pre-school, and had a private tutor; she speaks Russian every day with her parents. She self-assessed as follows: L5, S5, R5, and W4. She scored 100% on both reading and listening, and 98.77% on the grammar test. Judging from her responses to the two prompts, she was an Advanced High level speaker. She used rich and varied vocabulary both at the everyday level (what clothes to bring on a trip) and on a more advanced level (a book she has read). On the written prompt, she produced an essay at the Intermediate High level. She wrote in cursive and only made a few spelling mistakes. However, while the essay showed a firm command of varied and precise vocabulary, it was loosely organized and had no complex syntax.

Table 4 summarizes the results of the four students' self-reported language study and use, self-assessment of their proficiency in the four skills, and test results. It shows that students overestimate or underestimate their abilities.

Table 4 Second generation: the profiles of four students

Tests	Anna	Julia	Daniel	Maxine
Language study	2–3 years at home	Home and 1 year of pre-school	Less than a year	12 years of Russian instruction
Language use	Every day with parents	Sometimes with parents	Sometimes with parents	Every day with parents
Speaking self-assessment	4	3	4	5
Speaking test	IH	IM	IM	AH
Reading self-assessment	3	2	2	5
Reading test	81%	86%	33%	100%
Listening self-assessment	5	4	5	5
Listening test	60%	80%	20%	100%
Writing self-assessment	3	3	0	4
Writing test	IM	IM	Cannot write	IH

7 Discussion

This study shows the importance of an integrated test as student outcomes differ skill by skill. A strong performance on a speaking test does not guarantee similar results on reading, listening or writing. Listening results were particularly unexpected. While students' self-rating indicated that they considered listening comprehension their strongest skill, similar to many HL learners of all languages who describe their listening comprehension as "native-like" (Carreira & Kagan, 2011), the test results did not confirm this self-assessment but instead indicated that listening and reading both posed difficulties. Grammar subtest results confirmed that "some areas of [HL speakers'] grammatical knowledge" are "more vulnerable" than others in HL language (Montrul, 2010, p. 295). The main grammatical errors were as follows:

- the nominal system: use of cases and declension of nouns, adjectives, and pronouns; noun-adjective agreement;
- verbs: verbs of motion without (unidirectional/multidirectional) and with prefixes; reflexive verbs; use of infinitive; aspect;
- collocations: time expressions;
- complex sentences: use of conjunctions *ли/если* [if], *чтобы/что* [in order/that], *который* [which], *кто* [who]; if-clauses;
- participles (a form necessary for comprehension but not production)

While grammar subtest results give an insight into the gaps in the students' knowledge of morphosyntax and indicate the areas where intervention is necessary, they do not preclude students from carrying out the task at the Intermediate level in

speaking and writing. Paucity of vocabulary, mixing registers, and lack of understanding of pragmatics is a greater obstacle to successful performance. These features hold students back and do not allow them to function at the Advanced or higher levels. Martin, Swender and Rivera-Martinez (2013) and Swender, Martin, Rivera-Martinez, and Kagan (2014), studies of Russian and Spanish HL speakers, confirm these results. In the area of lexical breadth, students demonstrated their ability to use everyday vocabulary, but lacked "precise vocabulary" (Zyzik, 2016, p. 30).

The study cases show that the amount of schooling has a clear effect on the students' performance. The student who had considerably more instruction in Russian (Maxine) reached Advanced ranges in speaking and came close to Advanced in writing.

We did not intend to test for this outcome but it became obvious that one feature of HL performance that distinguishes these students from L2 learners is that even at the Intermediate level they can be "understood by native speakers of the language, including those unaccustomed to non-native speech" (ACTFL, 2012, p. 5). The same is true of writing: while not rising above Intermediate range, native speakers could easily understand the students' written samples.

8 Curricular Implications

As previously mentioned, we based our discussion of HL learners' curricular needs on the results of the placement test. The four students whose background information and test results were analyzed in detail are representative of our current cohort of students in the HL classes. UCLA offers a special track to HL learners whose speaking proficiency falls in the Intermediate range. The curriculum targets proficiency in all four skills and uses ACTFL (2012) proficiency guidelines to determine what functions students should be able to perform.

HL curriculum is macro-based. For example, it teaches "grammar and vocabulary as dictated by function or context" and "instruction proceeds from the general message or the big ideas to analysis of linguistic building blocks" (Carreira, 2016, p. 125). This approach allows instructors to avoid re-teaching what learners can do already, but rather uses their incoming proficiencies as a springboard. The integrative nature of the placement test provides instructors with a broad picture of student proficiencies in each skill that needs to be incorporated in the curriculum. Our results show that listening comprehension is not as strong as students themselves and some researchers believe (Kagan & Dillon, 2009). As such, educators should not neglect including work on listening comprehension; it needs to be one of the foci of the curriculum.

While research on advancing HL learners' proficiencies has not been extensive (Montrul, 2013), if such learners are offered consistent and appropriate instruction, there are studies that show that HL learners can progress to Advanced High or

Superior level more rapidly than L2 learners (Davidson & Lekic, 2013). Such progress needs to be the target of instruction in HL programs, so that motivated students have the support to reach their full potential.

9 Conclusions and Further Research

The purpose of the study was to investigate the kinds of tests to administer to HL learners, with the objective of more fully assessing their communicative competencies. We hope we have shown convincingly that need analyses for HL learners require testing all four skills, as well as grammatical competence and breadth of vocabulary. Integrated tests of the kind we described help placement and provide information that can lead to adjusting curriculum for both group and individual needs.

While Russian data is the focus of this chapter, other languages may benefit from our results and curricular suggestions. A Russian HL learner is similar to a typical HL learner of other immigrant languages who,

> (1) acquired English in early childhood, after acquiring the HL"; (2) has limited exposure to the HL outside the home; (3) has relatively strong aural and oral skills but limited literacy skills; (4) has positive HL attitudes and experiences; and (5) studies the HL mainly to connect with communities of speakers in the United States and to gain insights into his or her roots. (Carreira & Kagan, 2011, p. 81)

A collection of papers *Language Diversity in the USA* (Potowski 2010) explores immigration patterns and language needs of the speakers of 10 United States minority languages with the largest numbers of speakers. The volume makes it clear that while each language community is different, defined by the reasons and character of immigration, each share many features related to language loss and maintenance. One could reach the same conclusion upon examining the special issue of the *Heritage Language Journal* edited by Lo Bianco and Peyton (2013). The volume contributors explored the vitality of heritage languages in the United States, and they persuasively argued that in adjusting for linguistic differences, one could develop curricula to address the multiple language similarities, meeting the needs of various language groups.

Research into HL listening and reading practices and strategies is sorely missing from the studies of HL competencies. Such studies will allow us to offer HL learners the instruction that would not only help them preserve their home language but also advance it to high levels of proficiency. It would be particularly fruitful to conduct parallel research on reading strategies in both alphabetic (Latin and non-Latin based) and non-alphabetic languages. Once such data are available, language educators will be able to base curriculum design for HL learners on solid data and not on assumptions. Moreover, we could significantly advance our knowledge of HL instructional needs by conducting parallel studies in several languages and developing a curricular outline that would not be language specific but rather an instructional blueprint.

Appendices

Appendix 1: Difficulties Encountered by the Test Takers on the Grammar Test

- Participles—35.11%
- Case system—21.17%
- Time expressions—19.86
- Verbs of motion (uni/multi-directional, prefix) 21.88%—15.6%/18.75%
- Perfective/Imperfective forms (aspect)—15.96%
- Complex sentences (ли/если, чтобы/что, который)—14.04%
- Reflexive verbs—13.47%
- Use of infinitive—8.51%

Appendix 2: Case Studies: Transcripts of Recordings

The transcripts of the students' recordings are followed by English translation. Mistakes are noted in parentheses. FR: Full Russian.

Prompt 1. Your Russian friend is coming to the US to visit next winter. Describe in detail the weather in your city at this time of the year. Give your friend advice what clothes to pack.

Anna

Duration of speech: 0:46

В Южной Калифорнии даже в зимой хорошая погода, так что, когда вы будете приезжать, вы можете с собой только брать джинсы и легкая куртка, и а то обычно, когда я это нашу мне уже тепло. В градусах Фаренгейтах обычно тут 60–70 градусов, иногда даже больше. И обычно дождь так часто не бывает, ну ты все равно можешь с собой брать куртку для дождя, но, как я сказала, обычно у нас сильный дождь только один или два раза бывает и так у нас ну солнце и даже иногда жарко.

The weather in Southern California is good even in winter (uses preposition 'in' while there is no preposition in Russian), so when you will be coming (wrong verbal aspect) you can only take jeans and a light jacket, and even then usually when I wear all of this, it's already warm (meaning is not clear). In Fahrenheit degrees usually we have 60–70 degrees, sometimes even more. And usually rain does not happen so often, well, anyway you can take a jacket for rain, but, as I said, usually we have heavy rain only once or twice, and well it's sunny and even hot sometimes.

Julia
Duration of speech: 0:55

Я очень рад, что ты можешь приехать ко мне в Америку. У нас здесь в Лос Анджелесе очень теплая погода, иногда дождь, но иногда просто солнце. Так что принеси легкая куртка и, может быть, шарф. Валенки здесь не надо, здесь теплая погода, и дождь очень редко. Снег здесь вообще не бывает, но когда солнце, принеси солнечные очки. Надо зонт тоже, потому что иногда дождь. Если мы пойдем куда- то на лыжи, принеси куртку и валенки.

I am very glad (uses masculine) you can come to see me in America. Here is Los Angeles the weather is very warm, sometimes it rains, but sometimes it's just the sun (SR: солнечно – sunny) So bring (this word is used when walking only) a light jacket and may be a scarf. You don't need felt boots (valenki – a particular kind of peasant felt boots, it is unlikely that the student has ever seen them, other than in movies), it's warm and it rains very rarely. It never snows, but when it's sunny, bring sun glasses. One needs an umbrella too because there is rain sometimes (FL: идет дождь – it rains) If we go somewhere to ski (omitted the verb and used the Accusative Case instead of the Prepositional).

Daniel
Duration of speech: 0:34

Мой хороший друг приезжает ко мне из Питера. Он живет в очень холодный город. И он сейчас зимой одевается в свитеры, пиджаке, теплые сапоги. Он ко мне приезжает здесь, в Лос Анджелесе. У нас очень тепло. У нас постоянно солнце, редко дождь, снегу совсем нет. Значит пусть он приезжает с рубашкой, с шорт, даже плавки, можно в океан плавать или в бассейн. Значит, ему не надо тепло одеваться, пусть приезжает, и очень приятно.

My good friend is coming to see me from Peter (colloquial for St. Petersburg) He lives in a very cold city (uses the Nominative Case instead of the Prepositional). And he now in winter wears sweaters, a dress jacket (the ending of the Prepositional Case instead of the Accusative), warm boots. He comes to me here in Los Angeles (the Prepositional Case instead of the Accusative). It is very warm here. We always have the sun, it rains rarely, there is no snow. So let him just come with a shirt, shorts (uses the Genitive Case instead of the Instrumental), even swimming trunks, one can swim in the ocean (the Accusative Case instead of the Prepositional) or in a swimming pool (the Accusative Case instead of the Prepositional). So he doesn't need to be warmly dressed, let him come and visit, and it is very pleasant (not clear what 'pleasant' refers to).

Maxine
Duration of speech: 1:23

Зимой бывает дождливо и холодно, но по сравнению с Москвой намного теплее. Снега нет, и температура довольно таки стабильная. Бывает сильный ветер, так лучше одеваться потеплее. В школу я ношу теплый свитер или куртку. Я рекомендую взять колготки, брюки, куртку, пару свитеров и пару футболок, потому что температура часто меняется тоже. Не бойся брать несколько летних нарядов, типа одежды, например, платьев или шортов,

потому что у нас часто бывают дни, когда солнце выходит и тепло. Когда моя московская бабушка приезжает, она вообще редко одевает теплую одежду. Обязательно возьми зонт, он тебе понадобится от дождя. Я люблю ходить на каток и хотела бы сходить с тобой. Возьми свои коньки тоже тогда. Брать варежки не надо, оставляй их домой, потому что снега, как я уже сказала, нет. Шарф часто помогает, потому что он помогает от ветра, который утром обычно довольно таки сильный. И не забывай сапоги, у нас бывает мокро после дождя. И не хотелось бы, чтобы твои туфли намокли.

It can be rainy and cold in winter, but much warmer in comparison to Moscow. There is no snow and the temperature is rather stable. Sometimes it can be quite windy, so it's better to be warmly dressed. When I go to school, I put on a warm sweater or a jacket. You should take leggings, pants, a jacket, a couple of sweaters and a couple of t-shirts because the temperature can frequently change too. When my Moscow grandmother (a correct Russian turn of phrase) visits, she rarely wears anything warm. Make sure to take an umbrella, you'll need it against rain. I like to skate and I would like to go skating with you. Bring your skates as well. No need to take mittens, leave them at home because, as I said already, there is no snow. A scarf helps because it helps against the wind, which can be in the morning, quite strong. And don't forget to bring boots, it can be wet after the rain. We wouldn't want for you to get your shoes wet.

Prompt 2. You have read a book that you found interesting and you want to recommend it to a friend.

Anna

Duration of speech: 1:23

Летом я читала сирию, и первая книжка моя, мне кажется самая лучшая. Главная характер она была девчонка, и мне это было интересно, потому что обычно, когда я читаю книги, главный характер - это всегда мужики. И это было интересно читать с, ну с женщиной. Было еще интересно, потому что все остальные характеры имели очень интересные таланты. Каждый должен был иметь эти таланты, чтоб ну победить других в такой игре. И в этих играх они должны были, только один может победить. И там, наверное, 2–4 люди должны в арене сражаться, и только один может жить. Ну, в конце книги, я не хочу этот, ну сказать, что случилось, ну, короче, если вы смотрели кино про эту книжку, значит вы может уже знаете. Ну еще, если вам нравиться книга, которая имеет кино, вы точно должны эту прочитать, потому что мне кажется, книжка лучше, чем кино.

Last summer I read a series, and the first book was, it seems to me the best. The main character (this is a calque from English; FR: герой; also lack of agreement) she was a girl (informal for 'girl') and it was interesting for me because usually when I read b ooks the main character it's always guys (using a highly colloquial word, the same she used in her oral sample). And it was interesting to read well with a woman. It was also interesting because all the other characters had very interesting talents. Each of them had to have these talents to defeat the others in such a game. And in these games, they had to, only one could win. And it looked like 2–4 people

had to fight in an arena, only one could win. Well at the end of the book I don't want to say well what happened, well in a word, if you watched the film about this book that means you already know may be. Well and if you like a book which has a film you really need to read it because I think the book is better than the film.

Julia
Duration of speech: 1:04

Я только что прочитал очень интересную книжку. И я думаю, что тебе очень понравится. Она романтика, но у нее тоже есть исторические факты. Там есть монстры и вампиры, и она очень страшная иногда, но это очень интересно. Я должна была читать эту книжку для школе. Так что я не очень хотела в начале, но когда я ее уже прочитала, я была очень рада, потому что это была одной из самой лучшей книжки, чтобы я когда- то читала. И я думаю, что тебе очень понравится. Она будет фильм скоро, и я хочу ее смотреть тоже.

I just read a very interesting book. And I think that you will like it. It is a romance, but it also has historic facts. There are monsters and vampires and it is scary sometimes, but this is very interesting. I had to read the book for school (using the word that means K-12, not university//incorrect agreement with the preposition. FR: для школы) So I didn't want to very much at first, but when I finished reading, I was very glad because it was one of the best books (wrong case) that I ever read. And I think that you will really like it.

It will be a movie soon and I want to see it too.

Daniel
Duration of speech: 0:35

Я прочитал такой книгу, называется «The choice you make». Эта книга про жизнь и про выборы, которые ты доверяешь жизни, и почему ты берёшь этот выбор. Эта книга, у неё есть реально характер человека, который написал, очень умный. И он так пишет, как будто ты реально в это веришь, и ты реально хочешь это знать, это понимать. Эта книга, ты открываешь с первой страницы и просто не можешь перестать читать. Я часами просто сидел и читал, и читал, даже не хотел себе чай налить, потому что ты просто такой сил, такой характер, власть.

I read such a book (incorrect agreement, he skips который) called … This book is about life and choices that you trust the life and why you take this choice (not very clear). This is the book, it has a real character (wrong word) of the man who wrote it, very intelligent. And he writes so as if you believe in it and you really want to know it, understand it. This book, you open the first page and can't stop reading. I spent hours just sitting and reading and reading, even did not want to pour myself a cup of tea, because you just such strength, such character, power (not clear).

Maxine
Duration of speech: 1:42

Два года назад я прочитала очень интересную книгу. Она называется Мастер и Маргарита. Ее написал автор Михаил Булгаков. Мои родители очень любят эту книгу, потому что квартира, которая описана в этой книге,

находится очень близко к квартире, в которой они жили, когда они жили в Москве. Так они чувствовали, как будто они жили очень близко к героям, которые описаны в этой книге. В этой книге много фантазии, и еще очень много любви и страсти. Часть книги описывает убийство Иисуса Христоса, и как его предали, и как римляне над ним издевались. А другая часть книги описывает любовь между колдуном, которого называют Мастер, и обычной женщины, которую зовут Маргарит. Они влюбляются, но понимают, что им будет очень тяжело быть вместе. В этом заключается страсть, которая описана. Маргарита берет на себя много терпения [this is a non-Russian turn of phrase, but not an English calque either] и начинает участвовать к колдунстве [the correct Russian word is колдовство], в котором участвует Мастер и его друзья. Прилетают за ней ведьмы, и она намазывает на свое тело какую то мазь, с которой она превращается в чуть ли не в такое святое существо. Мне эта книга очень понравилась, потому что то в ней было и описано настоящее жизнь, и фантастическая жизнь, которая существует только из-за любви между двух человек. Я очень рекомендую тебе прочитать ее тоже.

Two years ago I read a very interesting book. Its title is Master and Margarita. It was written by Mikhail Bulgakov. My parents love this book because the apartment described in this book is very close to the apartment where they lived when they lived in Moscow. So they felt as if they lived very close to the characters when they lived in Moscow. In this book there is a lot of imagination/fantasy and also a lot of love and passion. One part of the book describes the murder of Jesus Christ and how he was betrayed and how the Romans tortured him. The other part describes the love between a magician called master and a simple woman whose name is Margarit. They fall in love, but they understand that it will be very difficult for them to be together. This is the part about passion which is described. Margarita takes upon herself [direct translation of a non-Russian turn of phrase; FR: проявляет много терпения] and she starts participating in the magic practiced by Master and his friends. Witches come for her and she dabs some ointment on her body which turns her into some kind of a holy [wrong word; FR: магическое] being. I liked this book a lot because it describes real life and fantastic life which exists only because of love between two people. I highly recommend that you read the book.

References

ACTFL. (2012). *Proficiency guidelines: Speaking, listening, reading, writing*. Retrieved from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Andrews, D. (1998). *Sociocultural perspectives on language change in diaspora. Soviet immigrants in the United States*. Philadelphia, PA: John Benjamins.

Carreira, M. (2016). Supporting heritage language learners through macrobased teaching: Foundational principles and implementation strategies for heritage language and mixed classes. In M. Fairclough & S. M. Beaudrie (Eds.), *Innovative strategies for heritage language teaching: A practical guide for the classroom* (pp. 123–142). Washington, DC: Georgetown University Press.

Carreira, M., & Kagan, O. (2011). The results of the National Heritage Language Survey: Implications for teaching, curriculum design, and professional development. *Foreign Language Annals, 44*(1), 40–64. https://doi.org/10.1111/j.1944-9720.2010.01118.x

Davidson, D., & Lekic, M. (2013). The heritage and non-heritage learner in the overseas immersion context: Comparing learning outcomes and target-language utilization in the Russian flagship. *Heritage Language Journal, 10*(2), 226–252.

Fairclough, M. (2012). A working model for assessing Spanish heritage language learners' language proficiency through a placement exam. *Heritage Language Journal, 9*(1), 121–138.

He, A. W. (2016). Heritage language learning and socialization. In P. A. Duff & S. May (Eds.), *Language socialization, encyclopedia of language and education* (pp. 1–12). New York, NY: Springer.

Isurin, L. (2011). *Russian diaspora: Culture, identity, and language change* (Vol. 99). New York, NY: Walter de Gruyter.

Kagan, O., & Dillon, K. (2006). Russian heritage learners: So what happens now? *Slavic and East European Journal (50th Anniversary Issue), 50*(1), 83–96. https://doi.org/10.2307/20459235

Kagan, O., & Dillon, K. (2009). The professional development of teachers of heritage language learners: A matrix. In M. Anderson & A. Lazaraton (Eds.), *Bridging context, making connections: Selected papers from the fifth international conference on language teacher education* (pp. 155–175). Minneapolis, MN: Center for Advanced Research on Language Acquisition.

Kagan, O., & Dillon, K. (2010). Russian in the USA. In K. Potowski (Ed.), *Language diversity in the USA* (pp. 179–194). Cambridge, UK: Cambridge University Press.

Kagan, O., & Friedman, D. (2003). Using the OPI to place heritage learners of Russian. *Foreign Language Annals, 36*(4), 536–545. https://doi.org/10.1111/j.1944-9720.2003.tb02143.x

Lo Bianco, J., & Peyton, J. K. (Eds.) (2013). Special issue on language vitality in the U.S. *Heritage Language Journal, 10*(3).

Martin, C., With Swender, E., & Rivera-Martinez, M. (2013). Assessing the oral proficiency of heritage speakers according to the ACTFL proficiency guidelines 2012—speaking1. *Heritage Language Journal, 10*(2), 73–87.

Montrul, S. (2010). Dominant language transfer in adult second language learners and heritage speakers. *Second Language Research, 26*(3), 293–327. https://doi.org/10.1177/0267658310365768

Montrul, S. (2013). How "native" are heritage speakers? *Heritage Language Journal, 10*(2), 15–39.

Polinsky, M. (2006). Incomplete acquisition: American Russian. *Journal of Slavic Linguistics, 14*, 161–219. Retrieved from http://www.jstor.org/stable/24599616

Polinsky, M. (2008). Heritage language narratives. In D. Brinton, O. Kagan, & S. Bauckus (Eds.), *Heritage language education: A new field emerging*. New York, NY: Routledge.

Polinsky, M., & Kagan, O. (2007). Heritage languages: In the 'wild' and in the classroom. *Compass of Language and Linguistics, 1*(5), 368–395. Retrieved from http://scholar.iq.harvard.edu/files/scholar/uploads/11/Offprint.pdf

Rumbaut, R. G., Massey, D. S., & Bean, F. D. (2006). Linguistic life expectancies: Immigrant language retention in southern California. *Population and Development Review, 32*(3), 447–460. https://doi.org/10.1111/j.1728-4457.2006.00132.x

Schwartz Caballero, A. M. (2014). Preparing teachers to work with heritage language learners. In T. G. Wiley, J. K. Peyton, D. Christian, S. C. K. Moore, & N. Liu (Eds.), *Handbook of heritage, community, and Native American languages in the United States: Research, policy, and educational practice* (pp. 359–369). New York, NY: Routledge.

Sohn, S.-O., & Shin, S.-K. (2007). True beginners, false beginners, and fake beginners: Placement challenges for Korean heritage speakers. *Foreign Language Annals, 40*(3), 407–418. https://doi.org/10.1111/j.1944-9720.2007.tb02866.x

Swender, E., Martin, C. L., Rivera-Martinez, M., & Kagan, O. E. (2014). Exploring oral proficiency profiles of heritage speakers of Russian and Spanish. *Foreign Language Annals, 47*(3), 423–446. https://doi.org/10.1111/flan.12098

Valdés, G. (1989). Teaching Spanish to Hispanic Bilinguals: A look at oral proficiency testing and the proficiency movement. *Hispania, 72*, 392–401. https://doi.org/10.2307/343163

Valdés, G. (2000). *Introduction. Spanish for native speakers, Volume 1. AATSP professional development series handbook for teachers K-16*. New York, NY: Harcourt College Publishers.

Zemskaya (2001). Земская, Е. А., & Гловинская, М. Я. (2001). *Язык русского зарубежья: Общие процессы и речевые портреты* (Vol. 53). Вена. [Jazyk Russkogo Zarubezhja [Language of Russian emigration]. *Moscow/Vienna: Wiener Slawisticher Almanach*.].

Zyzik, E. (2016). Toward a prototype model of the heritage language learner. In M. Fairclough & S. M. Beadrie (Eds.), *Innovative strategies for heritage language teaching: A practical guide for the classroom* (pp. 19–38). Washington DC: George Washington University.

Olga Kagan was a professor in the UCLA Department of Slavic, East European and Eurasian Languages and Cultures and director of the National Heritage Language Resource Center. She has published textbooks of Russian both as a foreign language and as a heritage language. Her textbook of Russian as a Heritage Language, Russian for Russians, received a book award from the American Association of Teachers of Russian and Eastern European Languages (AATSEEL). In 2015 she received an MLA Award for Distinguished Service to the Profession.

Anna Kudyma is a senior lecturer and TA Supervisor in the UCLA Department of UCLA Slavic, East European and Eurasian Languages and Cultures. She holds a MA in Russian language pedagogy and a Ph.D. in linguistics. Her primary interests are language pedagogy and computer-assisted language learning. She is the coauthor of five textbooks, including Учимся писать по-русски: экспресс-курс для двуязычных взрослых… [We are learning to write in Russian: An express-course for bilingual adults] (Zlatoust, Saint Petersburg 2011).

Modern-Day Foreign Language Majors: Their Goals, Attainment, and Fit Within a Twenty-First Century Curriculum

Paula Winke, Susan M. Gass, and Emily S. Heidrich

Abstract In 1967, John Carroll produced a seminal research report that overviewed the proficiency levels of foreign languages majors at U.S. colleges and universities with the goal to capture and record the state of foreign language instruction in the United States at the university and college level. This chapter revisits the status of foreign language proficiency amongst majors with data from language majors from three large state universities. Data collected in areas of listening, speaking, and reading are compared with the data of Carroll. Fifty years later, a similar picture emerges with speaking and listening skills falling behind other skills. What is different, however, is the general picture of what it means to be a major, with the majority of students today declaring multiple majors as opposed to the single "language/literature" major of the past. A second area of investigation concerned the possible predictors of success amongst language majors. 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 is important is when language learning begins, with greater progress being made in college-level courses when language learning begins before tertiary education.

Keywords Foreign language major · Double major · Proficiency · Speaking · Reading · Listening

P. Winke (✉) · S. M. Gass
Second Language Studies Program, Michigan State University, East Lansing, MI, USA
e-mail: winke@msu.edu; gass@msu.edu

E. S. Heidrich
Center for Language Teaching Advancement, Michigan State University,
East Lansing, MI, USA
e-mail: heidric6@msu.edu

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_6

1 Introduction

In 1967, John Carroll produced a seminal research report that overviewed the proficiency levels of foreign languages majors at U.S. colleges and universities. Carroll's goal was to capture and record the state of foreign language instruction in the United States at the university and college level. For the study, in the spring of 1965, Carroll tested a nation-wide sample of 2523 seniors majoring in five foreign languages (French, German, Italian, Russian, and Spanish). In the current chapter, 50 years later, we revisit the status of foreign language proficiency amongst majors with data from three large state universities (Michigan State University, University of Minnesota and University of Utah[1]). With federal funding to conduct language proficiency assessments over a three-year period (2014–2016), data were collected from majors and non-majors. In this chapter, we report only on the data from majors so that we can make comparisons between our data and those of Carroll. The results presented in this chapter will allow for a better understanding of the language major in an early twenty-first century context.

In the first part of the chapter, we consider the proficiency data from 3 years of testing and compare those data with Carroll's results. For this analysis, we analyzed data from French, Russian, and Spanish given that these three languages were the only languages common across Carroll's assessments and our assessments. Therefore, we view this part as a partial replication of the work done by Carroll and as an opportunity to take the "foreign language major's temperature" in the twenty-first century. In the second part of the chapter, we expand the language base to include Chinese (in addition to French, Spanish, and Russian) and consider only the data on majors in these languages from one university (Michigan State University [MSU]) due to the extensive background data collected from MSU students. We report language attainment results related to background variables on gender, heritage status, and study abroad experience. We conclude the chapter with a retrospective of Carroll's data and how the situation today differs from the situation of 50 years ago.

2 MLA Database: Bachelor Degrees, 1967–2015

We begin by looking at numbers of students in the United States earning bachelor degrees (Fig. 1) beginning with data from shortly after Carroll's study and ending with data shortly before the end of our data collection period (U.S. Department of Education, 2016a, b). As can be seen, in 1972–73 (not long after the publication of Carroll's study), Spanish degrees surpassed French, and Spanish has remained the dominant language major ever since with a large upswing beginning in the late

[1] Because the University of Utah used tests that used a different scoring system, we opted to limit the results presented here to those from Michigan State University and the University of Minnesota.

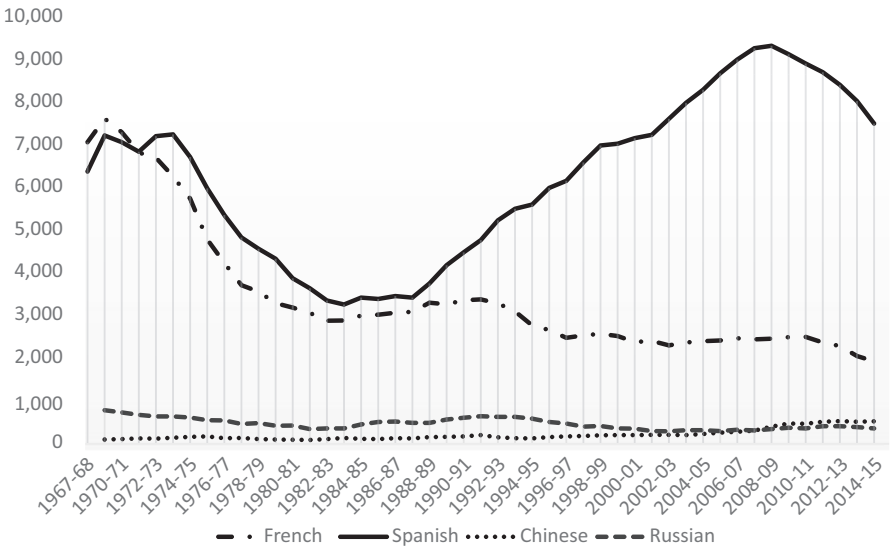


Fig. 1 Bachelor degrees granted by Postsecondary Institutions in Chinese, French, Russian, and Spanish from 1967–2015. (Data from U.S. Department of Education, 2016a, b)

1980s. Another interesting trend is the general downturn of foreign language study for a period of about 15 years (1972–1987), after which Spanish became the dominant language of the four illustrated, with very little change in Russian and Chinese.

As we will discuss below, the concept of a major is quite different today than it was 50 years ago, when most students had only one major. In today's world, it is quite common to see students major in more than one subject matter (e.g., language and mechanical engineering, or language and a business-related field), making the direct comparison with Carroll's data less than straightforward. The figure above counts all bachelor degrees granted in a particular language, regardless of a student's status as a sole language major or someone with multiple majors.

3 Carroll's 1967 Study and Beyond

Carroll used the *MLA Foreign Language Proficiency Tests* in four skills (reading, writing, listening, speaking) to test foreign language majors. The total time for the battery of tests was two hours. In addition, students filled out a broad background questionnaire and a 30 minute *Modern Language Aptitude Test* (short form). There were numerous findings, but for our immediate purposes, we note two: First, speaking and the audiolingual skill of listening were generally low in comparison to reading and writing; and second, study abroad had a significant impact on attainment. Carroll further found that a language-learning-start in elementary school and/or heritage language status increased one's chance of higher-level competency.

Table 1 FL majors in the U.S. 1964–1965, and numbers tested (a subset)

Language	Total in U.S.[a]	No. in participating institutions	No. Seniors Tested	Percent of Total in U.S.	Percent of Total in Participating Institutions
French	5043	2287	1270	25.2	55.5
Russian	556	331	105	18.9	31.7
Spanish	4178	1900	968	23.2	50.9

[a]This table includes only students in institutions defined as being in the 1962–63 population that was used as a basis for drawing the sample. It excludes 60 students listed as majoring in a Romance language, some of whom may have appeared in the sample, but who are not included here because no information is available on the language in which they were tested

Table 1 is an overview of the sample from Carroll's study based on the three of the four languages discussed in his chapter. As can be seen, French had the largest number of majors tested, followed by Spanish, with Russian having the smallest number.

Carroll compared student achievement on the *MLA Foreign Language Proficiency Tests* to the ratings from the Foreign Service Institute (FSI), the scale that had "been used as [a] common basis for comparing skills and for comparing languages" (p. 134). The FSI scale at that time ranged from 1 (elementary proficiency) to 5 (native or bilingual proficiency), with 3 being the minimum proficiency needed to work in a professional setting (see the Interagency Language Roundtable skill level descriptions and the scale history at http://www.govtilr.org/). In Carroll's comparison of the achievement of students on the MLA tests to this overall rating of proficiency, he found that the "median graduate with a foreign language major can speak and comprehend the language only at about an FSI Speaking rating of "2+", that is, somewhere between a 'limited working proficiency' and a 'minimum professional proficiency'" (p. 134). By his calculations, both Spanish and French students would be rated closer to a 3 on the FSI for reading. On the other hand, students of Russian were at a "limited working proficiency" capacity, coming in just below FSI 2 for both speaking and reading.

Over the years, researchers have followed in Carroll's footsteps and investigated aspects of foreign language learning on college campuses to shed a more refined light on foreign language proficiency development (and factors that affect it) at the college-level in the United States (Bernhardt & Brillantes, 2014; Clément & Kruidenier, 1983; Holmquist, 1993; Lafford, 2004; Magnan, 1986; Oller & Nagato, 1974; Rifkin, 2005; Robinson, Rivers, & Brecht, 2006; Rosengrant, 1987; Schumann, 1975; Spada, 1986; Spolsky, 1969; Tschirner, 1996, 2016; Wong & Van Patten, 2003).

But one issue that researchers have not re-investigated since 1967 is how being a foreign language major (or minor) affects attainment and opportunities in foreign language learning. In other words, what level of attainment can one expect from foreign language majors in the early twenty-first century. This is important because Carroll's metric of foreign language learning in 1967 was focused on the major; but, as noted earlier, being a language-only major today may be rare. In modern-day higher education, there are competing demands and majoring in more than one area to increase employment opportunities and to provide a wider breadth of knowledge

(Urlaub, 2014) is commonplace. As a result, there is a lack of a desire to complete a major with a literary-theory focus (Kym, 2011) because such work is often seen as impractical for employment beyond continuing on to graduate-level literary study. In fact, as noted above, we will discuss a slightly different foreign-language-student profile, namely one that holds a double major.

4 Database for Current Study

In 2014, Michigan State University, along with the University of Minnesota and the University of Utah, received federal grants from the National Security Education Program's Flagship Program to undertake a broad-based testing program to include proficiency assessments of foreign language students in the skills of speaking, reading, and listening. The numbers of tests administered differed across the three universities, and the languages selected also differed, but the grant programs were similar: Each university had the goal of measuring the proficiency levels of students across all four years of their undergraduate curricula and across the language programs being studied.

At Michigan State University, students were tested in Chinese, French, Russian, and Spanish, whereas at the University of Minnesota students were tested in Arabic, French, German, Korean, Portuguese, Russian, and Spanish, and at the University of Utah, the languages assessed were Arabic, Chinese, Korean, Portuguese, and Russian. Thus, the scope of the testing for this three-university grant was much broader than Carroll's study, as he tested only majors. As Carroll noted, "[t]he primary purpose of this study was to measure in meaningful terms the foreign language proficiency levels attained at time of graduation by American college students who 'major' in French, German, Italian, Russian, or Spanish" (p. 131). In this chapter, because we are only looking at majors, we limited our analysis to proficiency scores from students enrolled in third and fourth year language classes who had declared majors in the languages assessed. The data come from proficiency assessments in spring 2015, 2016, and 2017. To be included in our analysis, students had to have taken all three ACTFL language proficiency tests (reading, speaking, and listening). If they took these tests more than once, we included only their most recent set of tests. In sum, our analysis is based on 884 majors, 22 in Russian, 227 in French, and 635 in Spanish.

The tests we used were based on the standards of the American Council on the Teaching of Foreign Languages (ACTFL).[2] For speaking, we used the Oral Proficiency Interview – Computerized (OPIc); for reading, the ACTFL Reading

[2] Carroll served as a consultant in developing the FSI scale, later revised by the Interagency Language Roundtable (ILR) and refined so it could be applied consistently by various raters. In the early 1980s, the American Council on the Teaching of Foreign Languages (ACTFL) created proficiency guidelines, with the Guidelines officially appearing in 1986. While the ILR and ACTFL scales are not direct equivalents (the ILR is used for measurement of professional ability, as opposed to ACTFL, which was aimed at the academic community), a general sense of proficiency can be gleaned from both, allowing us to compare where majors are today.

Proficiency Test (RPT); and for listening, the ACTFL Listening Proficiency Test (LPT). These tests are on-demand tests that are taken on the computer online, administered by the company Language Testing International (https://www.languagetesting.com/). In the case of the RPT and LPT, the tests are automatically computer-graded, and scores are generated immediately upon completion of the test. For this grant project, the scores were given to the test taker, and also to the language programs. The OPIc requires the student to respond to questions delivered by a virtual "partner" (a computer avatar) instead of a live interviewer, as in a traditional OPI test. Students are rated in any one of four broad categories: (1) Novice (2) Intermediate (3) Advanced or (4) Superior, with levels Novice through Advanced each containing three sub-levels: Low, Mid, and High (e.g., a student could be assigned "Novice Mid (NM)" or "Intermediate High (IH);" see ACTFL (2012) for more information about the scale descriptors).

The results from the current analysis are presented in Table 2 and graphed in Fig. 2. Following Kenyon and Malabonga (2001), to calculate means, we transformed achievement levels into a series of ranked scores such that "10" represented the highest level attainable on ACTFL measures (i.e., superior/S) and "1" represented the lowest (i.e., novice low/NL). Levels in between were coded accordingly in one-point increments. See Chap. 9 by Tigchelaar in this volume for further information on this issue.

A diverse picture emerges when looking at each of these languages individually. The results of the Spanish students (also the largest *n* size) mirror Carroll's results most closely. Students of Spanish had test results that were strongest in reading skills, followed by listening, then speaking. The reading scores of French students were similarly strong, but the average speaking score for French students was slightly higher than the listening scores. The Russian majors displayed more proficiency in their speaking skills, followed closely by reading. The listening skills of the Russian students were an average of a full level below their speaking skills.

5 Language Major or Multiple Majors?

As mentioned above, many college students 50 years ago had a single major, as opposed to students now who may have multiple majors, with a foreign language major being one, a trend that has been documented in the literature since as early as

Table 2 Mean proficiency level in listening, speaking, and reading for majors in French, Russian, and Spanish

	Mean (S.D.)		
Language	Listening	Speaking	Reading
French (n = 227)	5.58 (1.37)	5.81 (1.53)	6.22 (8.64)
Russian (n = 22)	3.64 (1.43)	4.64 (1.33)	4.50 (1.54)
Spanish (n = 635)	5.45 (1.38)	5.20 (1.14)	6.49 (1.36)

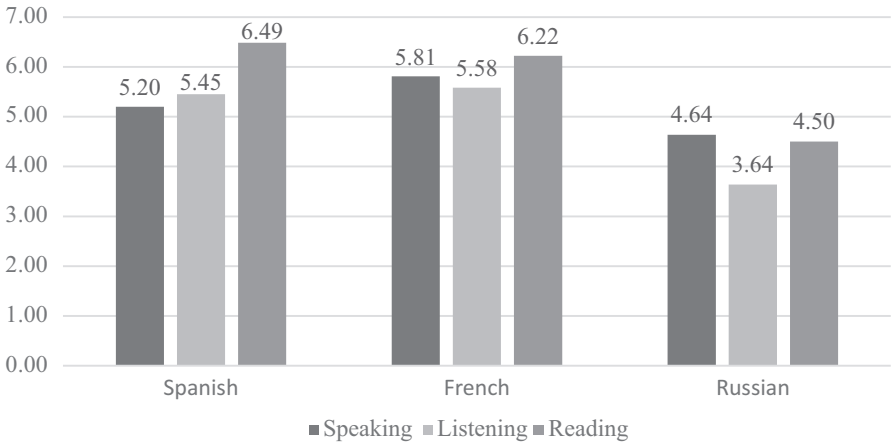


Fig. 2 Graphic representation of mean Proficiency level in speaking, listening, and reading for majors in Spanish, French, and Russian

the 1980s (see, for example, Herman, 1987), with more recent literature finding that "second majors" are more often paired with language degrees than any other higher education degree, except for "area studies" (Lusin, 2009). We wanted to see if the students who had one or more majors differed in their performance on the proficiency tests. Based on the information on majors from the registrar's office at Michigan State University, 125 of the 884 majors from Michigan State examined in this study were identified as language-only majors, which means that they majored in foreign languages only, whether one or, in some cases, more than one; the remaining students, totaling 759, were classified as hybrid language majors, as they had at least one parallel or secondary major besides a foreign language. Overall, the language-only major group seemed to outperform the hybrid language group on all three skills being assessed—listening, speaking, and reading—according to the assessment outcomes that were aligned with the ACTFL language proficiency guidelines.

Of the hybrid group, only 14.76% scored at the Advanced level or higher (i.e., Superior, Advanced High, Advanced Mid, and Advanced Low) on the OPIc, with the majority falling in the Intermediate range (80.63%), whereas of the language-only group, 32% obtained a score at or above AL and another 32% achieved IH. Similar performance differences were also seen on the LPT, where the language-only major (46.40%) was found to be nearly twice as likely to achieve an Advanced level or higher as the hybrid (26.35%) group. The hybrid group closed this performance gap to some extent on the RPT: While 52.44% and 44.14% of the hybrid group scored at the Advanced and Intermediate range, respectively, the corresponding percentages in the language-only major group were 70.40% and 27.20% respectively.

A more detailed understanding of the score distribution among language-only and hybrid language majors is provided in Tables 3, 4, and 5, where the student's

Table 3 ACTFL OPIc levels by major

Group	Language	S	AH	AM	AL	IH	IM	IL	NH	NM	NL	**Total**
Hybrid language major	French	0	6	17	28	40	47	44	4	0	0	**186**
	Russian	0	0	1	1	3	3	10	1	1	0	**20**
	Spanish	0	4	11	44	111	229	125	27	1	1	**553**
	Total	**0**	**10**	**29**	**73**	**154**	**279**	**179**	**32**	**2**	**1**	**759**
Language-only major	French	1	6	5	10	7	7	4	1	0	0	**41**
	Russian	0	0	0	0	0	1	1	0	0	0	**2**
	Spanish	0	0	3	15	33	21	10	0	0	0	**82**
	Total	**1**	**6**	**8**	**25**	**40**	**29**	**15**	**1**	**0**	**0**	**125**

Note: For clarity reasons, we did not include in the Tables 3, 4, or 5 a double language major who took tests in both Spanish and French. His Spanish and French RPT levels were AL and IH, respectively

Table 4 ACTFL LPT levels by major

Group	Language	S	AH	AM	AL	IH	IM	IL	NH	NM	NL	**Total**
Hybrid language major	French	0	0	3	56	38	44	26	19	0	0	**186**
	Russian	0	0	0	1	1	3	7	5	2	1	**20**
	Spanish	1	2	12	125	101	170	86	53	3	0	**553**
	Total	**1**	**2**	**15**	**182**	**140**	**217**	**119**	**77**	**5**	**1**	**759**
Language-only major	French	0	1	0	18	7	9	3	3	0	0	**41**
	Russian	0	0	0	0	0	0	0	0	2	0	**2**
	Spanish	1	1	4	33	14	16	10	3	0	0	**82**
	Total	**1**	**2**	**4**	**51**	**21**	**25**	**13**	**6**	**2**	**0**	**125**

score information (i.e., OPIc, LPT, and RPT) has been broken down by group and language. Of the three foreign languages, the French program had the highest proportion of language-only majors (18.06%)—one out of five French majors was identified as a language-only major—followed by Spanish (12.91%) and Russian (9.09%). Regarding the score distribution of language-only and hybrid language majors within each language program, we decided to focus discussion only on French and Spanish due to the small number of language-only majors in Russian ($N = 2$).

On the OPIc, language-only majors in both French and Spanish tended to be twice as likely to score at the level of AL or higher as hybrid-language majors in the corresponding programs, despite the fact that the French major in general outperformed the Spanish major in terms of the percentage of Advanced achievers (French [Advanced scorers] = 32.15% vs. Spanish [Advanced scorers] = 12.13%). Unlike the OPIc, the French and Spanish students performed to a large extent alike on the LPT (French [Advanced scorers] = 34.36% vs. Spanish [Advanced scorers] = 28.19%), yet when the effect of major was taken into account, a greater performance discrepancy was observed in the Spanish group between language-only and hybrid language majors than in the French group. For instance, the share of high-achieving students (i.e., Advanced levels or higher) in language-only majors

Table 5 ACTFL RPT levels by major

Group	Language	S	AH	AM	AL	IH	IM	IL	NH	NM	NL	**Total**
Hybrid language major	French	0	2	24	63	36	33	20	7	0	1	**186**
	Russian	0	0	0	3	0	11	1	4	1	0	**20**
	Spanish	12	9	69	216	102	103	29	13	0	0	**553**
	Total	**12**	**11**	**93**	**282**	**138**	**147**	**50**	**24**	**1**	**1**	**759**
Language-only major	French	0	0	7	22	9	3	0	0	0	0	**41**
	Russian	0	0	0	0	0	0	1	0	0	1	**2**
	Spanish	2	3	24	30	12	7	2	2	0	0	**82**
	Total	**2**	**3**	**31**	**52**	**21**	**10**	**3**	**2**	**0**	**1**	**125**

surpassed that of the hybrid language majors by 14.62% in French, whereas the difference was as much as 22.24% in Spanish. The RPT was the only test on which about half or more of the students in both language major types (language-only and hybrid language) achieved at or above the Advanced level. A closer examination showed that language-only majors again performed better than hybrid language majors on reading: In the Spanish program, 71.95% of the language-only majors reached the Advanced level or higher, while only 55.33% of the hybrid-language majors did, whereas in the French program, 70.73% of the language-only majors reached Advanced or higher, and only the 47.85% of the hybrid-language majors reached an Advanced or higher level.

To summarize, the majority of the sampled foreign language majors had stronger reading skills than speaking or listening skills, although this differed across different language programs. Regardless of the type of language skill under examination, language-only majors demonstrated stronger proficiency than hybrid language majors both within and across different foreign language majors. In the next section, we investigate the factors that contribute to the higher proficiency gains in the language-only major group (over the hybrid-language major group). For example, are language-only majors obtaining higher proficiency levels because they study abroad more often than hybrid-language majors do? Or is it because they comprise more heritage language learners than the hybrid-language major group does? Or might it be because language-only majors are more motivated to do well in their language classes?

6 Predictors of Proficiency

6.1 Background on Predictors of Proficiency Analysis

Research on second/foreign language acquisition has suggested an array of factors that might affect learning outcomes. Similar to Carroll's 1967 study, we were interested in knowing to what extent heritage-speaker status and study-abroad experience contribute to predicting the outcomes of foreign language acquisition. We

added on an examination of several aspects of motivation. We briefly explain the coding of each of the three categories of independent variables (heritage status, motivation, and study abroad), below.

Due to the lack of an explicit heritage-speaker indicator, we collapsed three categorical variables measuring out-of-school language exposure (i.e., family members, friends, and communities) into one measure of heritage language exposure. Students scored either a 1 or a 0 on these survey items, depending on whether they had received language input through the heritage-related source specified in each item. The sum of scores on these three variables constituted the student's final heritage-speaker score. In total, four levels were attainable, ranging from non-heritage speaker (i.e., a score of 0), heritage speaker by one standard (i.e., a score of 1), and up to heritage speaker by all three standards (i.e., a score of 3).

The motivation variable consisted of nine indictors (9 binary variables) describing the purpose for which the student decided to learn a foreign language at the college level. In the context of foreign language learning, it seemed plausible that some types of motivation would have greater influence on learning outcomes than others, as found by prior researchers (for a review, see Ushioda & Dörnyei, 2012). In this study, each of the nine motivation types was a binary indication of that motivation: the nine were not mutually exclusive, which means that if one student was motivated in multiple ways to learn a foreign language, he or she could indicate that by choosing multiple motivations.

The last group of predictor variables contained only one categorical indicator of study-abroad experience. Students obtained a 1 or a 0 on this variable depending on the presence or absence of a study-abroad program in their past experience.

The three dependent variables in this analysis are the 1 to 10 scores achieved on the proficiency tests by the students in speaking (OPIc), listening (LPT), and reading (RPT). We used the same Kenyon and Malabonga (2001) scale as above (transforming results to a 10 point, ranked-ordered scale). Because we had a large body of independent variables to test, and because our goal was to find the most parsimonious models, we used backward regression in our analysis of the data using SPSS 23.0. The analysis began with the full set of independent variables, and in each step, the variable associated with the least reduction in overall R-square was removed until every independent variable left in the model had a significant p value. In this way, the highest overall R-square was guaranteed upon the elimination of redundant model predictors. Such analysis was performed independently for each of the three dependent variables.

6.2 Results

In total, there were 270 unique student cases (majors at MSU with full test and background data) in this dataset. All were all in their senior years when they took the tests. Descriptive analysis of the data showed that over 50% of the students ($n = 153$) had been on a study-abroad program, but only 26.9% ($n = 70$) had received

exposure in the target foreign language through family members (n = 56), friends (n = 10), or communities (n = 4). Variation was also present among the students in terms of their purposes for learning the target language (their motivation). The most common reasons listed by the students included learning the language for (1) professional purposes (N = 230; 85.2%), (2) expanding cultural knowledge (n = 207; 76.7%), and (3) traveling to a country where the target language is spoken (n = 206; 76.3%), followed by slightly less common ones such as learning the language (4) for fun (n = 189; 70%) and (5) for the purpose of completing a graduation requirement (n = 123, 45.6%). The least popular motivation categories were found to be (6) communicating with friends (n = 66; 24.4%), (7) preparing for studying abroad (n = 46; 17%), and (8) learning about one's heritage (n = 27; 10%).

Summary statistics on the students' test scores are displayed in Table 6. The N sizes for all three tests were smaller than 270 due to the presence of missing values. In the OPIc, one test taker received an "AR" (above range) score and 29 students received "BR" ratings (below range): these were coded as missing data, totaling 30 cases. The missing data in the LPT were composed of 59 unreported test scores and 65 BRs, whereas in the RPT, there were 68 unreported test scores and 57 BRs, and all of these were likewise treated as missing. A comparison of mean scores (only with those without missing data) across tests showed that the students performed slightly better on the RPT (M = 6.19) than on the OPIc (M = 5.02) or the LPT (M = 5.25), although with greater variation on the RPT than on the other two tests. The skewness and kurtosis indices, together with the minimal and the maximal values, indicated the scores on all three tests were relatively normally distributed and were spread out along the entire ACTFL scale represented by 1 to 10.

To save space, we do not explain the multiple steps involved in each regression analysis (these can be obtained by emailing the lead author). However, we present in Table 7 the R-square, unstandardized B values, standardized beta values, and p-values that were produced in the last step. Sample sizes varied across the analysis of the OPIc, LPT, and RPT data due to the unequal numbers of missing values on each measure. Only a small subset of the independent variables turned out to be significant predictors in the regression models tested. Coincidentally, each backward regression analysis left three significant predictors, and except for the purpose of completing a language requirement, all the significant independent variables displayed substantial predictive power for two language skills. We now take a closer look at the results of each of the three (OPIc, LPT, and RPT) regression analyses.

Regarding the OPIc, three significant predictors were revealed: study-abroad experience, level of heritage language exposure; and learning the language for the purpose of travel. Together these three variables explained 24.8% of the total

Table 6 Summary statistics for OPIc, LPT, and RPT scores

	N	*Min*	*Max*	*M*	*SD*	Skewness	Kurtosis
OPIc	240	1	9	5.02	1.38	0.281	0.225
LPT	146	1	8	5.25	1.63	−0.901	0.421
RPT	145	1	10	6.19	1.81	−0.772	0.773

Table 7 Summary of results for regression analyses with OPIc, LPT, and RPT scores

Significant predictors		Assessment (N, R-Square)		
		RPT (145, 0.182)	**LPT** (146, 0.217)	**OPIc** (240, 0.248)
Study abroad	*B*	0.783	NS	0.669
	Beta	.214		.241
	p value	.006		<.001
Heritage level	*B*	NS	0.636	0.658
	Beta		.265	.306
	p value		.001	<.001
Learning language for fun	*B*	1.253	0.754	NS
	Beta	.306	.216	
	p value	<.001	.006	
Learning language for travel	*B*	NS	1.068	0.698
	Beta		.265	.218
	p value		.001	<.001
Learning language to complete requirement	*B*	−0.658	NS	NS
	Beta	−.182		
	p value	.018		

Note: p values are significant below .05, NS = not significant, *B* = standardized Beta value

variance in the outcome variable (OPIc level). The unstandardized *B* values showed that the three factors of (a) having study-abroad experience (as opposed to not having such experience), (b) having an additional source of heritage language exposure (as opposed to not having it), and (c) learning the language for the purpose of traveling to a country where the language is spoken (as opposed to not learning the language for such a reason) were associated with (a) 0.669, (b) 0.658, and (c) 0.698 point increases in the OPIc score, respectively. The numeric increase, when translated into real life, denotes approximately a full level jump on the ACTFL proficiency scale. For example, with all other things being equal, a student who is learning a foreign language for travelling reasons tends (in this data set) to score one level higher on the 1 to 10 ACTFL OPIc scale than a student who does not have such motivations. Such a difference might sound small when considering the two ends of the scale, such as NL versus NM or AM versus AH, but when it comes to the middle range, real-world values attached to the scores might lead to distinct consequences for stakeholders. For example, K-12 French and Spanish foreign language teacher candidates in Michigan currently need to achieve a rating of at least AL on the ACTFL OPI or OPIc speaking test to be eligible for official state certification to teach, so the difference between the rating of an IH and AL is an important one.

One's level of heritage language exposure and learning the language for both travel and fun explained a significant 21.7% of the total variance in the LPT scores. While learning the language for the purpose of traveling to a country where the language is spoken (as opposed to not learning the language for such a reason) was associated with a 1.068 point increase on the LPT, having an additional source of heritage language exposure or learning the language for fun (as opposed to not learning the language for such reason) was associated with a 0.636 or a 0.754 point increase, respectively. The real-world interpretations for the numeric increase in OPIc apply in the case of LPT as well.

The only negative predictor was found in the model for the RPT scores. While study-abroad experience and learning the language for fun (as opposed to not learning the language for fun) both contributed positively to the improvement in reading skills, learning for the purpose of completing graduation requirement (as opposed to the absence of such a motivation) was associated with a 0.658 point *decrease* in the RPT score. At first glance, a negative motivation-based predictor seemed counterintuitive because generally, we would expect the presence of a measure of motivation to exert positive effects on learning outcomes. However, an examination of the different *types* of motivation reveals an important distinction between learning for the purpose of completing graduation requirement and other learning purposes, particularly those that were found to be significant and positive predictors in the regression analyses. Contrary to learning a language for fun or travelling abroad, where the main motivation comes intrinsically from a learner's genuine interest in a foreign language, country, or culture, learning the language for the purpose of meeting graduation requirement is supported by extrinsic motivation imposed on the learner from the outside world (for more on intrinsic versus extrinsic motivation in second language learning, see Ryan & Deci, 2000), which in this case appears to be a negative stick, rather than a positive carrot. The results thus underline the benefits of intrinsic motivation above and beyond extrinsic motivation (in this case) and suggest that foreign language programs at the college level should aim to foster more intrinsic motivation in students, especially because intrinsic motivation appears to support (or go hand and hand with) learner autonomy, a trait essential for sustaining motivation to learn and for promoting active participation in language learning classrooms (Ushioda, 2011).

7 Discussion

When comparing Carroll's 1967 results to our present-day results, the findings from the first part of our study shows a remarkably similar picture when considering proficiency abilities in the skills of reading, listening, and speaking. Reading skills still largely surpass other skills for those graduating with a major in a language. It is likely that the current upper-level emphasis on reading literature, particularly in third and fourth year classes, accounts for this result. This (third and fourth year) is when literary theory is often taught, and it is sometimes taught through the reading

of canonical, "standard-setting works of literature" (Saussy, 2005, p. 17). This approach to language learning is known as the "canon approach" (Saussy, p. 18), and one rationalization for this approach is that a language learner should be "able to anticipate in one's mind the probable reactions of a native speaker," and that, according to Saussy, can only come through the learning of the classics (p. 21). On the other hand, the high reading yet lower scores in listening and speaking found in this study may reject a purely canonical approach to language learning at the upper levels and rather endorse the growing realization that it is time to make "more room in the major for nonliterary courses," as described by Jrade (2009, p. 86), with foreign language programs needing "the inclusion of learning experiences that draw on cultural studies, film, and service-learning opportunities; and practical courses that tie into a student's professional aspirations" (p. 87). Such recommendations align with those from Pope (2008), who asked if language programs could reinvent the language major so that it includes "real seminars, research groups, discussion groups, exhibits, practical projects, and so on" (p. 25). He challenged language programs to change and suggested that they can do so by asking themselves questions like the following:

- Do we ask ourselves what the needs of our particular students are?
- Have we tracked what our students do with their majors?
- Have we asked ourselves what skills they [students] want to have and what information they need?

Pope continued to note that in addition to studying abroad, language majors today may need additional experiences, such as joint ventures with local K-12 schools, community colleges, or other universities. Doing so may provide proficiency growth that mirrors the growth gained during study abroad, an endeavor that is often too expensive for undergraduates, or not possible, especially given the burden of required, on-campus courses needed to fulfill the requirements of their other (non-language) major or majors.

In this study with college foreign language majors, we found that study abroad, heritage status, and intrinsic language-learning motivation contributed to higher proficiency attainment. Looking at the three variables, we now zero in on the one that language programs can most probably and easily address: intrinsic motivation. As described by Byrnes (1988), most foreign language majors declare their major in the second half of their sophomore year. She noted that it may be good for foreign language programs to encourage language learners to declare a language major earlier so that they move earlier from fulfilling a language requirement to taking a foreign language on a volunteer or elective status (that is, they switch from extrinsic to intrinsic language learning motivation). She noted that "language programs should exert every effort to identify majors early on, preferably upon entry during the freshman year" (p. 37). Such efforts could help programs identify enough students for a majors-track, or at least a majors-club, within the language program, which could help students obtain, early-on, a sense of camaraderie, belonging, and autonomy in learning. In addition, the kinds of suggestions made by Jrade (2009) mentioned above (e.g., including a greater emphasis on cultural studies, film, and

service learning) are precisely the kinds of activities that are likely to increase motivation because they tap into students' existing interests.

As in Carroll's (1967) study, we also found that study abroad impacts language learning. This is no surprise, as many studies have found the same results; this has not changed in over 50 years. But what has changed is the duration of study abroad programs: Study abroad programs have shifted overall from majority academic-year programs, to shorter, one semester or even 8-, 6-, 4-, or 2-week intensive summer study abroad programs (Davidson, 2010; Dwyer, 2004). Again, this shift may be due in part to the changing nature of the foreign language student: As shown in this study, few are solely language majors. Most majors in our study were hybrid majors, who often must take pre-requisite courses on campus to fulfil the requirements for their non-language majors alongside those of their language major, requirements that may preclude them from the ability to study a full semester or academic year abroad without impacting their length of overall study. A silver lining on the study abroad literature is that even though shorter programs provide less and fewer linguistic gains (as evidenced by Davidson, 2010; Dwyer, 2004), they still provide a motivational boost that positively impacts language learning overall and for a sustained period of time (Kinginger, 2013). The gains can be in the creation of social networks that sustain learning and connection to the culture, which in turn spur increased or broader opportunities for language development post study abroad. (See Sanz & Morales-Front, 2018, for a discussion on study abroad and second language learning.)

A second finding of Carroll's that garners continued support in our data is the positive impact that starting language learning early has on ultimate proficiency. As noted by Duvick (2002), oftentimes in the United States the foreign language majors are made in the high school foreign language class. High school language learners are anxious to continue studying in college, to participate in study abroad programs, and even major in the language. Duvick (p. 78) noted that high school students may look at a college foreign language program and ask early, even before admission, about what they will be able to do with the language on campus, and their parents may ask about what their adult children will be able to do with the foreign language major post graduation. Duvick noted that the programs need to be prepared for such questions. He wrote that it is increasingly clear that:

> Foreign language programs are strengthened when they can answer that prospective student's question, when they can provide opportunities for students to link their interest in foreign language and culture (in its broadest sense) to distinct career paths (p. 78).

He opined that one thing programs can do preemptively is to enter into collaborative arrangements with other academic units. And we also believe this would be wise. Foreign language programs should collaborate with the academic units that commonly share (hybrid) majors with the language programs. Indeed, some language and non-language programs are collaborating at Michigan State University to offer special-topic and interdisciplinary study-abroad programs that integrate language instruction and hands-on practical and content learning, such as a summer exchange program to Quito, which includes a program sponsored by the Spanish program within the College of Arts and Letters and the Anthropology program

within the College of Natural Sciences. Hybrid anthropology and Spanish majors can spend a summer with a home stay family in Quito while also taking part in experiential (on-location) anthropological research. Such programming and collaboration make study abroad possible for more majors, especially hybrid majors, as they can, conceivably, gain credit that may be honored across their two majors (a two-for-one benefit).

We further see the benefits in starting early, as outlined by Carroll (1967): Another paper by Isbell, Winke, and Gass (2018) found that having taken the foreign language prior to college entrance helps one with overall proficiency. Students, the longitudinal study shows, who come into programs with prior learning can advance (grow) in their proficiency more than those who start in college. This underscores the importance of high schools as a venue for recruiting foreign language majors, but note that Kym (2011) warned that high school language programs are dwindling; thus, university and college language programs must not solely rely on them as consistent or sustainable feeders. We too have heard this warning cry, as in Michigan new computer programming classes in high schools are substituting for or even replacing foreign language requirements, and such trends may only continue as programming becomes more important and high-school programming curricula are created and put into place through legislation, which is at times backed by technology companies. Foreign language programs may be pushed in coming years to be viewed more and more as programs on the sidelines, those relegated to a broad humanities-based education, unless the programs can tout themselves as offering strong, interdisciplinary majors, ones partnered with other programs in science, technology, health, communications, and math. Such links with other programs may make hybrid language majors with a foot in STEM (science, technology, economics, and math) and other areas more globally focused, and graduates from such programs more locally and internationally employable.

Our research, like Carroll's (1967), questions a strong, traditional emphasis on literature for majors. It is likely that in some foreign language programs, a strong emphasis on literature in the upper levels coincides with less upper-level instruction in speaking and listening. Byrnes (1988) lamented that an emphasis on literature in upper-level courses before students are ready (proficiency-wise) to take such classes often leads to the teachers reverting to English in class, a problematic result that may still occur, as evidenced more recently by Zyzik and Polio (2008). Zyzik and Polio observed Spanish literature courses at Michigan State University. They categorized the type of interactions that occurred between the professors and the students, and noted that there was a lack of opportunity for the students to speak in anything but short utterances. Instead, teacher-talk dominated the lessons, and instructors were concerned with covering the content in a limited amount of time, and asked students questions to check their comprehension. Zyzik and Polio's observations combined with our proficiency measures of majors at the same institution seem to suggest that students in upper level foreign language classes are, indeed, still language learners, and that they need linguistic, socio-pragmatic, as well as content-based instruction. They need practice using the foreign language across all skills: listening, speaking, reading, and most probably, writing, although

we did not assess that skill in this study: See Bernhardt et al. (2015) for information on how college-level program directors can develop and sustain writing proficiency assessment at their institutions.

Teaching all skills and content-areas across the entire four-year, undergraduate curriculum may help to advance robust foreign language learning, one that produces students who are truly *Advanced* in their skills, a designation that eludes many college graduates majoring in a foreign language today (Tschirner, 2016). The goal is not to take literature and culture *out* of the language program, but rather, as explained by Kym (2011, p. 44) to greatly increase the emphasis on language learning in the current literature and culture classes. By definition, advanced language learners can use the language contextually in sophisticated ways (Ortega & Byrnes, 2008, p. 8); Advancedness in a foreign language is associated with "aspects of literacy, to diverse manifestations of cultural competence, choice among registers and multiple speech community repertoires, voice, and identity in cross-cultural communicative settings," they wrote. Thus, it is obvious that majors, who are, as a goal, to become Advanced language learners and beyond, need more than instruction on canonical texts: They need (a) interactive classes, (b) the ability to join active research groups and real-world (online or face-to-face) group discussions, (c) the facility, materials, and space to create hands-on exhibits, and (d) concrete experience in creating practical projects for other language learners and community members that use or need to use the language: that is, the entire rich world of language-learning equipment and experiences that Pope (2008), Jrade (2009), and Zyzik and Polio (2008) called for.

One large difference between Carroll's data and ours was the nationwide scope of his and the local nature of ours. His study covered 15 medium to large institutions who participated voluntarily. He had a sophisticated sampling procedure, but did rely on voluntary institutional buy-in. And, within those institutions, students volunteered to take his test battery. In our case, MSU 'volunteered' in the sense that the PIs applied for funding to undertake this study, but different from Carroll's students who volunteered, our students were required to take the tests as part of course requirements. We relied on department chairs, language program coordinators, and individual instructors to allow us access to their classes. Even though students were required to participate in the testing as part of class, it was the instructor who determined whether participation was part of a student's grade or, perhaps, extra credit.

We, of course, do not know if any differences between our study and those of Carroll are due to the students' differences in volunteering, but we do note that Carroll attempted to determine whether the volunteer versus non-volunteer status was significant. He examined the transcripts of the 237 foreign language majors in his participating institutions to determine their foreign language grade point averages (GPAs). He compared those GPAs with 284 students who had not volunteered, and he found a significant difference in only 6 of the 15 institutions with stronger GPAs amongst the tested students. Perhaps most interesting was his finding that, contrary to what one might expect, those who opted to take the test were not always the stronger students. In two university settings (with large numbers of students tested), those who opted to take the tests had lower grade point averages than those who did not test.

We also want to note that there is a discussion on whether there has been a drift in the ACTFL scale in how it is mapped onto the ILR scale; that is, whether the scale-level correspondences have shifted over the years since the ACTFL scale was first conceived (and created in reference to the ILR scale) in the 1980s (personal communication, Liskin-Gasparro, Oct. 27, 2017). An original interpretation of the ACTFL scale in the early 80s was that Superior on the ACTFL scale was a 3 (working-level ability) on the ILR scale, but now, conventionally, a 3 on the ILR scale is considered to map onto Advanced Low on the ACTFL scale. Whether this drift is real is a matter of empirical investigation (and could be studied using archives of rated speaking test data). But if it is real, then applied linguists may need to know the size and scope of the drift to best understand how to compare today's proficiency data with data from 50 years ago.

These issues aside, our study shows that Carroll's (1967) work was groundbreaking, but his work was not finished in 1967. Many of the questions he pursued then are being pursued now. How far can we go? What can foreign language majors obtain in terms of foreign language proficiency? How prepared are they for a globalized and industrialized world? What pushes them forward, and what pushes them back? We add nuances to these questions, such as how do the changing natures of foreign language programs, curricula, and the students themselves contribute to the attainment of language proficiency at the tertiary level? We believe that while we found many common threads across the two studies that bridge 50 years of research, more research is needed. More pictures of proficiency at the college level are needed, and different cameras (tests, observation protocols, self-assessments, and portfolios) should be used so that our results can be triangulated and tested for methodological rigor. This is important, because researchers (e.g., Liskin-Gasparro, 1995) have long suggested that majors have language skills that are difficult to assess, and some of the genres they are knowledgeable about are best assessed through extensive portfolios or senior research theses. When the assessments and robust depictions of proficiency are strung together, we may eventually be able to see our moving trajectory rather than just our moment-in-time trends.

References

ACTFL. (2012). *ACTFL proficiency guidelines*. Alexandria, VA: ACTFL. Available from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Bernhardt, E., & Brillantes, M. (2014). The development, management, and costs of a large-scale foreign language assessment program. In N. Mills & J. Norris (Eds.), *Innovation and accountability in language program evaluation* (pp. 41–61). Boston, MA: Cengage Learning.

Bernhardt, E., Molitoris, J., Romeo, K., Lin, N., & Valderrama, P. (2015). Designing and sustaining a foreign language writing proficiency assessment program at the postsecondary level. *Foreign Language Annals, 48*(3), 329–349. https://doi.org/10.1111/flan.12153

Byrnes, H. (1988). How do you get there from here? Articulating the foreign language major program. *ADFL Bulletin, 20*(1), 35–38. https://doi.org/10.1632/adfl.20.1.35

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

Clément, R., & Kruidenier, B. G. (1983). Orientations in second language acquisition: The effects of ethnicity, milieu, and target language on their emergence. *Language Learning, 33*(3), 273–291. https://doi.org/10.1111/j.1467-1770.1983.tb00542.x

Davidson, D. E. (2010). Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals, 43*(1), 6–26. https://doi.org/10.1111/j.1944-9720.2010.01057.x

Duvick, R. J. (2002). Sustaining foreign language enrollments through collaboration: An interdisciplinary major. *ADFL Bulletin, 33*(2), 78–80. https://doi.org/10.1632/adfl.33.2.78

Dwyer, M. M. (2004). More is better: The impact of study abroad program duration. *Frontiers: The Interdisciplinary Journal of Study Abroad, 10*, 151–164.

Herman, G. (1987). From dual to single track: Revision of the undergraduate French major at the University of California, Davis. *ADFL Bulletin, 18*(3), 25–27. https://doi.org/10.1632/adfl.18.3.25

Holmquist, J. C. (1993). Social and psychological correlates of achievement: Spanish at Temple University. *The Modern Language Journal, 77*(1), 34–44. https://doi.org/10.1111/j.1540-4781.1993.tb01942.x

Isbell, D., Winke, P., & Gass, S. (2018). Using the ACTFL OPIc to assess proficiency and monitor progress in a tertiary foreign languages program. *Language Testing*. https://doi.org/10.1177/0265532218798139

Jrade, C. L. (2009). Assessing the present foreign language major and offering strategies to improve it. *ADFL Bulletin, 41*(2), 83–87. https://doi.org/10.1632/adfl.41.2.83

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology, 5*(2), 60–83. Retrieved from http://llt.msu.edu/vol5num2/kenyon/

Kinginger, C. (Ed.). (2013). *Social and cultural aspects of language learning in study abroad*. Amsterdam: John Benjamins.

Kym, A. (2011). Curricular change and the major. *ADFL Bulletin, 41*(3), 43–47. https://doi.org/10.1632/adfl.41.3.43

Lafford, B. A. (2004). The effect of the context of learning on the use of communication strategies by learners of Spanish as a second language. *Studies in Second Language Acquisition, 26*(2), 201–225. https://doi.org/10.1017/S0272263104262039

Liskin-Gasparro, J. E. (1995). Practical approaches to outcomes assessment: The undergraduate major in foreign languages and literatures. *ADFL Bulletin, 26*(2), 21–27. https://doi.org/10.1632/adfl.26.2.21

Lusin, N. (2009). Are you counting second majors in foreign languages? *ADFL Bulletin, 41*(2), 105–107. https://doi.org/10.1632/adfl.41.2.105

Magnan, S. S. (1986). Assessing speaking proficiency in the undergraduate curriculum: Data from French. *Foreign Language Annals, 19*(5), 429–438. https://doi.org/10.1111/j.1944-9720.1986.tb01031.x

Oller, J. W., & Nagato, N. (1974). The long-term effects of FLES: An experiment. *The Modern Language Journal, 58*(1–2), 15–19. https://doi.org/10.1111/j.1540-4781.1974.tb05072.x

Ortega, L., & Byrnes, H. (Eds.). (2008). *The longitudinal study of advanced L2 capacities*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pope, R. D. (2008). The major in foreign languages: A four-pronged meditation. *ADFL Bulletin, 40*(1), 24–26. https://doi.org/10.1632/adfl.40.1.24

Rifkin, B. (2005). A ceiling effect in traditional classroom foreign language instruction: Data from Russian. *The Modern Language Journal, 89*(1), 3–18. https://doi.org/10.1111/j.0026-7902.2005.00262.x

Robinson, J. P., Rivers, W. P., & Brecht, R. D. (2006). Speaking foreign languages in the United States: Correlates, trends, and possible consequences. *The Modern Language Journal, 90*(4), 457–472. https://doi.org/10.1111/j.1540-4781.2006.00462.x

Rosengrant, S. F. (1987). Error patterns in written Russian. *The Modern Language Journal, 71*(2), 138–146. https://doi.org/10.1111/j.1540-4781.1987.tb01595.x

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*, 54–67. https://doi.org/10.1080/17501229.2011.577536

Sanz, C., & Morales-Front, A. (Eds.). (2018). *The Routledge handbook of study abroad research and practice*. New York: Routledge.

Saussy, H. (2005). Language and literature on the pedagogical continuum; or, life begins after proficiency. *ADFL Bulletin, 36*(2), 17–21. https://doi.org/10.1632/adfl.36.2.17

Schumann, J. H. (1975). Affective factors and the problem of age in second language acquisition. *Language Learning, 25*(2), 209–235. https://doi.org/10.1111/j.1467-1770.1975.tb00242.x

Spada, N. (1986). The interaction between type of contact and type of instruction: Some effects on the L2 proficiency of adult learners. *Studies in Second Language Acquisition, 8*(02), 181–199. https://doi.org/10.1017/S0272263100006070

Spolsky, B. (1969). Attitudinal aspects of second language learning. *Language Learning, 19*(3–4), 271–275. https://doi.org/10.1111/j.1467-1770.1969.tb00468.x

Tschirner, E. (1996). Scope and sequence: Rethinking beginning foreign language instruction. *The Modern Language Journal, 80*(1), 1–14. https://doi.org/10.1111/j.1540-4781.1996.tb01132.x

Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals, 49*(2), 201–223. https://doi.org/10.1111/flan.12198

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. (2016a). *Digest of Education Statistics* [Web only table] Table 325.57 Degrees in French, German, Italian, and Spanish language and literature conferred by postsecondary institutions, by level of degree: Selected years, 1949–50 through 2014–15. Retrieved from: https://nces.ed.gov/programs/digest/d16/tables/dt16_325.57.asp?current=yes

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. (2016b). *Digest of Education Statistics* [Web only table] Table 325.59 Degrees in Arabic, Chinese, Korean, and Russian language and literature conferred by postsecondary institutions, by level of degree: 1969–70 through 2014–15. Retrieved from: https://nces.ed.gov/programs/digest/d16/tables/dt16_325.59.asp?current=yes

Urlaub, P. (2014). Departmental contexts and foreign language majors. *ADFL Bulletin, 43*(1), 123–134. https://doi.org/10.1632/adfl.43.1.123

Ushioda, E. (2011). Why autonomy? Insights from motivation theory and research. *Innovation in Language Learning and Teaching, 5*(2), 221–232. https://doi.org/10.1080/17501229.2011.577536

Ushioda, E., & Dörnyei, Z. (2012). Motivation. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 396–409). New York: Routledge.

Wong, W., & Van Patten, B. (2003). The evidence is IN: Drills are OUT. *Foreign Language Annals, 36*(3), 403–423. https://doi.org/10.1111/j.1944-9720.2003.tb02123.x

Zyzik, E. C., & Polio, C. (2008). Incidental focus on form in Spanish literature courses. *The Modern Language Journal, 92*, 50–73. https://doi.org/10.1111/j.1540-4781.2008.00686.x

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*.

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*.

Emily S. Heidrich is a Project Manager and Academic Specialist at the Center for Language Teaching Advancement at Michigan State University. She received her PhD in German and Second Language Acquisition from the University of Wisconsin - Madison. Her research interests include foreign language proficiency, educational technology, curriculum design, and education abroad topics.

Part III
Assessments and Learning Outcomes

In Advanced L2 Reading Proficiency Assessments, Should the Question Language Be in the L1 or the L2?: Does It Make a Difference?

Troy L. Cox, Jennifer Bown, and Teresa R. Bell

Abstract When investigating foreign language (FL) proficiency in reading in higher education, one must first determine what proficient reading entails and how to operationalize it. The American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines provide a starting point in this process, but they do not provide instructions for assessing reading. Clifford and Cox (Foreign Lang Ann 46(1):45–61, 2013) define proficient reading as "the active, automatic, far-transfer process of using one's internalized language and culture expectancy system to efficiently comprehend an authentic text for the purpose for which it was written (p. 50)." According to this definition, reading is an asynchronous, written two-way interaction between author and reader, in which the reader's primary task is to comprehend the author's intent. However, since the cognitive processes involved in reading cannot be directly observed, researchers use observable tasks (e.g., answering questions, reading aloud, etc.) to make inferences about the FL learner's reading proficiency. Shohamy (Lang Test 1(2):147–170, 1984) notes that this reliance on indirect methods of assessment places a "heavy burden on the testing method and therefore may create greater variations in scores obtained as a result of these methods" (p. 149). Thus, researching how test method affects test scores is paramount to ensure that any variance in scores is due to differences in proficiency rather than choice of test method. In designing tasks to assess reading comprehension, the issue of question language (QL) arises. That is, scholars must decide whether the QL should be in the same language as the reading passage—the learners' second language (L2) or in the native language (L1) of the learner. When the QL is in the L1, it is easier to infer what the reader has understood. When the QL is in the L2, the responses are dependent on the examinees' comprehension of both the questions and the text. However, as L2 learners gain reading proficiency, they should also better be able to comprehend questions in the L2. The present study sought to fill

T. L. Cox (✉)
Center for Language Studies, Brigham Young University, Provo, UT, USA
e-mail: troyc@byu.edu

J. Bown · T. R. Bell
Department of German and Russian, Brigham Young University, Provo, UT, USA
e-mail: jennifer_bown@byu.edu; tbell@byu.edu

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_7

these gaps in the research literature by examining the effect of QL on the scores of advanced readers of Russian on a criterion-referenced test of reading proficiency. Understanding the effect of QL on readers with Advanced-level proficiency will allow practitioners to make more informed decisions about design of reading assessments in general and of high-stakes, criterion-referenced tests of reading proficiency in particular.

Keywords Russian · Russian reading proficiency · Reading proficiency · Question language · Reading proficiency test · Learner perception

1 Introduction

When investigating foreign language (FL) proficiency in reading in higher education, one must first determine what proficient reading entails and how to operationalize it. The American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines provide a starting point in this process, but they do not provide instructions for assessing reading. Clifford and Cox (2013) define proficient reading as "the active, automatic, far-transfer process of using one's internalized language and culture expectancy system to efficiently comprehend an authentic text for the purpose for which it was written" (p. 50). According to this definition, reading is an asynchronous, written two-way interaction between author and reader, in which the reader's primary task is to comprehend the author's intent. However, since the cognitive processes involved in reading cannot be directly observed, researchers use observable tasks (e.g., answering questions, reading aloud, etc.) to make inferences about the FL learner's reading proficiency. Shohamy (1984) notes that this reliance on indirect methods of assessment places a "heavy burden on the testing method and therefore may create greater variations in scores obtained as a result of these methods" (p. 149). Thus, researching how test method affects test scores is paramount to ensure that any variance in scores is due to differences in proficiency rather than choice of test method.

In designing tasks to assess reading comprehension, the issue of question language (QL) arises. That is, scholars must decide whether the QL should be in the same language as the reading passage—the learners' second language (L2) or in the native language (L1) of the learner. When the QL is in the L1, it is easier to infer what the reader has understood. When the QL is in the L2, the responses are dependent on the examinees' comprehension of both the questions and the text. However, as L2 learners gain reading proficiency, they should also better be able to comprehend questions in the L2.

The relationship between QL and reading comprehension scores was first studied by Shohamy in 1984. In her large-scale study involving 655 Israeli high school students learning English as an L2, Shohamy tested the effect of multiple choice questions in both L1 and L2 and open-ended questions in both L1 and L2 on learners' scores. She found that in addition to test method (i.e., multiple-choice, written

recall, etc.), QL had a significant effect on students' scores with students scoring lower on tests with questions in the L2. However, she noted that the effect on learners' scores diminished as the learners' skills increased, positing that the difference may be erased entirely for highly proficient L2 readers.

Since Shohamy's study, QL has received sparse treatment from researchers. While some research has examined the effect of QL on examinee scores on norm-referenced tests (Brantmeier, 2006; Godev, Martinez-Gibson, & Toris, 2002, Gordon & Hanauer, 1995; Lee, 1986; Poh & Hock, 1979; Shohamy, 1984), little attention has been given to the effects of QL on criterion-referenced proficiency tests. Proficiency exams are often high-stakes and summative in nature, thus it behooves researchers to understand how the QL can affect learners' scores.

The few existing studies investigating the effect of QL on test scores have primarily focused on students at the beginning levels of language learning (Godev et al., 2002; Lee, 1986; Nevo, 1989; Poh & Hock, 1979; Shohamy, 1984) largely ignoring advanced-level readers (but see Brantmeier, 2006) for whom the effect of the QL may be less significant. Furthermore, the research that has been conducted has focused on commonly-taught L2 s such as English, French, and Spanish.

The present study sought to fill these gaps in the research literature by examining the effect of QL on the scores of advanced readers of Russian on a criterion-referenced test of reading proficiency. Understanding the effect of QL on readers with Advanced-level proficiency will allow practitioners to make more informed decisions about design of reading assessments in general and of high-stakes, criterion-referenced tests of reading proficiency in particular. This study also considered students' affective reactions to the QL.

1.1 Reading Comprehension and Question Language

As a receptive skill, reading comprehension is an internal process dependent on many internal and external factors. The purpose of reading, for instance, affects students' internal processing (Linderholm & van den Broek, 2002; Lorch, Lorch, & Klusewitz, 1993). When readers read with different goals—such as for enjoyment, learning, to evaluation, etc.—they use different internal comprehension processes. In addition, studies have also shown that background knowledge affects learners' reading processes (Anderson, 1991; Brantmeier, 2005; Bügel & Buunk, 1996; Shiotsu & Weir, 2007).

Assessing reading comprehension, whether in a learner's L1 or L2, poses particular challenges. Because comprehension processes take place internally, researchers must infer reading ability through external measures. The difficulty for researchers is that "by attempting to observe the reader's response, we are bound in some way to affect that response" (Harrison & Dolan, 1979, p. 13). For this reason, scholars have considered how different question types affect the comprehension process. In L2 testing, QL becomes another facet that can affect the test takers, their scores, and their attitudes.

Reading Proficiency In 2012, ACTFL released its most recent Reading Proficiency Guidelines which describe five major levels of proficiency that represent a geometric progression of reading skills (e.g., Distinguished, Superior, Advanced, Intermediate, and Novice) (ACTFL, 2012).[1] A study by Clifford and Cox (2013) validated these Guidelines, using a test design which aligned author purpose, reader purpose, and text characteristics. As the authors note, "the fact that a reader can get the main idea (an Intermediate-level task) of a text generated for an Advanced communication purpose does not indicate Advanced reading ability" (p. 60). Instead, at the Advanced level, learners should be able to extract details from the text. Moreover, an Advanced-level text must exhibit characteristics of the Advanced level: the vocabulary must go beyond the high-frequency vocabulary of Intermediate-level texts, and the topics should be of broader interest than those at the Intermediate level.

As Table 1 demonstrates, at the Advanced level, readers can comprehend the details as well as the main ideas of texts, and at the Superior level, readers must be able to read between the lines, determining tone and stance. The range of vocabulary required to perform at this level across a wide variety of topics is significant. As such, the vocabulary used in L2 questions, theoretically, should not pose difficulties for the Advanced- or Superior-level reader, though the vocabulary used in L2 questions quite likely might pose difficulties for Novice- and Intermediate-level learners.

Effects of Question Language on Test Scores Relatively little attention has been given to the effect of QL on reading test scores and even less has been given to the effect of QL on the test scores of advanced-level learners. Moreover, interpreting the results of the prior literature is complicated by the variety of design variables and instruments used in prior studies. See Table 2 for more detailed information on the participants and the types of questions examined. For example, some researchers have examined the effect of QL using multiple choice questions (Gordon & Hanauer, 1995; Nevo, 1989; Poh & Hock, 1979; Shohamy, 1984), while others have used open-ended questions (Godev, Martínez-Gibson, & Toris, 2002; Gordon & Hanauer, 1995; Shohamy, 1984), written recall (Brantmeier, 2006; Lee, 1986), or think aloud protocols (Gordon & Hanauer, 1995). Findings suggest that the type of question affects the difficulty of the task as much as the QL.

In spite of the differences in design, prior research on QL generally suggests that questions in the L1, especially multiple choice questions, are easier to answer than are questions in the L2 (Gordon & Hanauer, 1995; Lee, 1986; Poh & Hock, 1979; Shohamy, 1984), and open-ended questions in the L2 are the most difficult to answer (Godev et al., 2002; Gordon & Hanauer, 1995; Lee, 1986; Shohamy, 1984). In the latter case, the difficulty of responding to open-ended questions in the L2 might be attributed to issues of production, rather than comprehension.

[1] We use capital letters to refer to the ACTFL levels. Elsewhere, lowercase is used for generic labels of learners' abilities.

Table 1 Clifford's assessment criteria—Reading (2016, p. 26)

ACTFL level	Reader task	Author purpose	Conditions		
			Text type	Content	Accuracy expectations
Superior	Understand literal and figurative meanings by reading both "the lines" and "between the lines." Recognize the author's tone and unstated positions. Evaluate the adequacy of arguments made.	Evaluate situations, concepts, and conflicting ideas. Present and support arguments and/or hypotheses with both factual and abstract reasoning.	Multiple-paragraph prose on a variety of professional or abstract subjects such as found in editorials, formal papers, and professional writing.	Multiple, well-organized abstract concepts interlaced with attitudes and feelings. Social/cultural/political issues with abstract aspects and supporting facts presented as well. Most allusions and references are explained by their context.	Understand the facts; the details; and the author's opinion, tone, and attitude.
Advanced	Understand the facts and supporting details including any causal temporal and spatial relationships.	Convey structured, factual information, supporting details, and factual relationships in extended narratives and descriptions.	News reports, magazine articles, short stories, human interest features, and instructional and descriptive materials.	Concrete information about real-world phenomenon with supporting details, as well as interrelated facts about world, local, and personal events.	Grasp both the main ideas and the supporting details.
Intermediate	Understand the main idea, orient oneself by identifying the topic or main idea.	Orient by communicating one or more general ideas.	Very simple announcements, ads, and personal notes.	Information about places, times, people, etc. that are associated with everyday events, personal invitations, or general information.	Recognize the main idea and some broad, categorical distinctions.
Novice	Recognize some random items in a list or short text.	List, enumerate.	Lists, simple tables.	Sparse or random; format or external context may reveal internal relationships.	Correctly recognize some words.

Table 2 Previous QL research

Authors, Year	Level of student	L2	L1	Question types	Participants
Poh and Hock (1979)	Post High School	English	Bahasa Malay	MC	Students described as post fifth form learners of English (n = 39) (fifth form is equivalent to senior year in high school). These students were entering the university.
Shohamy (1984)	Low, Int, and High (High School)	English	Hebrew	MC and Open ended	Students in 12th grade in a high school setting in Israel (n = 655), and they were divided into proficiency groups of low, intermediate, and high.
Lee (1986)	Beg and Int (University)	Spanish	English	Written Recall	Students enrolled in four different semester level (n = 320; 80 participants per level). Spanish classes at Michigan State and University of Michigan for first and second years.
Nevo (1989)	Intermediate (High School)	French	Hebrew	MC	Students in tenth grade in High School in Israel (n = 42).
Gordon and Hanauer (1995)	10th Graders (High School)	English	Hebrew	Think-aloud protocols (MC and open ended)	Students in 10th grade (n = 28) studying English in Israel.
Godev et al., (2002)	Intermediate (University)	Spanish	English	Open ended- with different language for stem and answer	Students in third-semester (intermediate) of Spanish at a university (n = 28).
Brantmeier (2006)	Advanced (University)	Spanish	English	Written Recall	Students enrolled in an advanced-level Spanish grammar and composition course at a private university in the Midwest (n = 106).

The negative effect of QL on test scores may, however, be mitigated by higher reading proficiency. Shohamy (1984) found that the negative effects of L2 questions on test scores diminished as the students' reading skills increased. She posited that advanced L2 learners have acquired a broad enough vocabulary that testing them in the L2 does not impede their performance. Similarly, Brantmeier (2006), in a study

of the effect of task language on the written recalls of 66 advanced-level learners of Spanish, found that QL accounted for only 3% of the variance in the performance of 66 advanced learners of Spanish on a written recall task. However, a sizable 28% of the variance in L2 written recall was attributed to learners with lower levels of reading proficiency, as measured by student scores on the "Romance Languages and Literatures Online Placement Exam." Scholars suggest that the L1 plays a larger role in L2 reading for novice level readers (Corder, 1978; Upton, 1997; Upton & Thompson, 2001). As readers gain proficiency in the L2, they rely less on their L1 to process texts. Thus Bernhardt (2005) asserts that, until readers reach the "highest L2 proficiency/fluency levels" (p.141), assessment should take place in the L1.

Impact of QL on Strategies and Affect Of further interest in testing language proficiency is ensuring that systematic score variance is not due to extraneous factors such as strategies or affect (e.g., confidence, motivation, etc.). In fact, a few studies suggest that the QL may affect the strategies that learners employ while reading and processing. For example, Gordon and Hanauer (1995) found that multiple choice questions offer information to the learners that they may rely on to respond to questions. Nevo (1989) noted that readers faced with multiple choice questions in the L2 were more likely to guess by attempting to match words and phrases from the text and from the questions.

The question of learners' affective responses to QL in tests of reading comprehension has largely been ignored in the research. Shohamy (1984) suggests that test items in the L2 may increase learners' anxiety levels, thus indirectly leading to lower scores, however, she did not empirically test this hypothesis. One study of learners' attitudes towards QL in *listening comprehension* may provide some preliminary insights. As part of a study to examine the difficulty of test questions in the L1 or the L2, Filipi (2012) also surveyed her participants to ascertain their attitudes towards the test questions. She found that a majority of beginning and intermediate students of French (N = 154) and Japanese (N = 194) preferred questions in the L1, generally finding the questions in the L1 to be harder. Whether this holds true in reading comprehension has yet to be seen.

This study sought to shed additional light on the issue of QL, by focusing on advanced-level learners of Russian, a less commonly taught language that has heretofore not been included in studies on QL. Not only does Russian use a non-Roman alphabet, but it is also typologically quite different from English (the L1 used in the majority of the QL studies) and thus may pose unique challenges to the L2 reader. Moreover, we sought to understand learners' affective responses to the QL. The following questions guided our research:

- What effect does QL have on reading comprehension test scores among advanced learners of Russian?
- What are the attitudes of advanced learners of Russian toward QL?

2 Methods

To explore the different effects of QL on reading comprehension scores of advanced learners of Russian and their attitudes, two instruments were created: a reading comprehension exam and an attitudinal survey. Participants were then recruited from upper division (third-year) Russian courses. A counter-balanced design was employed after which the resulting data were analyzed.

2.1 Reading Comprehension Exam

The reading comprehension exam used items that had been previously validated for ACTFL reading proficiency assessments (Clifford & Cox, 2013) in which each L2 reading passage contained a single question (in English, the L1). The instrument consisted of four Advanced-level passages in which the author's purpose was *to inform* and the readers' task was *to understand* details, and sixteen Superior passages in which the authors' purpose was *to persuade* and the readers' task was *to infer* the authors' argument. The existing multiple-choice questions were translated from English into Russian by university faculty (two native and one native speaker of English with Superior-level proficiency in Russian) in order to equalize the item difficulty.

Two forms of the test were created: one in which the examinees responded to the questions in Russian first and the other in which examinees saw the questions in English first. To control for the influence of ordering effects, the order of the questions was constant between the two test forms independent of the QL. The test consisted of two ten-question parts resulting in an exam that was 20 questions. The structure of each part started with one Advanced question followed by eight Superior and ended with another Advanced. This attempted to mimic the structure of an ACTFL Oral Proficiency Interview (OPI) in which examinees warm up and cool down with easier items and attempt the more difficult items in the middle. Once the test was created and the item ordering fixed, two test forms were created using a counterbalanced design. Form A had part 1 with the QL in Russian and part 2 with the QL in English. Form B had part 1 with the QL in English and part 2 with the QL in Russian.

2.2 Attitudinal Survey

To measure participant attitudes two steps were involved. First, after every question, participants were asked to use a 100-point slider scale to rate their confidence in answering the item correctly (very unconfident to very confident) and their anxiety

level with the question (very low to very high). Following the reading test, the participants were asked to complete a post-test survey. This survey was created for this study and consisted of both Likert-scale and open-ended questions. Only the open-ended questions that asked for opinions on the QL were used for this study.

2.3 Participants

The participants were 64 male (N = 51) and female (N = 13) students who were enrolled in a third year Russian language course. The average age of the participants was 21.74 (SD = 1.11) and they had all previously lived in a Russian-speaking country for an average of 22.6 months (SD = 2.92). While this group of students did not take an external proficiency test, prior OPI testing of students in this population has revealed an average speaking proficiency of Advanced-mid. Moreover, the students in the course read a great deal of material at the Advanced level. Thus, it was assumed that they were at least Advanced-level readers. All participants were given $20 for participating, and in order to incentivize the participants to do their best on the exam, an additional $5 compensation was offered for scoring in the Superior range.

2.4 Counterbalanced Design

Participants were randomly placed into two groups based on their arrival to the testing site in order to use a counterbalanced design. Counterbalancing occurs when two or more groups receive the same treatment. To avoid confounding due to ordering effects, the first group took form A with part 1 in English and part 2 in Russian, and the second group took form B with part 1 in Russian and part 2 in English. Without counterbalancing, some might argue that the differences in performance were confounded by the passages that were used instead of the QL. Counterbalanced designs use a repeated measures ANOVA with one between group variable (e.g., group membership) and one within-subject variable (e.g., QL). To answer the first research question, the aforementioned ANOVA was used on the test scores with the dependent varaiable as the test score on each part (correct out of 10 possible points), the within subjects independent variable as the QL of the part of the test and the between subjects variable as the group the participants were randomly assigned. To answer the second research question, another two repeated measures ANOVA were conducted employing the same independent variables. Confidence and anxiety were calculated by averaging the participants' responses on each post-question survey (see Fig. 1) for the two parts of the test. The open-ended responses on the post exam survey were also analyzed for possible trends.

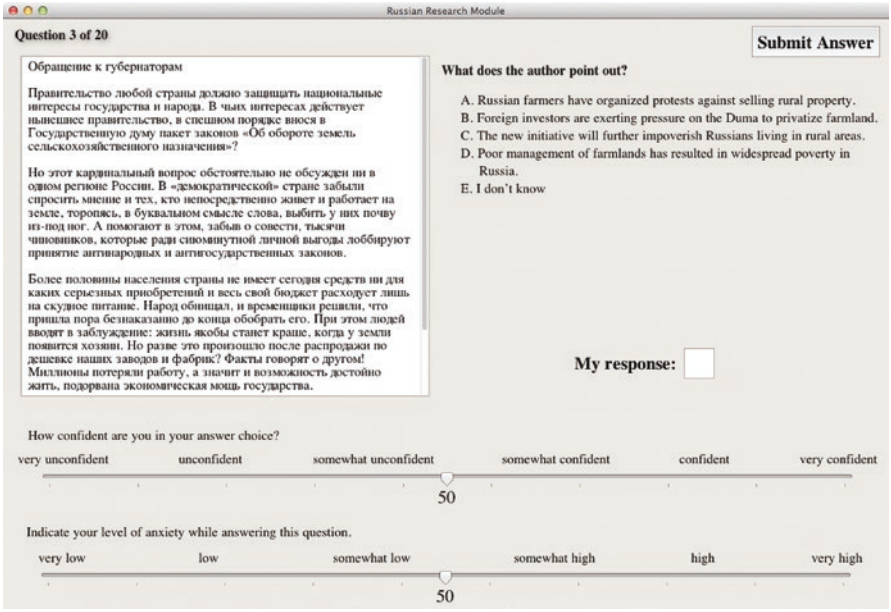


Fig. 1 Screenshot—passage and questions

3 Results

The participants scored substantially higher on items in which the QL was English rather than Russian (see Table 3). A repeated measures ANOVA found that while group 1 performed better than group 2 [$F(1, 62) = 451.88$, $p < .001$] with a large effect size (partial $\eta^2 = .88$), but more importantly, there was no interaction with the between subject variable of group [$F(1, 62) = .22$, $p = .75$] (see Fig. 2) and QL. Thus it did not matter whether participants saw the English questions in part 1 or part 2. Regardless of the actual reading passage, when the QL was in English participants scored higher [$F(1, 62) = 21.47$, $p < .001$,] with a large effect size (partial $\eta^2 = .26$).

Participants were also more confident in answering the question correctly and less anxious when the QL was English rather than Russian with a mean difference of 4.30 in their confidence level and a mean difference of 3.75 in terms of anxiety (Tables 4 and 5).

Intriguingly, the relationship between confidence and actual reading comprehension was stronger when students responded to the questions in Russian ($r = .53$, $p < .001$) as opposed to responding in English ($r = .28$, $p = .027$). In both cases, learners were overconfident in their ability. That is that they assumed they answered the question correctly when they did not. However, they were even more overconfident when the QL was English.

Attitudinal Survey A repeated measures ANOVA found that in the between subjects variable of group, there were no significant differences with either confidence

Table 3 Descriptive statistics of ability level by QL

	English			Russian		
	Group 1	Group 2	Total	Group 1	Group 2	Total
Mean	5.83	5.26	5.53	4.73	4.00	4.34
N	30	34	64	30	34	64
95% Confidence Mean (Min)	4.93	4.51	4.96	4.05	3.29	3.85
Interval for (Max)	6.73	6.02	6.10	5.42	4.71	4.84
Median	6.00	5.00	5.50	5.00	4.00	4.00
Std. Deviation	2.41	2.17	2.28	1.84	2.04	1.97
Minimum	1.00	0.00	0.00	1.00	0.00	0.00
Maximum	10.00	9.00	10.00	8.00	8.00	8.00

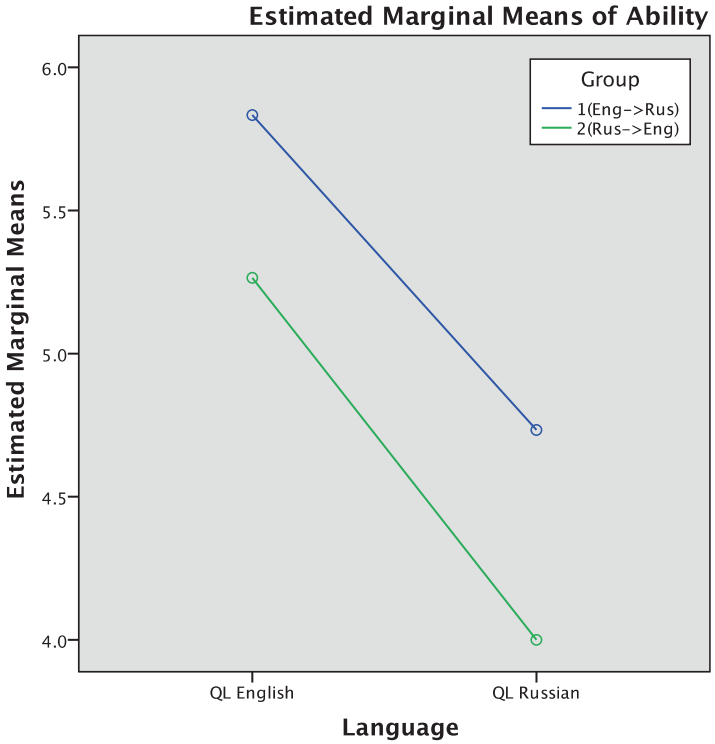


Fig. 2 Estimated marginal means of reading ability by group and QL

($F(1,62) = .002$, $p = .96$, $\eta^2 = .000$) and anxiety ($F(1,62) = .568$, $p = .45$, $\eta^2 = .009$). We found no interaction with the between subject variable of group and QL ($F(1,62) = .26$, $p = .61$, $\eta^2 = .004$) in terms of confidence. When the QL was English, participants were more confident (the within-subject variable) in answering the question correctly ($F(1, 62) = 16.33$ $p < .001$, $\eta^2 = .26$).

Table 4 Descriptive statistics of ability level by confidence by QL

	English			Russian		
	Group 1	Group 2	Total	Group 1	Group 2	Total
Mean	61.29	60.50	60.87	55.68	56.14	55.93
N	56.17	56.06	57.60	48.93	50.95	51.85
95% Confidence Mean (Min)	66.42	64.94	64.14	62.44	61.32	60.00
Interval for (Max)	60.92	60.28	60.58	55.85	56.19	56.03
Median	59.60	58.45	59.15	51.90	54.95	52.80
Std. Deviation	13.73	12.72	13.10	18.08	14.86	16.32
Minimum	31.20	38.90	31.20	7.40	24.30	7.40
Maximum	98.70	87.30	98.70	97.30	86.70	97.30

Table 5 Descriptive statistics of ability level by anxiety by QL

	English			Russian		
	Group 1	Group 2	Total	Group 1	Group 2	Total
Mean	43.02	38.13	40.42	44.60	44.16	44.36
N	37.86	32.77	36.74	39.26	39.22	40.84
95% Confidence Mean (Min)	48.17	43.49	44.11	49.94	49.10	47.89
Interval for (Max)	47.80	44.70	47.45	49.25	48.40	49.25
Median	13.81	15.37	14.75	14.30	14.16	14.11
Std. Deviation	0.00	1.80	0.00	0.00	5.20	0.00
Minimum	60.90	60.00	60.90	63.10	68.40	68.40
Maximum	43.02	38.13	40.42	44.60	44.16	44.36

However, we did find an interaction ($F(1,62) = 8.11$, $p = .006$, $\eta^2 = .116$) between group and QL in terms of anxiety. The group that saw the English QLs first had a slightly higher level of anxiety when they subsequently encountered the Russian QL, but the group that saw the Russian QLs first reported much less anxiety when the QL switched to English. Perhaps this indicates a sense of a relief in better comprehending what was being asked of them. With both groups, though, the QL in English resulted in less anxiety ($F(1, 62) = 23.75$ $p < .001$, $\eta^2 = .28$). (see Fig. 3).

Prior to answering the open-ended questions, students were asked to respond to the following statement "I prefer having the questions in Russian." Their average on a seven-point Likert scale was 3.46 (sd = 1.18, 95%CI [3.17, 3.75]) indicating no strong preference for QL. Thirty-one students responded to the optional open-ended questions for a response rate of 50%. These responses shed some light on learners' preferences with regards to QL.

Preferring Questions in Russian Eighteen of the 31 comments were from those that preferred questions in L2. Responses of participants who preferred the questions in Russian seemed to fall into three major categories: (1) naturalness, (2) vocabulary strategies, and (3) motivation.

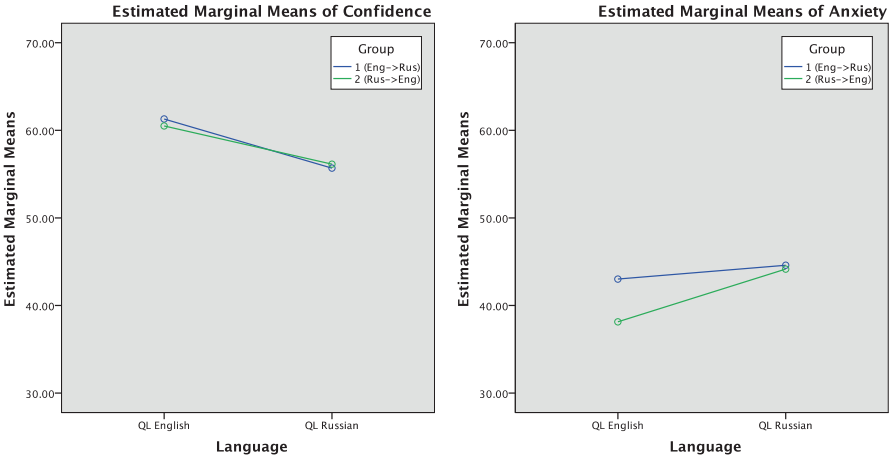


Fig. 3 Estimated marginal means of confidence and anxiety level by group and QL

Naturalness. Three of the 18 comments made reference to questions in L2 seeming more natural. One of the higher performing students who preferred questions in Russian commented that "[i]t's more natural to discuss Russian passages in Russian." The naturalness of an L2 activity is certainly a reflection of an advanced language learner's mindset. Research has shown that advanced language learners begin to think more and more in the L2, making tasks in the L2 more "natural." By contrast, beginning learners are less likely to consider activities in the L2 as natural when they are grappling with all of the newness of learning a language. This comment echoes Upton's (1997) and Upton and Thompson's (2001) findings indicating that students rely less on L1 in reading comprehension as they progress toward advanced levels. Interestingly, this student who appreciates the naturalness of the Russian questions scored the exact same in both languages. Two other comments echoed this sentiment: "Russian answers seem a bit easier since the text was in Russian" and "it was easier for me to just answer in Russian, largely because I just had to think in one language instead of switching between two."

The act of switching, perhaps, relates to the perceived naturalness of the test instrument. In addition to being less natural, switching between languages may be cumbersome, potentially violating the terms of a good test question as defined by Shohamy. A good test, she says, is one where "the method has very little effect on the trait," and a bad test is one where the method "has a strong effect on the trait being measured and consequently on the test takers' scores on such tests" (p. 147). If code-switching is perceived to place additional strain on test takers, L1 questions might invalidate the assessment instrument. Curiously, this student who did not like "switching" back and forth actually scored higher when questions were in English.

Strategies involving vocabulary. Several strategies surfaced as explanation for preferring questions in L2. One student noted that "[Russian questions] helped with words that I wasn't sure about in the text although it took me longer to figure it [the

answer] out." This statement suggests that multiple choice questions in the L2 facilitated use of problem-solving strategies. Gordon and Hanauer (1995) similarly suggested that students might learn from multiple choice questions, since they offer the most information to the test taker and serve as a rich source of information in helping the test taker answer questions. Nevo (1989) noted a similar pattern, finding that L2 questions led to more guessing or matching of similar words and phrases among a group of intermediate students of French. However, intuition suggests that matching may be more of a concern with Novice- and Intermediate-level learners, whose comprehension skills are not fully developed. Though one student in this study reported "matching" as a strategy, noting that "it is easier to match vocabulary from the question to the passage than to make good guesses using the English words," the student in question scored lower on the assessment with questions in the L2 than in the L1. If question items are well-designed, "matching" may become less viable as a test taking strategy.

Motivation. Some students who preferred the questions in Russian cited motivation as a factor in their choice. As one student explained, "if it's a Russian exam, I would like to answer the questions in Russian. It gives me an incentive to want to learn more vocabulary if I am going to take an exam as a Russian would." This statement implies that question items in the L2 may enjoy greater face validity among Advanced-level speakers than questions in the L1. Moreover, the assessment instrument itself apparently served as a purpose-driven and meaningful experience for this language learner, further motivating the student to improve vocabulary and reading skills.

Preferring Questions in English Though students on the whole scored lower when questions were posed in L2, the prior section illustrates that questions in the L2 posed some benefits to the students in the form of naturalness, strategies, and motivation. However, some students preferred questions in their L1. In analyzing their responses, two major themes emerged—strategies involving vocabulary and general difficulty of the question.

Strategies involving vocabulary. In regards to the vocabulary of the questions one student responded, "I mean, I know if I had a wider vocabulary, it (Russian questions) would not be a problem, so up to a point yes (I would prefer questions in Russian)." Another student stated that "at…times the English was good if I didn't know an important word." This statement supports Shohamy's (1984) and Bernhardt's (2005) concerns about testing L2 reading in the context of learners' "impoverished second language skills." (p 141). In fact, Shohamy (1984) asserts that for novice and intermediate learners, "presenting the questions in L1 may be considered more ethical, since the decision maker obtains information on the test taker's ability to understand the L2 text, without a carry-over from the language of the questions" (p. 158).

Whereas learners' inability to comprehend the question items in the L2 may have negatively affected their reading test scores, L1 question items may have offered test takers clues as to the general meaning of the text, as Shohamy (1984) posits. One student reported, "if there is a word you don't know in the passage, an English ques-

tion could help you figure it out." This strategy appears similar to the L2 vocabulary matching, cited above. However, the strategies used to puzzle out new vocabulary may have been quite different when the questions were posed in the L1. Matching of L2 words may involve less comprehension than matching of L1 words to L2 words. Moreover, Godev et al. (2002) found that cognates, quasi-cognates, false cognates, and quasi-false cognates can either aid or lead the test taker astray when questions are in the L1. Overall, the student responses indicate that carefully constructed test items are important to ensure that questions do not provide too much information that can lead to strategic guessing.

General difficulty of the questions. Some comments spoke to the difficulty of the questions: "It was harder in Russian," and "honestly, the questions in Russian were easy, but the answers in Russian were difficult." The latter comment may initially seem like commentary on the construction of the multiple choice items from the exam or even multiple choice items in general. However, this seems unlikely in light of the fact that no such comment was made in reference to the questions in English which had been formerly arbitrated by experts and rated at comparable difficulty to the questions in Russian. This leaves aspects that are specific to QL as a basis for establishing personal preference.

4 Conclusion

This research study investigated the effect QL has on reading comprehension test scores among advanced learners of Russian as well as learners' preferences regarding QL. Our findings corroborate previous research indicating that questions in the L1 are easier for students. Shohamy (1984) and Bernhardt (2005) hypothesized that questions in the L2 are appropriate at the advanced levels of reading proficiency, however implicit in this assumption is that the passage, question, and examinee proficiency levels are aligned. Our data suggest that L1 questions are easier even for advanced-level learners, when responding to texts and questions that may be beyond their actual proficiency level. It may be that the QL has less of an impact when the learners' reading proficiency matches that of the intended passage and question difficulty.

We also examined students' preferences for QL in the L1 or L2. In this study, unlike in Filipi's (2012) study with lower-level learners, participants reported no strong preference for QL, in spite of the fact that questions in the L2 proved more difficult. This ambivalence towards QL suggests that questions in the L2 may enjoy greater face validity than questions in the L1 for advanced-level speakers. Face validity is defined as an individual's subjective view of the validity of an assessment, or in other words, the test taker's belief about whether or not the assessment is a fair measure of knowledge or ability (Holden, 2010). The students' lack of a strong preference may indicate their beliefs that, as advanced speakers, they *should* be able to handle questions in the L2. In fact, at least one student indicated that questions in the L2 appeared more authentic and therefore more motivating.

Even though students expressed no strong preference for QL, their confidence levels in answering correctly were generally higher when responding to questions in English than in Russian. However, that confidence was frequently misplaced with students being generally overconfident in their abilities, but tending to be more so when the QL was English. Even for Advanced-level learners, questions in the L1 may serve as a "security blanket," making them overconfident in their comprehension.

These findings are intriguing in light of Shohamy's study, suggesting that more advanced readers were "hardly affected" (p. 157) by the QL, leading Shohamy to conclude that "high and low-level" students may process L2 data differently. Drawing on Corder (1978), she suggested that learners rely more on the L1 in the beginning phases of language learning. In fact, she posits that "[i]n the beginning phases, the native language is the only linguistic system from which the learner can draw" (p. 158). Nevertheless, even advanced learners appear to feel more confident, even if over confident, in their comprehension when questions are posed in the L1.

At least one of the comments in our qualitative study discussed the difficulty associated with switching between the L1 and the L2. If, as Shohamy (1984) and Corder (1978) posit, advanced level learners draw primarily from the L2 linguistic system, switching between languages may cause psychological strain. Even if this strain does not adversely affect performance on an assessment instrument, learners may believe that code switching does. Such a belief would challenge the face validity of the instrument.

In developing criterion-referenced tests, it is important to consider QL and cut scores. Advanced- and Superior-level readers are expected to have a much broader base of vocabulary, allowing them to more easily comprehend questions in the L2. If test designers insist on presenting questions in the L1, cut scores may need to be higher to certify Advanced- and Superior-level proficiency, since it appears that questions in the L1 are easier even for advanced-level readers.

4.1 Limitations and Suggestions for Future Research

The primary limitations of this study involve a possible mismatch between the learners' reading proficiency levels and the test items. The research instrument predominately comprised Superior-level items, whereas the learners may have been at the Advanced level, or possibly below. Participants were invited to take part in the study based on their enrollment in an advanced-level course in Russian cultural history and their extended time abroad in Russian-speaking countries. In the end, however, the overall scores on the reading comprehension exam were unexpectedly low, suggesting that the multiple choice items may have been above most students' proficiency level (overall mean = 49.35%). Since the test items had been empirically validated for the Superior level, the learners' performance may be evidence that they were not actually Superior-level readers.

To better understand the effect of QL on Advanced- and Superior-level readers, researchers should first establish the proficiency level of the learners, using a criterion-referenced test and then match the instrument to the learners' level. The study would have benefitted from dividing participants into groups of high and low ability based on prior proficiency measures in order to better interpret the results. In cases where some students preferred L1 and some students preferred the L2, it would have been illuminating to know find out if there was a correlation between previously determined proficiency level and scores on the reading comprehension exam and preference for either QL. Future research could also investigate the effect of QL for items below, at, and above the learners' established proficiency levels.

Additionally, future research that investigates question-related variables, such as vocabulary, strategies, motivation, and naturalness, would contribute to an understanding learner attitudes toward QL. These areas could be examined in terms of preference, difficulty, and validity. The present study only addressed QL preference and found that preference and difficulty regarding QL do not necessarily correlate. More research is needed to understand what factors contribute to the "difficulty" of a passage and item.

Student comments about the difficulty of items in the L1 or L2 only hinted at their processing and test-taking strategies. Multiple choice questions on reading comprehension exams contain a great deal of information and thus may help the test taker to answer questions correctly (Godev et al., 2002). The information contained in the questions in the present study invited participants to implement the strategy of comparing the question information with the passage information in order to learn more information. From the survey comments it is apparent that the strategy of using questions in L1 as well as L2 to learn information was employed at least to some degree. Whether that information actually helped in answering the questions correctly was undetermined.

In the present study, we considered students' preferences for QL as well as their confidence in responding to questions posed in the L1 or the L2. However, Shohamy (1984) has hypothesized that L2 questions may cause anxiety, particularly for low-level learners. What impact anxiety might have for learners of any level remains to be determined. Other affective variables, such as confidence and self-efficacy, along with their relationship with QL may also prove to be useful areas of research.

More research is needed in this area with larger sample sizes and with a wider variety of L2 s. The QL research to date has focused on French, Spanish, and English, and while Shohamy examined QLs in Hebrew with English passages, this is the first known study to examine a less commonly taught language with a different orthographic system. As FL research suggests, each language has a unique interaction and relationship with the L1, and the effect of different writing systems such as Arabic, Chinese, Japanese, Korean, etc. should be explored. Moreover, the nature of QL research may be such that outcomes among other language pairs may lead to substantially different results than those previously found, and the interpretation of QL research should be considered in this light.

4.2 Implications for Testing and Teaching

What language should be used when testing L2 reading comprehension? Unfortunately, the answer is not entirely clear-cut and is likely dependent on the testing situation and population as well as on practical considerations. Though it may appear that the QL should be in the L2 for Advanced- and Superior-level items, there are situations in which the L1 may be preferable. For example, certain professions, particularly in government work, require a high degree of bilingual fluency. Many language professionals are required to read in the L2 and report on that reading in the L1. In such cases, requiring learners to switch languages during an assessment instrument is a valid means of assessing their ability to perform their jobs. Additionally, in the design of reading assessments for less commonly taught languages, finding testing experts with enough expertise to ensure the quality of test items may be difficult, if not impossible. In such situations, passages can be translated into a common L1, and the questions can be evaluated in that L1. On the other hand, heritage speakers of a language or students who do not speak the L1 of the dominant population may be at a disadvantage when asked to respond to questions in a third language. In order to make reading proficiency tests available to anyone regardless of L1, the instruments cannot be dependent on bilingualism.

In the end, the issue of QL remains unsolved. However this study does yield some implications for testing. First, it is important to establish the reading proficiency of learners first with criterion-referenced testing before testing items in L1 and L2. Doing so will eliminate the question of whether the test items might be too difficult for learners. Second, test designers should construct items with test tasking strategies in mind. Well-constructed items to avoid matching or giving away information and are designed with test-taking strategies in mind. And third, the type of task required to answer a test item may have an effect on outcomes. Multiple choice may be useful for large scale norm- or criterion-referenced tests, but may not always be appropriate for classroom assessment.

This study represents a first attempt to investigate the effect of QL on the scores of Advanced-level readers of Russian. Although we were unable to definitively answer the question about which language should be used in assessing reading comprehension at the Advanced and Superior levels, this study has nonetheless contributed to our understanding in this area. Advanced-level readers of Russian generally reported that questions in the L2 were more difficult to answer leading to increased anxiety and decreased confidence than were questions in the L1. However, our study finds that the level of the learners may not be as important as the alignment of the learners' proficiency level and the difficulty of the reading passages and subsequent tasks. Decisions about QL should be made deliberately, taking into consideration the level of the participants and the level of the tasks that they are expected to perform.

References

ACTFL. (2012). *ACTFL proficiency guidelines 2012*. Alexandria, VA: American Council on the Teaching of Foreign Languages. Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal, 75*(4), 460–472. https://doi.org/10.1111/j.1540-4781.1991.tb05384.x

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25*, 133–150. https://doi.org/10.1017/s0267190505000073

Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension in Spanish. *The Modern Language Journal, 89*(1), 37–53. https://doi.org/10.1111/j.0026-7902.2005.00264.x

Brantmeier, C. (2006). The effect of language of assessment and L2 reading performance on advanced readers' recall. *The Reading Matrix, 6*(1), 1–17. Retrieved from: https://pages.wustl.edu/files/pages/imce/brantmeierlanguageresearch/effects_of_language_of_assesment_0.pdf

Bügel, K., & Buunk, B. P. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *The Modern Language Journal, 80*(1), 15–31. https://doi.org/10.2307/329055

Clifford, R. (2016). A rationale for criterion-referenced proficiency testing. *Foreign Language Annals, 49*(2), 224–234. https://doi.org/10.1111/flan.12201

Clifford, R., & Cox, T. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals, 46*(1), 45–61. https://doi.org/10.1111/flan.12033

Corder, S. P. (1978). Language learner language. In J. Richards (Ed.), *Understanding second and foreign language learning: Issues and approaches* (pp. 71–93). Rowley, MA: Newbury House.

Filipi, A. (2012). Do questions written in the target language make foreign langauge listening comprehension tests more difficult? *Language Testing, 29*(4), 511–532. https://doi.org/10.1177/0265532212441329

Godev, C. B., Martinez-Gibson, E. A., & Toris, C. C. (2002). Foreign language reading comprehension test: L1 versus L2 in open-ended questions. *Foreign Language Annals, 35*(2), 202–221. https://doi.org/10.1177/026553229601300205

Gordon, C. M., & Hanauer, D. (1995). The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly, 29*(2), 299. https://doi.org/10.2307/3587626

Harrison, C., & Dolan, T. (1979). Reading comprehension: A psychological viewpoint. In R. Mackay, B. Barkman, & R. R. Jordan (Eds.), *Reading in a second language* (pp. 13–23). Rowley, MA: Newbury House.

Holden, R. B. (2010). Face validity. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini encyclopedia of psychology* (4th ed., pp. 637–638). Hoboken, NJ: Wiley.

Lee, J. F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition, 8*(2), 201. https://doi.org/10.1017/s0272263100006082

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology, 94*(4), 778–784. https://doi.org/10.1037/0022-0663.94.4.778

Lorch, R. F., Lorch, E. P., & Klusewitz, M. A. (1993). College students' conditional knowledge about reading. *Journal of Education al Psychology, 91*, 239–252. https://doi.org/10.1037/0022-0663.85.2.239

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*(2), 199–215. https://doi.org/10.1177/026553228900600206

Poh, T., & Hock, L. (1979). The performance of a group of Malay-medium students in an English reading comprehension test. *RELC Journal, 10*(1), 81–89. https://doi.org/10.1177/003368279010001 06

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test perofmrance. *Language Testing, 24*(1), 99–128. https://doi.org/10.1177/0265532207071513

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comrehension. *Language Testing, 1*(2), 147–170. https://doi.org/10.1177/026553228400100203

Upton, T. (1997). First and second language use in reading comprehension strategies of Japanese ESL students. *Teaching English as a Second or Foreign Language – The Electronic Journal for English as a Second Language, 3*(1), 1–23. Retrieved from: http://www.tesl-ej.org/wordpress/issues/volume3/ej09/ej09a3/

Upton, T., & Thompson, L. (2001). The role of the first language in second language reading. *Studies in Second Language Acquisition, 23*, 469–495.

Troy L. Cox, PhD, is a faculty member in the Linguistics Department at Brigham Young University and serves as the Associate Director of the Center for Language Studies. He is also a certified American Council on the Teaching of Foreign Languages (ACTFL) oral proficiency trainer and has used his testing expertise as a forensic linguist and in test development projects in a number of different languages. His research interests include proficiency testing, the integration of technology with assessment, objective measurement and self-assessment.

Jennifer Bown is Associate Professor of Russian at Brigham Young University, where she serves as graduate coordinator for the Second Language Teaching program. Her research interests include literacy development in foreign languages, as well as affective issues related to language learning.

Teresa R. Bell is Associate Professor of German and Second Language Acquisition at Brigham Young University, USA. Her research focuses on effective language teaching and learning. She serves as the American Council on the Teaching of Foreign Languages (ACTFL) Program Review Coordinator for the Council for the Accreditation of Educator Preparation (CAEP).

Proficiency vs. Performance: What Do the Tests Show?

Fernando Rubio and Jane F. Hacking

This research was supported in part by a grant from The Language Flagship.

Abstract Research has shown consistently that after two semesters of instruction, students in post-secondary institutions show only Novice levels of proficiency as measured by the ACTFL scale. Even after four semesters, proficiency does not always reach the Intermediate level, especially in listening. These findings are troubling both for students and for practitioners. Although pedagogical or curricular weaknesses could explain these results, this chapter explores an alternative explanation that revolves around the nature of the tests used. We argue that the nature of the existing proficiency tests makes them inadequate for Novice learners since they measure a type of linguistic competence that is inconsistent with what language learners at the lower levels are able to do. We also argue that the lackluster results observed in listening may be due to a problem of test validity. The existing tests of listening proficiency may not be the right tools to measure the multi-modal processes involved in real-life listening comprehension.

Keywords Assessment · Validity · Task-based · Testing · Proficiency · Performance · Language

The American Council on the Teaching of Foreign Languages (ACTFL) published its first proficiency guidelines in 1986, with updated versions published in 1999, 2001 and 2012. ACTFL defines the guidelines as "descriptions of what individuals

F. Rubio (✉)
Second Language Teaching and Research Center, University of Utah, Salt Lake City, UT, USA
e-mail: Fernando.Rubio@utah.edu

J. F. Hacking
World Languages and Cultures, University of Utah, Salt Lake City, UT, USA
e-mail: j.hacking@utah.edu

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_8

can do with language in terms of speaking, writing, listening, and reading in real-world situations in a spontaneous and non-rehearsed context" (ACTFL, 2012a, p.2). They were developed based on the experience of governmental agencies with oral assessment and following the descriptors of language proficiency used by the Interagency Language Roundtable (ILR). The guidelines are designed to be used in the evaluation of functional language ability and describe a range of proficiency that goes from that of an educated native speaker to a level of no functional ability. Although they neither describe how languages are learned, nor prescribe how they should be taught, for more than 30 years since their publication, the Guidelines have progressively spread through the language teaching profession in the United States to become the main measure of the success of a language program. Many programs require proof of proficiency at a certain level, typically by means of an ACTFL test, in order to meet a graduation requirement or earn an academic certificate. Numerous post-secondary institutions gauge the success of their language programs based on students' level of proficiency measured according to the ACTFL guidelines.

Proficiency is defined by ACTFL as "the ability to use language in real world situations in a spontaneous interaction and nonrehearsed context and in a manner acceptable and appropriate to native speakers of the language" (ACTFL, 2012b, p.4). This is in contrast to the definition of performance, which is "the ability to use language that has been learned and practiced in an instructional setting" and is used "within familiar contexts and content areas" (ACTFL, 2012b, p.4). Although ACTFL published a parallel set of Performance Guidelines for K-12 in 1998, followed by an updated version (labeled Performance Descriptors) for K-16 in 2012, the notion of performance has primarily remained a K-12 concept that has received very little attention in post-secondary education.

ACTFL explains the difference between performance and proficiency as a factor of the context in which a certain function is performed and the degree of control that the learner exhibits over the function. For example, a student who has been practicing mock job interviews in a language class, may evidence the ability to ask and answer some basic job-related questions. This learner would then show performance at the Intermediate level by virtue of the ability to perform one or more Intermediate-level functions in a particular situation that has been previously rehearsed. That, however, does not guarantee that this learner would be able to perform the same functions in a different context (e.g., ask and answer questions in a health-related conversation with a doctor). As ACTFL puts it, "in an instructional environment, the content and tasks are controlled, resulting in higher expectations of learners' performance compared to how they perform in a non-instructional environment" (ACTFL, 2012b, p. 3). The assumption is that sustained performance at a certain level "points to" proficiency at that level. So, a student that is able to perform the functions of the Intermediate level over a wide variety of previously practiced contexts, is likely to be able to show Intermediate-level proficiency in an unrehearsed situation. Unlike the proficiency guidelines, which are designed to measure global functional ability, the performance descriptors illustrate what a learner is able to do with respect to a particular curriculum that has been taught and learned. In sum, both performance and proficiency describe linguistic behavior in language-use contexts; the difference is that proficiency refers to unrehearsed behavior in unpredictable situations, while performance refers to rehearsed behavior in controlled contexts.

This distinction between performance and proficiency is reflected in the testing instruments developed by ACTFL. There are ACTFL *proficiency* tests for speaking, writing, reading and listening, all developed around the proficiency guidelines. And there is a separate *performance* test—the ACTFL Assessment of Performance towards Proficiency in Languages (AAPPL)—that was developed with a K-12 focus and is based on the performance descriptors. AAPPL measures language learning based on the World-Readiness Standards for Language Learning. It assesses Interpersonal Listening/Speaking, Presentational Writing, Interpretive Reading, and Interpretive Listening.

According to ACTFL's description of performance and proficiency, through extensive practice learners progress along a continuum that goes from showing control of language features and functions under only very predictable conditions, to being able to perform those functions and exhibit those features in a sustained way regardless of content or context. There is, therefore, a connection, but also a clear difference between performance and proficiency. However, when one looks at the guidelines that describe the lower levels of proficiency in the ACTFL scale, one finds them much closer to the definition of performance than to proficiency. Table 1 (ACTFL, 2012a) shows the descriptions of proficiency at the Novice Mid sublevel, which is the level at which a learner exhibits the most prototypical Novice profile. Table 2 includes the performance descriptors for the Novice range. We have bolded the terms that are typically used to refer to performance, rather than proficiency.

It is evident from reading the descriptors in Table 1 and comparing them with Table 2 that learners at the Novice level of proficiency only have the ability to use the language in rehearsed, highly predictable situations and in essence, therefore, they can only show performance, rather than proficiency. In this chapter, we explore the consequences that this apparent overlap has for testing and curriculum.

Table 1 Proficiency descriptors for Novice Mid sublevel (ACTFL, 2012b)

Speaking	Speakers at the Novice Mid sublevel communicate minimally by using a number of isolated words and **memorized phrases** limited by the particular **context in which the language has been learned**. […] they may say only two or three words at a time or give an occasional **stock answer**. They pause frequently as they search for simple vocabulary or attempt to **recycle** their own and their interlocutor's words.
Writing	Writers at the Novice Mid sublevel can **reproduce from memory** a modest number of words and phrases in context. They can supply limited information on simple forms and documents, and other **basic biographical information**, such as names, numbers, and nationality. Novice Mid writers exhibit a high degree of accuracy when writing on well-practiced, **familiar topics** using limited **formulaic language**. With less familiar topics, there is a marked decrease in accuracy. […] There is little evidence of functional writing skills.
Listening	At the Novice Mid sublevel, listeners can recognize and begin to understand a number of **high-frequency, highly contextualized** words and phrases including aural **cognates** and borrowed words. Typically, they understand little more than one phrase at a time, and repetition may be required.
Reading	At the Novice Mid sublevel, readers […] can identify a number of **highly contextualized** words and phrases including **cognates** and borrowed words but rarely understand material that exceeds a single phrase.

Table 2 Performance descriptors for Novice level (ACTFL, 2012b)

Interpretive	Interpersonal	Presentational
Understands words, phrases, and **formulaic language** that have been **practiced** and **memorized** to get meaning of the same idea from simple, **highly-predictable** oral or written texts, with strong visual support.	Expresses self in conversations on very **familiar topics** using a variety of words, phrases, simple sentences and questions that have been **memorized**.	Communicates information on very **familiar topics** using a variety of words, phrases, and sentences that have been **memorized**.

1 Proficiency Level and Length of Study

The Foreign Service Institute (FSI) of the Department of State classifies languages based on their presumed level of difficulty for native English speakers.[1] According to this classification, there are three categories of languages based on the length of time that it takes a native speaker of English to reach a certain level of proficiency (Malone & Montee, 2010). Category I includes the Romance languages and others such as Dutch or Norwegian that require a comparable amount of time for English learners to master. Languages in Category II require approximately twice the amount of time to reach professional competence. This category includes Russian, Vietnamese, Turkish and Greek among others. Category III includes Arabic, Chinese, Japanese and Korean, which require about three times as much as the Category I languages to achieve professional competence. According to Liskin-Gasparro (1982), an English speaker needs a minimum of 240 h of instruction to reach the Intermediate level of proficiency in Category I languages and at least 480 h in languages that are more typologically distant from English. In the United States, the number of contact hours in introductory-level language courses varies from institution to institution, typically ranging from 3 to 5 contact hours per week. That means that, assuming a typical 30-week academic year, a student would be exposed to between 90 and 150 h of instruction in the language after one year and 180–300 after two years of instruction. This implies that the majority of the students enrolled in language courses at the post-secondary level in the United States are likely to still be in the Novice range of proficiency after one year and in some cases even after two years of instruction.

This scenario is confirmed by the results of a number of studies conducted over the past decade to measure the level of language proficiency of undergraduates in the United States using the ILR/ACTFL proficiency scale. Rifkin (2005) measured the level of proficiency in speaking, listening, reading and writing of undergraduate students of Russian who were enrolled in the summer immersion program of the Middlebury Russian school. A total of 352 students were assessed using the ACTFL Oral Proficiency Interview (OPI) and tests of listening, reading and writing that were designed based on the ACTFL guidelines. Students who had previous exposure

[1] Although the FSI language difficulty scale is often cited, it has never been empirically validated.

to Russian were given pre-immersion tests and all students were also tested at the end of the immersion program, which consisted of 140 h of instruction. The results of the pre-immersion tests show that students who had an average of 150 h of previous instruction in Russian had ratings of Novice High in all four skills. Those who had received 250 h of previous instruction were at the bottom of the Intermediate Low range in speaking and writing and still Novice Low in reading and listening. Students showed significant gains after the immersion experience and those gains were more evident in the receptive skills. Rifkin also compared the effects on proficiency of the two instructional models (regular classroom instruction vs. immersion). The results of his study indicate that the positive effect of the additional 140 h of immersion instruction is larger than would be predicted for 140 hours of non-immersion classroom instruction.

Watson & Wolfel (2015) analyzed the proficiency of 279 students participating in a semester abroad program. A prerequisite for participation in the program was completion of a minimum of 2 years of college foreign language courses or their equivalent. Students had to take three language proficiency tests: reading, listening and speaking. Reading and listening were assessed using the Defense Language Proficiency Test (DLPT), a computer-based proficiency test based on the ILR proficiency scale. Speaking proficiency was measured using the OPI. Learners represented seven languages that the authors divided into two groups according to difficulty. French, German, Portuguese and Spanish formed the "less difficult" category. The "more difficult" group was comprised of Arabic, Chinese and Russian. The results of the pre-study abroad tests showed that the majority of the students in the more difficult languages were still at the Novice level after 2 years of study (86% in listening, 88% in reading and 59% in speaking). In the less difficult languages, the results were considerably better. The percentage of students still at the Novice level after 2 years of instruction were as follows: 14% in listening, 8% in reading and speaking. Although the level of proficiency of the second group seems much higher than that reported in other similar studies and significantly better than that of the more difficult group, we do not know how many of those students had completed more than the required minimum of 2 years of previous instruction.

Tschirner (2016a) provides the most comprehensive overview of listening and reading proficiency of college level students across a variety of languages. For his study, Tschirner administered ACTFL RPTs and LPTs to more than 3000 students of French, German, Italian, Japanese, Portuguese, Russian and Spanish at 21 institutions of higher education in the United States. His goal was to determine the level of proficiency in those two skills at major milestones in the students' course of study, and also to look at the relationship between level of proficiency in the two skills. The results indicate that learners are able to reach advanced levels of proficiency in reading by the time of graduation, but not necessarily in listening. Of more interest for our purposes are his findings regarding levels of proficiency attained after 2 and 4 semesters. Tschirner found that, after 2 semesters, students were typically in the Novice range in both skills regardless of the language. The results after 4 semesters showed that students were reaching the Intermediate range

in reading only in the cognate languages, and that the average level of proficiency in listening was still in the Novice range for all languages (except Italian, which had a very small *n*).

2 Findings from the Flagship Proficiency Initiative

Similar results to those described in the previous section have been obtained as part of a large-scale assessment project funded by the Language Flagship. Under the auspices of the Flagship Proficiency Initiative, Michigan State University, the University of Minnesota and the University of Utah have documented levels of proficiency in speaking, reading and listening of several thousand undergraduate students enrolled in language courses at all levels from 1st- to 4th-year in Arabic, Chinese, French, German, Korean, Portuguese Russian, and Spanish. In this chapter, we report the data for students enrolled in second- and fourth-semester courses in Chinese, French, Russian and Spanish between the fall semester of 2014 and the spring semester of 2016 at all three institutions. We chose these languages because they provide robust enough samples and because they represent a range of levels of difficulty for native English speakers (Spanish and French are Category I languages, Russian is a Category II and Chinese is a Category III). The students enrolled in these courses were administered ACTFL proficiency tests of speaking, listening and reading after completing each semester of instruction. Speaking proficiency was measured using the Oral Proficiency Test by Computer (OPIc), which is a computer-delivered version of the ACTFL Oral Proficiency Interview (OPI). Reading and Listening proficiency were measured by means of the Reading Proficiency Test (RPT) and the Listening Proficiency Test (LPT) respectively (ACTFL, 2013, 2014); both are delivered by computer via the internet. All three tests are constructed based on the ACTFL Proficiency Guidelines 2012.

The OPIc replicates the structure of the OPI and uses a series of interactive and adaptive tasks to elicit a ratable sample of speech (ACTFL, 2012c). Test takers first complete a background survey and self-assessment. The test taker's answers to the background survey determine the pool of topics from which the computer will select the questions that will be generated. The self-assessment presents the test takers with six different descriptions of levels of proficiency and asks them to select the one that most accurately matches their level. Based on this response, the computer selects one of four possible forms of the OPIc (Form 1, Form 2, Form 3, or Form 4). Each form targets a range of levels from Novice Low to Superior. The OPIc is rated by certified OPIc raters.

The RPT and LPT are standardized tests for the global assessment of reading and listening ability in a language. They were developed and validated by the Institute for Test Research and Development at the University of Leipzig. Before taking the test, examinees (or their institution) determine what levels will be tested. Both tests have a number of different forms, each capable of assessing a range of levels from Novice through Superior. The reading or listening tasks can be at any of five sublev-

els: Intermediate Low, Intermediate Mid, Advanced Low, Advanced Mid and Superior. Each sublevel consists of five reading texts or listening passages accompanied by three tasks with four multiple-choice responses. Depending on the form of the test selected, an examinee will receive between 10 and 25 listening or reading passages. The appropriateness of the content area, length, organization, vocabulary, or purpose of the passages was determined in accordance with the respective descriptors in the ACTFL scale. Tasks vary from level to level. At the lower levels the tasks typically include global, detailed and selective questions, while at the higher levels they include global, detailed and inference questions. The complexity of the task is also aligned to the level of the passage. For example, a detailed or global question at the Intermediate-level can be answered by understanding single sentences, while the same type of question at the Advanced level requires understanding of complete paragraphs (Institute for Test Research and Test Development, (2013a, 2013b). Both the RPT and the LPT are machine-scored tests.

Table 3 shows the number of tests administered by skill and by year across the three institutions.

The data obtained from testing students after two and four semesters of instruction are summarized in Tables 4, 5, 6, 7, 8, 9, 10, and 11. The results were converted from ACTFL scores to an ordinal scale following the same conversion scale used in previous studies (e.g., Rifkin, 2005; Tschirner, 2016a), from Novice Low 1, to Superior 10. The unusually high maximum values found in some cases (up to 8 or 9) are due to outliers who were incorrectly placed in introductory-level courses. The results of the testing of 2nd-semester students are summarized in Tables 4, 5, 6, and 7. Both means and median scores for all languages in all three skills indicate that students at this level are consistently below the Intermediate range of proficiency. Similar to the findings of other studies, listening is the weakest skill in all cases. Not surprisingly, reading levels are significantly lower than speaking in the languages that do not use the Roman alphabet, but reading is higher than speaking in French and Spanish.

After four semesters of instruction (Tables 8, 9, 10, and 11), speaking levels are already in the intermediate range in French, Russian and Spanish, but not in Chinese. At this point, speaking is the strongest skill in all languages except for Spanish, where reading is slightly higher. Reading reaches the Intermediate level in the cognate languages, but it is still at the Novice level in Chinese and Russian. Listening still remains the weakest skill across languages and is still uniformly at the Novice level.

The results of the research reviewed above and these data from the Language Flagship Proficiency Initiative demonstrate that college students are not reaching the Intermediate level of proficiency after two semesters of instruction and, in many

Table 3 Number of tests administered in 2nd- and 4th-semester courses in Chinese, French, Russian and Spanish

	Semester 2	Semester 4
OPIc	724	886
RPT	726	1574
LPT	703	830
Total	**2153**	**3290**

Table 4 Chinese scores by skill—semester 2

	N	Minimum	Maximum	Mean	Std. Deviation	Median score
OPIc	55	1	9	2.75	1.377	2
RPT	49	1	5	1.49	.893	1
LPT	53	1	5	1.40	.840	1

Scores converted as follows: NL = 1, NM = 2, NH = 3, IL = 4, IM = 5, IH = 6, AL = 7, AM = 8, AH = 9, S = 10

Table 5 French scores by skill—semester 2

	N	Minimum	Maximum	Mean	Std. deviation	Median score
OPIc	241	1	5	3.06	1.107	3
RPT	243	1	5	3.07	1.229	3
LPT	220	1	5	2.48	1.199	2

Scores converted as follows: NL = 1, NM = 2, NH = 3, IL = 4, IM = 5, IH = 6, AL = 7, AM = 8, AH = 9, S = 10

Table 6 Russian scores by skill—semester 2

	N	Minimum	Maximum	Mean	Std. deviation	Median score
OPIc	86	1	8	3.28	1.214	3
RPT	89	1	5	1.94	1.300	1
LPT	86	1	5	1.80	1.166	1

Scores converted as follows: NL = 1, NM = 2, NH = 3, IL = 4, IM = 5, IH = 6, AL = 7, AM = 8, AH = 9, S = 10

Table 7 Spanish scores by skill—semester 2

	N	Minimum	Maximum	Mean	Std. deviation	Median score
OPIc	342	1	5	2.72	1.018	3
RPT	345	1	5	2.87	1.331	3
LPT	344	1	5	2.03	1.089	2

Scores converted as follows: NL = 1, NM = 2, NH = 3, IL = 4, IM = 5, IH = 6, AL = 7, AM = 8, AH = 9, S = 10

Table 8 Chinese scores by skill—semester 4

	N	Minimum	Maximum	Mean	Std. deviation	Median score
OPIc	68	1	8	3.34	1.522	3
RPT	67	1	7	2.03	1.314	2
LPT	64	1	5	1.89	1.370	1

Scores converted as follows: NL = 1, NM = 2, NH = 3, IL = 4, IM = 5, IH = 6, AL = 7, AM = 8, AH = 9, S = 10

Table 9 French scores by skill—semester 4

	N	Minimum	Maximum	Mean	Std. deviation	Median score
OPIc	284	1	8	4.15	1.145	4
RPT	260	1	7	4.08	1.408	4
LPT	255	1	7	3.41	1.334	4

Scores converted as follows: NL = 1, NM = 2, NH = 3, IL = 4, IM = 5, IH = 6, AL = 7, AM = 8, AH = 9, S = 10

Table 10 Russian scores by skill—semester 4

	N	Minimum	Maximum	Mean	Std. deviation	Median score
OPIc	99	1	8	4.36	1.281	4
RPT	94	1	7	3.32	1.453	4
LPT	97	1	7	3.05	1.439	3

Scores converted as follows: NL = 1, NM = 2, NH = 3, IL = 4, IM = 5, IH = 6, AL = 7, AM = 8, AH = 9, S = 10

Table 11 Spanish scores by skill—semester 4

	N	Minimum	Maximum	Mean	Std. deviation	Median score
OPIc	435	1	8	4.13	1.178	4
RPT	427	1	7	4.26	1.648	4
LPT	414	1	7	3.24	1.365	3

Scores converted as follows: NL = 1, NM = 2, NH = 3, IL = 4, IM = 5, IH = 6, AL = 7, AM = 8, AH = 9, S = 10

cases, not even after 4 semesters, particularly in listening. This has important curricular implications, since the majority of the students enrolled in language courses at postsecondary institutions populate first- and second-year courses, often to fulfill an institutional language requirement. According to the latest enrollments report published by the Modern Language Association (Goldberg, Looney, & Lusin, 2015), 83.3% of undergraduate language course enrollments were in introductory courses (first and second year). Thus, the results reported in this chapter are relevant for the vast majority of students in US higher education for whom the language learning experience is restricted to lower level language classes and does not result in any sort of functional proficiency. In the following sections, we attempt to answer two questions that arise from the findings of the research examined.

1. Is it appropriate to use proficiency tests with learners at the lower levels of proficiency?
2. Can the nature of the tests explain the lag in listening proficiency compared to other skills?

3 The (In)adequacy of Proficiency Tests

The basic premise of academic assessment is that the assessment will provide valid and reliable evidence of what the student can do in a non-testing situation. In this section, we try to answer our first question by examining the nature of the ACTFL tests to determine, to the extent that it is possible, whether they do achieve the goal of providing valid and reliable evidence of global language proficiency.

The ACTFL proficiency tests are a form of task-based language performance assessment (in the sense suggested by Brown, 2004) that are designed to provide evidence of proficiency.

Brown provides an excellent overview of some of the most crucial issues related to performance assessment. One of the main challenges that he points out in the development of performance assessments is how to address the complexity of the interactions between task characteristics, task conditions, and test-taker characteristics and how these interactions may affect students' performance on tasks (p. 102–122). Brown suggests using the assessment design framework of Evidence-Centered Design (ECD) proposed by Mislevy, Steinberg, and Almond (2002) as a potentially useful way to "solve the problems of complex interactions between task characteristics, task conditions, student characteristics, and so forth" (Brown, 2004, p. 115). Mislevy et al. propose a model to operationalize the components of a performance assessment so that we can first figure out the structure of the evidentiary argument (what do we want to say about students and what evidence do we need?) and then determine how to assemble the necessary elements to transform that argument into an assessment. There are four models in the ECD framework: a student model, an evidence model, a task model, and an assessment model. The student model specifies what we want to measure about students. The variable in the student model is the particular construct at the core of the assessment. In our case, the variable in the student model is proficiency. This variable has different values that are the different levels of proficiency. The task model determines how evidence will be elicited. According to Mislevy et al., a task model is "a schema for constructing and describing the situations in which examinees act" (p. 491). The link between the student model and the task model is the evidence model, which determines how achievement of a task is evaluated. Fig. 1 shows how the structure of the OPI can fit within the ECD framework as presented in Tschirner (2016b).

A detailed description of the complete ECD framework is beyond the scope of this chapter so, for the purposes of our discussion, we will focus here on how this framework may help us determine if the tests used to measure global language proficiency are valid measures when used with lower level learners.

In a task-based language test, the task model is what determines how evidence about language proficiency will be elicited. In the case of the ACTFL tests, the task model specifies the types of global tasks and functions that learners can perform at each level (describe, narrate, hypothesize, etc.), the range of content and contexts that they can handle, the text type that they are able to produce/process, etc. For example, the ability to show comprehension of a written passage that consists of

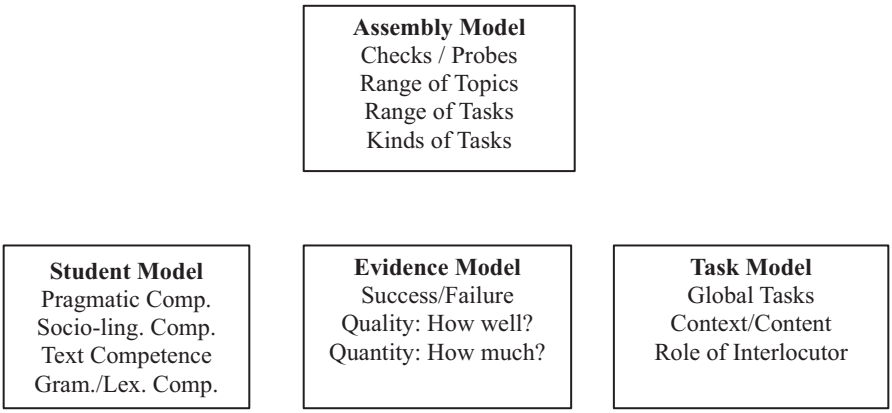


Fig. 1 The structure of the OPI in the ECD framework

simple sentences will provide information to the evidence model, which will determine whether that evidence successfully matches the construct of proficiency at the intermediate level, which is part of the student model.

We argue that the problem with the ACTFL proficiency tests when used with learners at the lower levels is that they are designed to measure global communicative competence, but the lower levels are defined as a lack of functional ability; that is, they are described as consistent with no proficiency or, at best, memorized performance. The variables of the task model (functions, text type, etc.) result in tasks that cannot elicit the type of performance that a Novice-level learner is capable of. A test of proficiency may not be the most appropriate tool to provide information about lower level learners since the only information that can be processed by the evidence model is that the student's performance on the tasks does not match the student model variable (proficiency). In the ECD framework and from a construct-centered perspective, a proficiency test does not work for lower level learners because they cannot show evidence of the student model variable—the construct—that is being measured. In essence, using a proficiency test to test a Novice learner would be akin to designing a driving test that measures your ability to drive under a variety of conditions (in heavy traffic, on a mountain road covered with snow, in the rain at night), and giving that test to someone who has only practiced driving in a straight line at low speeds on a road with no traffic.

According to Norris (2002), one of the key questions that needs to be asked before using a task-based language test is what we are going to do with the evidence that we gather, "what decisions will be made, what actions taken, what consequences sought" (p. 337). If the answer is that we want to make grading decisions at the individual level and curricular decisions at a programmatic level, the information about Novice-level examinees elicited through a proficiency test will not be very useful. An important reason why the ACTFL scale was adapted from the original ILR scale was to create additional sublevels that would reflect the reality of most

learners in academic settings, who would typically be ILR level 0 after a full year of study and often only 0+ after 2 years. ACTFL then published the Performance Descriptors to "provide more detailed and more granular information about language learners" (ACTFL, 2012b, p. 3). In conjunction with the Performance Descriptors, a performance scale was developed that divides up the Novice and Intermediate levels into additional sublevels (four for Novice and five for Intermediate). The motivation was that in the K-12 system learners would typically take several years to move through the Novice range and, therefore, a more granular scale would be better suited to show the progress being made and would provide more useful information to students and teachers. If, as research indicates, the situation is similar in the introductory language programs at the post-secondary level, a different type of test and a corresponding more granular scale would also be appropriate.

A test similar to the AAPPL measure described earlier but designed for adult learners, may be a better option for students in first- and second-year college courses, since it measures a learner's ability to perform a series of tasks with previously practiced content and context. In fact, ACTFL publishes the list of tasks and content that the AAPPL measure covers, which is an acknowledgement of the fact that the test is designed to measure practiced language-use tasks. In the ECD framework, Novice learners would be better served by a test in which the construct for the student model is simply success on specific tasks, rather than a construct of global language competence or ability.

4 The Problem with Listening

In view of the less positive results for listening proficiency, it would appear that listening ability develops at a slower pace than other skills. But is this the right conclusion to draw? Instead of concluding that there are (as yet not understood) psycholinguistic variables that may make listening more challenging, could there be additional explanations for these results? Research findings point to the type of learning context as perhaps a crucial variable in explaining the development of listening proficiency. For example, Tschirner (2016a) found that students who had spent a substantial amount of time (2 years) abroad in naturalistic, immersion settings had developed their proficiency to similarly high levels in reading and listening, unlike those without the immersion experience, for whom listening levels were significantly lower. Also, Davidson (2010) analyzing the proficiency gains of Russian learners studying abroad for periods ranging from 2 to 9 months found that only those who participated in the 9-month program were able to show significant gains in listening. And Rifkin (2005) shows that participation in an immersion program (a different context of learning from what they had previously experienced) resulted in proficiency gains for students especially in reading and listening. Taken together, the results of these studies seem to suggest that there is a relationship between what happens in the specific learning context (immersion vs. regular

classroom) and the development of listening proficiency, at least as it is measured by the ACTFL/ILR-based proficiency tests. Therefore, we attempt to answer our second question by looking at how the type of test used may be affecting the observed results.

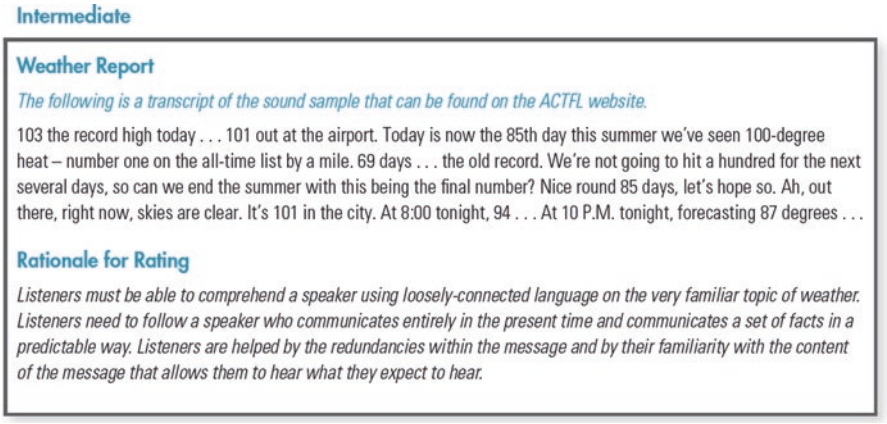
Mislevy et al. (2002) suggest that if we want an assessment to provide valid evidence of the student's abilities, we need to design it from both a task-centered and a construct-centered perspective (p. 493). In the next two sections, we discuss potential task- and construct-centered explanations for the lackluster results of listening proficiency tests.

4.1 A Task-Centered Explanation: Task Familiarity

A possible explanation for the general results of the testing described above could be that proficiency and performance at the lower levels (Novice and Intermediate) are in effect the same thing -- or rather that, in ACTFL terms, there is no proficiency at those levels, but rather only the ability to demonstrate control over features of the language that have been practiced extensively (that is, performance in the ACTFL definition). Lower-level learners demonstrate skilled performance of those tasks that they have been able to practice repeatedly. Most introductory-level language courses have as their main goal the development of oral proficiency and, consequently, dedicate significant time and attention to this skill. In contrast, a principled approach to the development of listening comprehension skills is not very common in most language classrooms and interpretive (as opposed to interpersonal) listening tasks are rare. Messick (1996) maintains that "[i]deally, the move from learning exercises to test exercises should be seamless" (p. 241), but as Tschirner (2016a) warns, "the emphasis on input and listening comprehension that characterized the early years of the communicative competence revolution in the 1970s and 1980s appears to have all but disappeared" (p. 219). Therefore, it is no surprise that student performance in listening comprehension tasks will lag behind that of reading and, particularly, speaking because of a difference in task familiarity: the tasks included in the OPIc are closer to what students do in the classroom than those included in the LPT.

4.2 A Construct-Centered Explanation: Construct Underrepresentation

As Mislevy et al. (2002) maintain, "[a] construct-centered approach helps us think through just what these performances in these situations can tell us about students, at a level above specific performances in specific situations" (p. 493). From a construct-centered perspective, a second potential validity issue that may explain the lower results in the listening tests has to do with the notion of construct under-representation. A valid task-based assessment should be made up of "engaging and worthy tasks


Intermediate

Weather Report

The following is a transcript of the sound sample that can be found on the ACTFL website.

103 the record high today . . . 101 out at the airport. Today is now the 85th day this summer we've seen 100-degree heat – number one on the all-time list by a mile. 69 days . . . the old record. We're not going to hit a hundred for the next several days, so can we end the summer with this being the final number? Nice round 85 days, let's hope so. Ah, out there, right now, skies are clear. It's 101 in the city. At 8:00 tonight, 94 . . . At 10 P.M. tonight, forecasting 87 degrees . . .

Rationale for Rating

Listeners must be able to comprehend a speaker using loosely-connected language on the very familiar topic of weather. Listeners need to follow a speaker who communicates entirely in the present time and communicates a set of facts in a predictable way. Listeners are helped by the redundancies within the message and by their familiarity with the content of the message that allows them to hear what they expect to hear.

Fig. 2 Example of Listening Passage. Reprinted from *ACTFL listening proficiency test (LPT). Familiarization manual and ACTFL proficiency guidelines 2012—listening*, by ACTFL, 2014

(usually involving multiple processes) in realistic settings or close simulations so that the tasks and processes, as well as available time and resources, parallel those in the real world" (Messick, 1996, p. 243). If a test is not assessing all of the construct and there are aspects of the construct that the test misses, we have a problem of construct under-representation. In real-life language-use situations, we rarely engage in interpretive listening that is devoid of any other contextual support; typically, when we listen, we have visual support. Therefore, our ability to infer meaning from an aural source in the real world is affected by the availability of other sources of information. For a twenty-first century learner, interpretive communication in the real world is a multi-modal process. From that perspective, the tasks used in the assessment of listening proficiency are not instances of language use that reflect all the processes involved in real-world language use. For example, Fig. 2 shows the transcript of an Intermediate-level sample listening passage from the LPT. This example is included in the LPT Familiarization Manual (ACTFL, 2014). Although the included rationale provides adequate justification for considering this an appropriate task to evaluate Intermediate-level listening skills, it is likely that in a real-life situation many listeners would require visual information to be able to process accurately the information included in such a short passage. This would be all the more true for learners who are actually still in the Novice range of proficiency.

5 Conclusions

Assessment of learning has been one of the central concerns facing higher education in recent years. There have been repeated demands by all stakeholders for colleges and universities to articulate clear learning objectives for curricula and offer concrete measures by which to assess learning. For example, The New Media

Consortium, which annually convenes a panel of experts in education to discuss the five-year horizon for the impact of technology in post-secondary education, identified a growing focus on measuring learning as one of the key short-term trends in its last report (Johnson et al., 2016). Whether or not we believe that the increased demand for external assessments of student learning is valid, is beyond the scope of this chapter. What is clear, is that there is increasing pressure to provide such evidence. Assessing students' language proficiency using standardized, nationally recognized tests is one way language departments can respond to the demand for accountability. And indeed, many programs have adopted the use of ACTFL tests for precisely this reason.

The increase in the use of third-party tests in language programs makes it all the more important to consider their efficacy, particularly if they are used at the lower levels of language instruction, e.g., at the end of a language requirement. One goal of the Flagship Proficiency Initiative grant which funded this research, was to determine the adequacy of existing assessment instruments. The data presented here suggest that proficiency tests may not always be the most appropriate instrument to assess language learning during the initial semesters of college instruction. We have argued that Novice ratings are in effect not consistent with the ethos of an instrument designed to measure learner proficiency since Novice ratings denote a learner that does not evidence functional ability in the language. If, as these data indicate, many students remain in the Novice range after two and sometimes even four semesters of language study, then an instrument predicated on demonstrating proficiency is not optimal. Rather, the adoption of a performance based assessment instrument (such as the AAPPL used in K-12 contexts), which is premised on the type of language behavior typical of Novice level learners and with finer gradations in ratings, might be more ecologically valid and provide more useful feedback to learners and language programs.

References

ACTFL. (2012a). *ACTFL proficiency guidelines 2012*. Alexandria, VA: American Council on the Teaching of Foreign Languages. Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

ACTFL. (2012b). *ACTFL performance descriptors for language learners 2012 edition*. Alexandria, VA: American Council on the Teaching of Foreign Languages.

ACTFL. (2012c). *ACTFL OPIc familiarization manual.* Retrieved from https://www.languagetesting.com/wp-content/uploads/2012/07/OPIc-Familiarization-Manual.pdf

ACTFL. (2013). *ACTFL reading proficiency test* (*RPT*). *Familiarization manual and ACTFL proficiency guidelines 2012—reading*. Retrieved from http://www.languagetesting.com/wp-content/uploads/2015/02/ACTFL_FamManual_Reading_2015.pdf

ACTFL. (2014). *ACTFL listening proficiency test* (*LPT*). *Familiarization manual and ACTFL proficiency guidelines 2012—listening*. Retrieved from http://www.languagetesting.com/wp-content/uploads/2015/02/ACTFL_FamManual_Listening_2015.pdf

Brown, J. D. (2004). Performance assessment: Existing literature and directions for research. *Second Language Studies, 22*(2), 91–139. http://hdl.handle.net/10125/40663

Davidson, D. E. (2010). Study abroad: When, how long, and with what results? Data from the Russian front. *Foreign Language Annals, 43*, 6–26. https://doi.org/10.1111/j.1944-9720.2010.01057.x

Goldberg, D., Looney, D., & Lusin, N. (2015). *Enrollments in languages other than English in United States institutions of higher education, Fall 2013*. Modern Language Association of America. Retrieved from https://www.mla.org/content/download/31180/1452509/EMB_enrllmnts_nonEngl_2013.pdf

Institute for Test Research and Test Development. (2013a). *Assessing evidence of validity of the ACTFL reading proficiency test (RPT)*. Retrieved from http://www.languagetesting.com/wp-content/uploads/2013/10/Technical-Report-ACTFL-RPT-for-publication.pdf

Institute for Test Research and Test Development. (2013b). *Assessing evidence of validity of the ACTFL listening proficiency test (LPT)*. Retrieved from http://www.languagetesting.com/wp-content/uploads/2013/10/Technical-Report-ACTFL-LPT-2013-for-publication.pdf

Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., & Hall, C. (2016). *NMC horizon report: 2016 higher education edition*. Austin, TX: The New Media Consortium.

Liskin-Gasparro, J. (1982). *ETS oral proficiency testing manual*. Princeton, NJ: Educational Testing Service.

Malone, M. E., & Montee, M. J. (2010). Oral proficiency assessment: Current approaches and applications for post-secondary foreign language programs. *Language and Linguistics Compass, 4*, 972–986. https://doi.org/10.1111/j.1749-818X.2010.00246.x

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256. https://doi.org/10.1177/026553229601300302

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496. https://doi.org/10.1191/0265532202lt241oa

Norris, J. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing, 19*(4), 337–346. https://doi.org/10.1191/0265532202lt234ed

Rifkin, B. (2005). A ceiling effect in traditional classroom foreign language instruction: Data from Russian. *Modern Language Journal, 89*, 3–18. https://doi.org/10.1111/j.0026-7902.2005.00262.x

Tschirner, E. (2016a). Listening and reading proficiency levels of college students. *Foreign Language Annals, 49*(2), 201–223. https://doi.org/10.1111/flan.12198

Tschirner, E. (2016b). *Task-based language assessment and testing for proficiency: Where do the twain meet?* A paper presented at the L.E.A.R.N. Workshop, Universities at Shady Grove, Rockville, MD, September 20–21, 2016.

Watson, J. R., & Wolfel, R. (2015). The intersection of language and culture in study abroad: Assessment and analysis of study abroad outcomes. *Frontiers: The Interdisciplinary Journal of Study Abroad, 25*, 57–72. Retrieved from: https://frontiersjournal.org/wp-content/uploads/2015/09/WATSON-WOLFEL-FrontiersXXV-TheIntersectionofLanguagenandCultureinStudyAbroad.pdf

Fernando Rubio is Professor of Spanish Linguistics at the University of Utah, where he also serves as the Co-Director of the Second Language Teaching and Research Center. His research interests include second language acquisition, language teaching methodology, and proficiency assessment. From 2014 to 2018 he served as the PI in the Language Flagship Proficiency Initiative grant at the University of Utah.

Jane F. Hacking is Associate Professor of Russian and Linguistics at the University of Utah, where she Co-Directs the Second Language Teaching and Research Center (L2TReC). Her research focuses on L2 phonology and the overall development of L2 proficiency. She received the 2017 award for Outstanding Contribution to the Profession from the American Association of Teachers of Slavic and East European Languages.

Exploring the Relationship Between Self-Assessments and OPIc Ratings of Oral Proficiency in French

Magda Tigchelaar

Abstract The present study analyzed the self-assessed spoken French language abilities that students said they 'can do' in relation to the ACTFL proficiency scores they received on an oral proficiency interview by computer (OPIc). A secondary aim was to assess different scales that have been used to convert OPIc ratings to numeric scores.

French university students (N = 216) of varying proficiency levels rated a series of can-do statements related to speaking skills. They then completed the ACTFL OPIc test, which was rated by certified ACTFL raters. A series of regression analyses showed that the strength of the relationship between self-assessment and OPIc ratings was strongly influenced by the type of numeric scale used: When data were ranked ordinally and analyzed using an ordinal regression, a majority (65%) of variance in OPIc scores was explained by self-assessment scores. Analyzed using linear regression, when scores were converted to equal-interval scales, self-assessment scores explained approximately 30% of variance. On a graduated scale that reflected the increasing distances between ACTFL (2012) proficiency levels, only 20% of variance was accounted for.

Keywords Self-assessment · Oral proficiency · Can-do statements · Concurrent validity · Correlation · Regression

1 Introduction

Research on self-assessment in second language (L2) learning has revealed that language learners are generally poor judges of their own performance, but that the use of can-do statements may help to sharpen their judgments (VanPatten, Trego, & Hopkins, 2015). One explanation for improvement comes from the movement toward assessment *for* language learning (Butler, 2016; Lee, 2016; Nikolov, 2016;

M. Tigchelaar (✉)
Western Michigan University, Kalamazoo, MI, USA
e-mail: magda.tigchelaar@wmich.edu

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_9

Purpura & Turner, 2014, 2015), which advocates for the use of can-do statements to push learners to gain awareness of their language abilities and deficiencies, allowing them to take a more active role in their assessment. The present study analyzed the self-assessed spoken French language abilities that students indicated they "can do" in relation to the ACTFL proficiency test scores they received on an oral proficiency interview by computer (OPIc).

In evaluating self-assessments of language proficiency, researchers commonly use correlation analyses to determine how well self-assessments relate to outside proficiency ratings (e.g., Ross, 1998). In order to do this, they must transform proficiency ratings into numeric values, which involves making a decision about the values to assign to each level. Several numeric scales exist that all propose different distances between proficiency levels (e.g., Brecht, Davidson & Ginsberg, 1995; Kenyon & Malabonga, 2001; Lange & Lowe, 1987; Meredith, 1990). The use of these differently weighted scales and the decisions that researchers make about which type of statistical analyses to perform with the data may influence the observation of the relationship between proficiency ratings and other variables. Thus, a secondary aim of the current study was to assess how the different scales that have been used to convert OPIc ratings to numeric scores can impact the strength of the relationship between proficiency ratings and self-assessments of spoken proficiency.

2 Literature Review

2.1 Self-Assessment of Oral Proficiency

One of the main areas of interest in research on self-assessment of oral proficiency has been concurrent validity. Specifically, researchers have considered how well self-assessment scores correlate with outside measures of L2 oral proficiency (Brown, Dewey & Cox, 2014; Malabonga, Kenyon & Carpenter, 2005; Ross, 1998; Trofimovich, Isaacs, Kennedy, Saito, & Crowther, 2014). These studies have produced a wide range of results: Some have shown strong correlations while others have shown weak or even non-significant relationships.

Some researchers have suggested that the different types of instruments used to conduct the self-assessments in these studies may in part explain why such a wide range of correlations has been reported in the literature (Brantmeier, 2006; Ross, 1998). For example, Trofimovich et al. (2014) considered English language learners' self-assessed ratings of how accented and comprehensible their speech was. They found weak correlations between the self-assessed measures and expert judgments (accent $r = .06$, $p = .50$; comprehensibility $r = .18$, $p = .03$). On the other hand, Brown et al. (2014) assessed the relationship between oral self-assessments using ACTFL (2015) can-do statements and oral proficiency interview (OPI) scores. They asked L2 Russian students to self-assess their oral proficiency prior to going on study abroad and after returning from study abroad and found significant

medium-sized correlations between both pre-study-abroad OPI scores and self-assessments (r = .27) and post-study-abroad OPI scores and self-assessments (r = .21). Comparing the instruments used in these two studies suggests that language learners may be better equipped to self-assess functional speaking skills (using can-do statements) than linguistic components of oral production, as the former scores showed stronger correlations with outside measures. As Brantmeier (2006) concluded, the use of criterion-referenced instruments such as can-do statements may help students better assess their speaking skills. However, more research in this area is needed to confirm this hypothesis.

One important use of self-assessment is to establish a starting point for test takers in computer adaptive test (CAT) contexts (Chalhoub-Deville & Deville, 1999). In the oral proficiency interview by computer (OPIc), discussed below, test takers begin by choosing one of five levels at which to begin the speaking test. One issue with this procedure is that if assessees over-estimate their abilities in the self-assessment, they may select a task or test form that is too difficult for them, making it difficult for outside raters to assess their proficiency. However, using self-assessments may help to guide test takers toward the appropriate starting point.

The most relevant study along this line of research (using self-assessment to guide test takers toward the appropriate starting point in a computer adaptive test) was conducted by Malabonga et al. (2005) at the Center for Applied Linguistics. The researchers designed a short self-assessment to guide computer-adaptive-speaking-test takers in choosing a starting level for their computerized oral assessment. The self-assessment was in the form of a questionnaire that included 18 questions. Based on their score on the self-assessment, one of four task levels was suggested for examinees to select for their first speaking task. The authors found that 92% of participants accurately used the self-assessment questionnaire and subsequently chose a starting level that was at an appropriate level of difficulty. They also found that the results of the self-assessments correlated strongly ($r = .88$) with the results of the oral proficiency test. One should note that this correlation has a certain amount of collinearity: That is, the outcome variable (the final test score) relied in part on the initial self-assessment outcome.

2.2 Measuring Oral Proficiency Using the ACTFL OPIc

As mentioned previously, many researchers who have investigated the validity of self-assessments of oral proficiency have looked at how well these assessments correlate with other-assessments of oral proficiency. Ross (1998) considered the wide range in correlation scores he found. He wrote that it is "important to consider that the criterion measures of speaking skill are likewise open to variation. Speaking skill is often assessed post hoc and holistically, by structural interviews that are biased towards formal control of grammar" (p. 9). These concerns can be addressed by using a well-studied, reliable and valid assessment of oral proficiency such as the ACTFL OPI or OPIc. Although these methods of assessing spoken proficiency have been critiqued

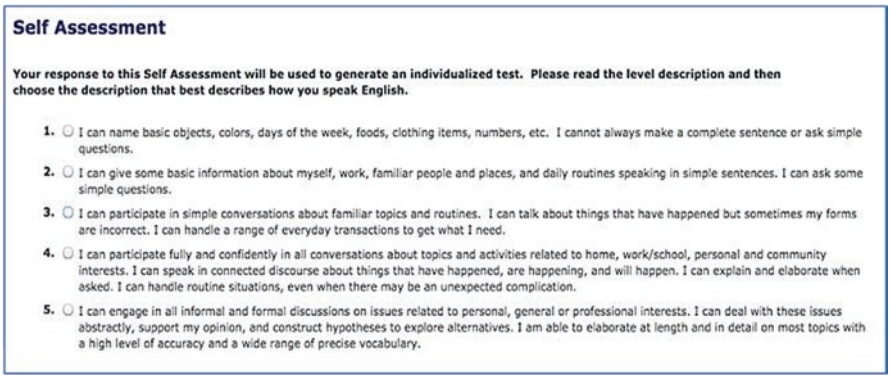


Fig. 1 ACTFL OPIc self-assessment. (Reprinted with permission of ACTFL)

(see, for example, Bachman, 1988; Bachman & Savignon, 1986; Malone & Montee, 2010), they have also been the focus of many studies that have established their reliability and validity (e.g., Dandonoli & Henning, 1990; Surface, Poncheri, & Bhavsar, 2008; Tschirner, Bärenfänger, & Wanner, 2012). For instance, in an investigation by an outside consulting service, Surface et al. (2008) found high inter-rater reliability, test-retest reliability, and construct validity for the ACTFL OPIc.

The ACTFL OPIc is a standardized speaking test that measures what language learners "can do with language…in real-world situations in a spontaneous and non-rehearsed context" (ACTFL, 2015). The test is administered over the Internet by an avatar that delivers questions to the test taker. The test can be considered to be somewhat adaptive as the test form generated depends on which level an examinee chooses when they complete a simple self-assessment of their oral proficiency, shown in Fig. 1. The test lasts 20–30 min and the resulting speech sample is recorded and rated by a certified ACTFL rater by comparing the OPIc performance to the ACTFL Proficiency Guidelines (2015). Ratings are given in terms of five major levels, or thresholds: Novice, Intermediate, Advanced, Superior and Distinguished. The first three thresholds are further subdivided into High, Mid and Low sublevels.

2.3 Converting ACTFL Proficiency Ratings to Numerical Scores

Researchers must convert proficiency ratings on descriptive scales, such as the ACTFL (2012) Proficiency Guidelines, into numerical scores that will lend themselves to statistical analyses. How to do this is an empirical question in and of itself. One option is to simply rank the hierarchy of proficiency levels on a scale from 1 (Novice-Low) to 9 (Advanced-High) or 10 (Superior). Kenyon and Malabonga (2001) used this approach in a study in which they compared test-takers' results on two different oral proficiency assessments. This type of conversion maintains the

ordinal nature of the ACTFL scale where the data are ranked from low to high. However, this conversion does not provide any information about the distance between the points on the scale (Field, 2009), and thus should be analyzed accordingly. Furthermore, Ross (1998) warned that comparing self-assessments to speaking assessments based on "noninterval rating scale criteria…could lead to a truncated correlation" (p. 9). When correlating two variables using Pearson's *r,* one of the assumptions is that the range of the data is not truncated. If there are differences in distance between the levels on the ACTFL scale, using an ordinal scale truncates, or condenses the range. This could lead to a correlation size that underestimates the true relationship between the two variables.

A second option is to use an existing conversion that proposes a measure of the distances between each of the sublevels in the scale (i.e., Low, Mid, High) or distances between proficiency thresholds (i.e., Novice, Intermediate, Advanced, Superior). One such scale was proposed by Lange and Lowe (1987), one of the authors of the ACTFL Proficiency Guidelines, and has been used by many researchers since (Dandonoli & Henning, 1990; Kenyon & Tschirner, 2000; Vande Berg, Connor-Linton, & Paige, 2009). This scale uses an increase of 0.2 points from Low to Mid, an increase of 0.5 points from Mid to High, and an increase of 0.3 points from High to the lowest level of the next threshold. These values suggest that to advance from Novice-Low to Novice-Mid, for example, represents a smaller increase in proficiency than to move from Novice-Mid to Novice-High or from Novice-High to Intermediate-Low. It also implies that moving from Novice-Low to Novice-Mid represents the same gain in proficiency as improving from Advanced-Low to Advanced-Mid. It is unclear why these measurements are used, as the authors do not provide a justification for the differences in distances between sublevels. Further, these measures do not appear to reflect the inverted pyramid shape that Lowe (1985) suggested to represent the ACTFL scale, since the distances between levels at the base of the scale are smaller than those at the top of the scale.

As the authors of the Boren awards report (Mason, Powers, & Donnelly, 2015) acknowledged, "the increasing width [of the pyramid] demonstrates that sublevel gains are not proportionate and that each sublevel advancement requires a greater amount of time and effort from the learner" (p. 12). In their study of oral proficiency gains after study abroad, however, the authors did not calibrate the scale they used to convert OPI ratings accordingly. Instead, they calculated oral language proficiency gains using the same scale as Brecht et al. (1995) and Davidson (2010). This scale posits an equal increase in sublevel scores for the Novice (1, 2, 3) Intermediate (5, 6, 7) and Advanced (9, 10, 11) levels with a one point increase between each of the thresholds.

According to Meredith (1990), a numeric representation of the levels on the pyramid "should reflect those unequal intervals with increasingly greater distances between the higher levels" (p. 289). He provided support for this theory by testing how well prior experience using Spanish as a foreign language, measured in months, could predict OPI ratings. He converted the ratings into five different numeric scales: two were equal-interval scales, and the remaining three were graduated scales. He found that Spanish speaking experience had greater predictive power for

OPI scores measured by the graduated scales than the equal-interval scales, and argued that the OPI ratings should be calibrated with increasing distances between levels when used for research purposes or for assigning grades.

Two issues are of note when considering the scales reviewed above. First, very little empirical evidence has been provided for the distances between proficiency levels proposed in each of the scales. No justification is provided by the creators of the Lange & Lowe (1987) or Boren scales for the distances between levels, and yet they are frequently used in language proficiency research. Meredith's (1990) scale was developed based on a single study, but this scale has not been widely used since and subsequent research has not validated the distances he proposed. A second issue is that most proficiency research conducted with the ACTFL scale involves a numeric conversion to either the ordinal scale or one of the weighted scales with the use of parametric statistical tests, even though the scales are not necessarily linear. Exceptions do exist, such as Thompson, Cox, and Knapp (2016), who used the ordinal scale and Spearman correlations to compare OPI and OPIc scores, but it is common practice to violate the assumptions and include an acknowledgment (e.g., Tschirner, 2016). In addition to the violation of the assumption of linearity, the assumption that the data will have a normal distribution and that measures of central tendency apply is problematic. This use of parametric statistics on non-linear data that are not normally distributed may influence the results that are reported by researchers.

In sum, the literature reviewed above highlights the importance of using appropriate instruments for language learners to perform meaningful self-assessments of their spoken language abilities, such as contextualized can-do statements that are related to established criterion like the ACTFL (2012) guidelines. Researchers have observed higher correlations between can-do self-assessments and ACTFL OPIc ratings than previous instruments used to measure oral proficiency (Brown et al., 2014). Another important issue in evaluating the strength of the relationship between self- and other-assessments lies in how proficiency ratings are converted to numeric scores, and specifically the lack of uniformity in the scales that have been used to accomplish this. With this in mind, I formulated two main research questions to guide this study.

2.4 Research Questions

1. What is the relationship between what students say they can do (self-assessment) and the ACFTL proficiency level they are assigned based on OPIc (other-assessment)? Specifically, how well do self-assessment scores predict OPIc ratings?
2. How do different conversions of OPIc ratings to numeric scores impact the observation of this relationship?

3 Method

3.1 Context and Participants

This study draws on data that was collected as part of the National Security Education Program's (NSEP, 2016) Language Flagship proficiency testing initiative at Michigan State University, which provided language testing to L2 learners at Michigan State University in four foreign languages over the course of 3 years. The participants in the present study were those who took the French speaking test at the beginning of the second year of testing (in Spring 2015). In order to represent a range of proficiency levels, students were selected from a number of intact French classes at four different class levels: FREN102 (second semester of university study; $N = 79$), FREN202 (fourth semester of study; $N = 65$), 300-level ($N = 42$) and 400-level ($N = 35$) French classes. A total of 221 participants completed both the self-assessment and the French oral proficiency interview for the time period under consideration. Of these, 94% (207) of the participants received an ACTFL OPIc rating based on their performance; 6% (14) of the participants were unrated because they either over-assessed their ability on the self-assessment (described below) and as a result took a test that was too difficult to generate a rating or were unrated due to technical difficulties.

3.2 Materials

The materials included a background questionnaire that collected data on participants' language learning experience (L1, class level at time of testing, classes they had completed in the language, heritage language learning experience, study abroad experience, other languages studied, high school language study experience). In addition, participants indicated their purpose for studying French and gave a Likert-scale (1–6) rating of how important studying the language was for them. They also indicated their age, gender, major, and minor (if they had one).

A second questionnaire included five sets of ten can-do statements that were selected by the principal investigator (PI, Paula Winke) and the Language Flagship Proficiency Team at MSU to represent the spoken ACTFL (2015) can-do statements that fall under the interpersonal communication and presentational modes. These statements were selected so that (when possible) only one skill was addressed per statement. Each statement was followed by a Likert scale where participants could rate their ability from one to four: 1 (I cannot do this yet), 2 (I can do this with much help), 3 (I can do this with some help), 4 (Yes, I can do this). The five sets of questionnaire statements were designed based on personal communications the team had with ACTFL and included the following ranges of proficiency levels: 1 (novice-low

to novice-high), 2 (intermediate-low to advanced-low), 3 (intermediate-mid to advanced-mid), 4 (intermediate-high to advanced-high) and 5 (advanced-low to superior). Each of the levels was accompanied by a brief, general description of the language abilities of learners whose proficiency falls within the range it represented. The descriptions and corresponding sets of can-do statements are presented in Appendix 1, and readers can access the questionnaire at https://msu.co1.qualtrics.com/jfe/form/SV_6hVFcyfYXkyW1sF. Additionally, more information on an earlier version of the survey and the how the five sets were presented are in a supplemental file from Tigchelaar, Bowles, Winke, and Gass (2017) that can be downloaded from the IRIS database at https://www.iris-database.org.

3.3 Procedure

The procedure that participants followed was similar to that of taking the ACTFL OPIc, with one modification: after completing the background questionnaire, participants proceeded to complete the can-do questionnaire before selecting the general description of their language ability. Each participant began the questionnaire at the first level, where they gave a Likert-scale rating from 1 to 4 for each of the 10 statements. If they rated at least 9 out of 10 of the statements as a 4, they were instructed to proceed to the next level of 10 questions on the questionnaire. They continued to rate can-do statements in this way until they reached a level where they assessed that they could no longer do 9 out of 10 statements. Based on the number of can-do statements participants rated as a 4 on the scale, the corresponding level was recommended for them to select. An OPIc test form was selected according to the participants' level choice. After taking the test, students' speech samples were rated according to the ACTFL (2012) guidelines by a certified rater and assigned a proficiency level. The breakdown of these ratings for students at each level is presented in Table 1.

Table 1 ACTFL ratings for participants based on class level

ACTFL rating		102 (N = 74)	202 (N = 64)	300-level (N = 39)	400-level (N = 30)
Novice	Low (N = 14)	12	2		
	Mid (N = 31)	23	8		
	High (N = 53)	23	24	6	
Intermediate	Low (N = 44)	14	16	11	3
	Mid (N = 31)	2	9	10	10
	High (N = 23)		5	9	9
Advanced	Low (N = 5)			1	4
	Mid (N = 5)			2	3
	High (N = 1)				1

Table 2 Scaled ACTFL proficiency ratings

ACTFL rating (N)		Ordinal scale	Boren scale	Lange & Lowe scale	Meredith scale
Novice	Low (14)	1	1	0.1	1
	Mid (31)	2	2	0.3	3
	High (53)	3	3	0.8	7
Intermediate	Low (44)	4	5	1.1	12
	Mid (31)	5	6	1.3	24
	High (23)	6	7	1.8	48
Advanced	Low (5)	7	9	2.1	96
	Mid (5)	8	10	2.3	128
	High (1)	9	11	2.8	256

3.4 Data Analysis

The self-assessment data were tabulated by tallying the Likert scale ratings (from 1 to 4) of the can-do statements on each of the five questionnaires, resulting in a total self-assessment score for each participant. In order to quantify the other-assessment data, the ACTFL proficiency ratings were converted to numeric scores based on four scales, represented in Table 2. The first was an ordinal scale ranking each proficiency level from 1 (Novice-Low) to 9 (Advanced-High). The second was the scale that was used to calculate oral language proficiency gains in the Boren Awards report (Mason et al., 2015), which has a one-point increase from Novice to Intermediate and Intermediate to Advanced levels. The third scale was the scale proposed by Lange & Lowe (1987), which increases by 0.2 points from Low to Mid, by 0.5 points from Mid to High, and by 0.3 points from High to the lowest level of the next threshold. The final scale was graduated "with increasingly greater points awarded for higher levels to reflect the inverted pyramid" (Meredith, 1990, p. 291) of the ACTFL (2012) scale.

Using the self-assessment and scaled other-assessment scores, I performed two types of regression analyses. First, I conducted an ordinal regression to see how well the self-assessment scores would predict the proficiency ratings on the ordinal scale. In addition, I conducted a series of linear regression analyses with self-assessment scores as predictors of proficiency ratings scaled using the Boren scale (Mason et al., 2015), the Lange & Lowe (1987) scale and the Meredith (1990) scale. Although the assumptions of linearity and normality of error distribution were violated, (see Appendix 2), I chose to use linear regressions since this mirrors common practices of proficiency researchers.

4 Results

The results that follow concern the relationship between French language learners' self-assessment of spoken proficiency and the ratings they received on their OPIc performance. In addition, the results show the predictive strength of self-assessment scores for OPIc ratings that were numerically scaled in four different ways.

Table 3 Descriptive statistics for self-assessments (total Likert score) and scaled proficiency ratings

	N	M (SD)	Minimum	Maximum	95% C.I.	
Self-assessment	221[a]	55.36 (45.84)	0	200	47.69	59.66
Ordinal scale[b]	207	3.82 (1.66)	1	9	3.59	4.04
Boren scale	207	4.41 (2.21)	1	11	4.10	4.70
Lange & Lowe scale	207	1.01 (0.56)	0.1	2.8	0.93	1.08
Meredith scale	207	20.69 (30.40)	1	256	16.56	24.82

[a]More data are reported for self-assessment scores than OPIc scores because 14 participants over-assessed their ability on the self-assessment and took an OPIc test that was beyond their proficiency level. Therefore, they did not receive an OPIc score
[b]Median score = 4.00

I operationalized the predictor variable, self-assessment, as the sum of the Likert ratings students provided on the can-do questionnaire. The reliability (Cronbach's alpha) of this 50-item assessment was .84. I operationalized the outcome variable, OPIc ratings, by converting the ratings onto four numeric scales. Descriptive statistics for the predictor variable and different measures of the outcome variable are in Table 3.

I first conducted an ordinal regression analysis to investigate how well self-assessment scores could predict OPIc ratings ranked on the ordinal scale. Ordinal regression assumes that the dependent variable is measured at the ordinal level (like the proficiency ratings scaled to the hierarchical scale) and that the predictor variables are either categorical or continuous (like the total Likert rating scores; Laerd Statistics, 2013). The analysis of the model fit indicates that the model including the self-assessment scores as a predictor of proficiency rating is a significant improvement over the fit of the null model with no predictors, χ^2 (54) = 209.56, $p < .001$. This result also indicates that an increase in total self-assessment score was associated with an increase in ACTFL proficiency rating. A Nagelkerke R^2 value of .646 indicates that the observed fitted model is a 65% improvement over the prediction of the null model, and that the model accounts for 65% of variance in scores. Because ordinal regression does not provide a correlation coefficient, to determine the strength of a correlation between the self-assessment scores and OPIc scores converted to the ordinal scale, I used a non-parametric test that ranks the data (Spearman's rho), $\rho = .64$, $p < .001$.

Next I conducted three linear regression analyses to evaluate the relationship between self-assessment scores and other-assessment scores that I converted to the three numeric scales. All of the models were statistically significant ($p < .001$) predictors of the outcome variable, presented in Table 4.

Self-assessment scores had a similar relationship to the OPIc scores that I converted to the Boren, and Lange and Lowe scales: I observed a positive, moderate correlation (between $R = .54$ and $R = .55$), and these two models accounted for nearly 30% of the variance in proficiency scores. On the other hand, the model with self-assessment scores predicting proficiency scores on the graduated scale proposed by Meredith (1990) accounted for approximately 20% of the variance in

Table 4 Linear regression analysis results

Model	R	R^2	β	β_1
Boren scale	.54	.29	.03	.54
Lange & Lowe scale	.55	.30	43.09	.55
Meredith scale	.46	.21	.28	.46

scores. The strength of the relationship between the predictor and outcome variables was still moderate ($R = .46$), though it was weaker than the relationship between the predictor variable and the ratings that were converted to the other three scales.

5 Discussion

With this research, I wanted to evaluate how well French learners' self-assessments using can-do statements could predict their performances on an ACTFL OPIc assessment. A secondary aim was to investigate how the relationship between self-assessment and other-assessment can be influenced by using some of the different scales that have been used to convert OPI ratings to numeric scores. The two research questions go hand in hand because the type of scale that is used directly impacts the observation of the relationship between self- and other-assessment. Differences between observed relationships may lead researchers to different interpretations about the usefulness of an instrument or about test takers' abilities based on their use of an instrument, which is problematic.

Research on self-assessment of oral proficiency has not produced consistent or conclusive results about the accuracy of language learners' judgments of their spoken abilities. Ross (1998) found a wide range of correlations in studies that compared self-assessment and other-assessment of oral proficiency and suggested that evaluating this productive skill is strongly influenced by external factors such as the instruments being used to conduct both the self- and other-assessments. More recent research has shown that the use of general questionnaires (e.g., Brantmeier, 2006) and fine-grained linguistic measures (e.g., Trofimovich et al., 2014) for self-assessment do not correlate strongly with outside measures. On the other hand, the use of self-assessment instruments that are criterion-referenced and that target functional, contextualized speaking skills using can-do statements have stronger correlations with other-assessments (e.g., Brown et al., 2014). These results suggest that the accuracy of self-assessment *depends on the type of instrument used*. Further study is needed to validate these findings. This will help to push the conversation beyond simply asking whether language learners are able to self-assess their abilities to a more nuanced discussion of what type of instruments allow for more accurate self-assessments.

The first research question revealed a moderate, positive relationship between self-assessment scores and OPIc proficiency ratings, with correlation coefficients between .46 and .64. The strongest relationship I observed was between the ratings

converted to the ordinal scale and the self-assessment, $\rho = .64$, which is considered large (Plonsky & Oswald, 2014). The other three scaled proficiency ratings had moderate correlations with self-assessment scores. Generally speaking, as self-assessment scores increased, so did participants' OPIc performance ratings. These findings are in line with Malabonga et al., (2005), who also found a strong relationship between self-assessments and OPI ratings, with a correlation of $r = .88$. This relationship and the one observed in the current study are stronger than that observed by Brown et al. (2014), who found a correlation of $r = .21$ for pre-study-abroad self-assessment scores and OPIc scores and $r = .27$ post-study-abroad. One possible explanation for this difference in correlation strength is the manner in which the can-do statements were modified across the two studies. In the case of Brown et al. (2014), the authors used statements from the NCSSFL-ACTFL Can-Do Statements (ACTFL, 2015) and modified them to reflect what participants could do before they studied abroad and what they were able to do after. They provided the example "I could exchange detailed information on topics within and beyond my fields of interest" (p. 269). Within this single statement, there are two abilities addressed: exchanging information about one's interests and exchanging information beyond one's interests. In the present study, the research team took care to select NCSSFL-ACTFL Can-Do Statements (ACTFL, 2015) that included only one skill per statement for the most part so that learners could rate distinct speaking skills for the interpersonal communication and presentation modes. This fine-tuning of the can-do statements may explain the stronger correlation scores between self-assessment and OPIc ratings.

Why is it that the correlations were weaker using the scaled proficiency ratings ($r = .46$–$.54$) than the ordinal scores ($\rho = .64$)? One possible explanation has do with how well the scales considered in the analysis represent increases in language proficiency. It may be more appropriate to rank language use from less proficient to more proficient than to quantify exactly how much better a given level is from another. This is what the numeric conversions propose to do, and the fact that weaker correlations are observed using these scores than the ordinal ranking suggests that more work needs to be done to better quantify increases in language proficiency along the ACTFL (2012) scale.

I also looked beyond correlation coefficients to investigate how well self-assessment scores could predict outcomes on the OPIc using regression analyses. The linear regressions accounted for 20–30% of the shared variance in OPIc and self-assessment scores, which can be considered moderate: Bachman (2004) gives an example of a similar shared variance (34%) between writing scores and a teacher's ranking of students in a class. He points out that with this amount of shared variance the two assessments likely do not measure exactly the same skills, but one could conclude that "the test and the classroom teacher rankings provide complementary information, and thus decide to use both" (p. 104). In terms of the present study, the observed shared variance indicates that the self-assessment instrument and the OPIc are not measuring exactly the same aspects of oral proficiency. However, as Green (2014) highlighted,

> No assessment task is entirely satisfactory. Each format has its own weaknesses. Rather than searching for one ideal task type, the assessment designer is better advised to include a reasonable variety in any test or classroom assessment system so that the failings of one format do not extend to the overall system. (p. 140)

In addition to standardized tests such as the OPIc, self-assessments using can-do statements can be incorporated to contribute to the variety of assessments that Green (2014) calls for. Further, they can be used to predict some of the variance in proficiency test scores. This can be helpful in computer adaptive test taking scenarios (like in Malabonga, et al., 2005), where students self-assess prior to selecting a starting point for a test.

The results of the ordinal regression analysis painted a slightly different picture: This regression model had a much larger R^2 value (.65) than the linear models. This finding leads to the discussion of the second research question: Maintaining the ordinal nature of the OPIc ratings with a numeric conversion and analyzing these data with an ordinal regression resulted in the model with the strongest predictive power. This result was more than double that of any of the linear regression analyses: the OPIc ratings that were scaled using the Boren scale (Mason et al., 2015) and the Lange & Lowe (1987) scale shared 29% and 30% of the variance, respectively, with self-assessment scores. Ratings that were scaled to the graduated scale proposed by Meredith (1990) shared only 20% of the variance with self-assessment scores. Depending on the scale and analysis used, the self-assessment and OPIc assessment can appear to measure a small fraction of overlapping aspects of oral proficiency (e.g., linear regression using the Meredith (1990) scale) or a large proportion (e.g., ordinal regression using the ordinal scale). This means that depending on one's purpose, researchers could cherry pick the most convincing (or unconvincing) result to show how well (or poorly) OPIc ratings relate to other-assessments. For example, using the data from this study can show that language learners' self-assessments are weak predictors (Meredith scale, $R^2 = .21$) or strong predictors (ordinal scale, $R^2 = .65$) of proficiency ratings, which might influence an instructor's or language program director's decision to use self-assessments or not.

The findings of this study contribute to the wider discussion of self-assessment in language learning and have implications for language assessment, instruction, and assessment research. The existing literature on self-assessment has shown that language learners are not able to accurately gauge their L2 proficiency, particularly when linguistic components of L2 speech are concerned (Trofimovich et al., 2014). Research on the use of can-do statements for self-assessment has shown stronger correlations between self- and other-assessment (Brown et al., 2014; Malabonga et al., 2005). The present study found correlation sizes that were between those found in previous studies, perhaps due to differences in the instrument used or the difference in population. Pedagogically speaking, self-assessment is a valuable tool as it can help to develop learner autonomy and can save language instructors time. The findings of the current study provide further incentive for language instructors to include can-do self-assessments for language learning (Butler, 2016; Lee, 2016; Nikolov, 2016; Purpura & Turner, 2015) and to evaluate classroom-based language learning (VanPatten, Trego & Hopkins, 2015). The moderate to strong predictive

validity of the self-assessment observed in the present study for OPIc scores also has implications for diagnostic and placement testing. This result provides support for the use of can-do statements as an initial diagnostic tool that can direct test takers and administrators toward the appropriate form of test for individual language learners (Chalhoub-Deville & Deville, 1999) and for using these statements for raising language users' awareness of their approximate proficiency level (Glover, 2011).

Finally, the findings of this study have implications for conducting research on language assessment. First, language learners *may be able* to more accurately self-assess using can-do statements, and particularly if they use statements that address one skill at a time. Secondly, this study showed that the type of scale used to convert proficiency ratings and the type of statistical analysis used have important impacts on the results. Using ordinal regression with proficiency ratings ranked on an ordinal scale resulted in far higher R^2 values than linear regression, and the linear regression analyses using scales that had similar intervals shared more variance with self-assessment scores than proficiency scores that were scaled to a graduated scale designed to reflect the inverted pyramid that represents the ACTFL proficiency levels. These observations are no more than that: *observations* of the relationship between self- and other-assessment. It is likely that the strength of the true relationship lies somewhere within the range of observed correlations and R^2 values. It was beyond the scope of this research to determine which scale most accurately reflects the distance between levels. Most likely, each individual measurement is a fairly good observation of the underlying true score: Each one has some measurement error (all measurements are just estimations, after all). Future research should build on the work of Meredith (1990) to determine whether there is one, more reliable scale that researchers can use to represent differences in proficiency levels.

The current research is limited in that the vast majority of the test takers in this research were at novice- and intermediate-level speaking proficiency. As Byrnes and Ortega (2008) highlighted, the study of advanced language learners is under-researched, and future research on self-assessment of speaking abilities should include more of this population. A second limitation of note is that this study only provides a cross-sectional view of self- and other-assessment. One way to address this limitation, however, is to conduct future research that considers the effect of time and assessment experience on language learners' ability to self-assess. Previous research has demonstrated that learners' self-assessments show improvements and become more refined over multiple rounds (Chen, 2008; Glover, 2011). The data collection for the Flagship Proficiency initiative is ongoing and tracks the assessment of the same participants year after year. Thus, it may be possible to compare self-assessments and proficiency outcomes from multiple years of testing. This may contribute to a more comprehensive picture of the relationship between self- and other-assessment.

Appendices

Appendix 1: ACTFL OPIc 1–5 Levels and Can-Do Statements

ACTFL OPIc level 1: I can name basic objects, colors, days of the week, foods, clothing items, numbers, etc. I cannot always make a complete sentence or ask simple questions.

	Can-do statements	ACTFL Levels	Mode
❑	I can say the date and the day of the week.	*NL*	*PS*
❑	I can list the months and seasons.	*NL*	*PS*
❑	I can say which sports I like and don't like.	*NM*	*PS*
❑	I can list my favorite free-time activities and those I don't like.	*NM*	*PS*
❑	I can state my favorite foods and drinks and those I don't like.	*NM*	*PS*
❑	I can talk about my school or where I work.	*NM*	*PS*
❑	I can talk about my room or office and what I have in it.	*NM*	*PS*
❑	I can list my classes and tell what time they start and end.	*NM*	*PS*
❑	I can answer questions about where I'm going or where I went.	*NM*	*IC*
❑	I can present information about something I learned in a class or at work.	*NH*	*PS*

ACTFL OPIc level 2: I can give some basic information about myself, work, familiar people and places, and daily routines speaking in simple sentences. I can ask some simple questions.

	Can-do statements	ACTFL Levels	Mode
❑	I can describe a school or workplace.	*IL*	*PS*
❑	I can describe a place I have visited or want to visit.	*IL*	*PS*
❑	I can ask for help at school, work, or in the community.	*IL*	*IC*
❑	I can talk about my daily routine.	*IM*	*IC*
❑	I can talk about my interests and hobbies.	*IM*	*IC*
❑	I can schedule an appointment.	*IM*	*IC*
❑	I can talk about my family history.	*IH*	*IC*
❑	I can plan an outing with a group of friends.	*IH*	*IC*
❑	I can explain why I was late to class or absent from work and arrange to make up the lost time.	*AL*	*IC*
❑	I can tell a friend how I'm going to replace an item that I borrowed and broke/lost.	*AL*	*IC*

ACTFL OPIc level 3: I can participate in simple conversations about familiar topics and routines. I can talk about things that have happened but sometimes my forms are incorrect. I can handle a range of everyday transactions to get what I need.

	Can-do statements	ACTFL Levels	Mode
❑	I can give some information about activities I did.	*IM*	*IC*
❑	I can talk about my favorite music, movies, and sports.	*IM*	*IC*
❑	I can describe a childhood or past experience.	*IM*	*PS*
❑	I can ask for and follow directions to get from one place to another.	*IH*	*IC*
❑	I can return an item I have purchased to a store.	*IH*	*IC*
❑	I can arrange for a make-up exam or reschedule an appointment.	*IH*	*IC*
❑	I can present an overview about my school, community, or workplace.	*AL*	*PS*
❑	I can compare different jobs and study programs in a conversation with a peer.	*AL*	*IC*
❑	I can discuss future plans, such as where I want to live and what I will be doing in the next few years.	*AM*	*IC*
❑	I can explain an injury or illness and manage to get help.	*AM*	*IC*

ACTFL OPIc level 4: I can participate in fully and confidently in all conversations about topics and activities related to home, work/school, personal and community interests. I can speak in connected discourse about things that have happened, are happening, and will happen. I can explain and elaborate when asked. I can handle routine situations, even when there may be an unexpected complication.

	Can-do statements	ACTFL Levels	Mode
❑	I can present ideas about something I have learned, such as a historical event, a famous person, or a current environmental issue.	*IH*	*PS*
❑	I can give a presentation about my interests, hobbies, lifestyle, or preferred activities.	*IH*	*PS*
❑	I can ask for and provide descriptions of places I know and also places I would like to visit.	*IH*	*IC*
❑	I can explain how life has changed since I was a child and respond to questions on the topic.	*AL*	*IC*
❑	I can discuss what is currently going on in another community or country.	*AL*	*IC*
❑	I can provide a rationale for the importance of certain classes, subjects, or training programs.	*AL*	*PS*
❑	I can talk about present challenges in my school or work life, such as paying for classes or dealing with difficult colleagues.	*AM*	*IC*
❑	I can exchange general information about leisure and travel, such as the world's most visited sites or most beautiful places to visit.	*AM*	*IC*
❑	I can give a presentation about cultural influences on society.	*AH*	*PS*
❑	I can participate in conversations on social or cultural questions relevant to speakers of this language.	*AH*	*IC*

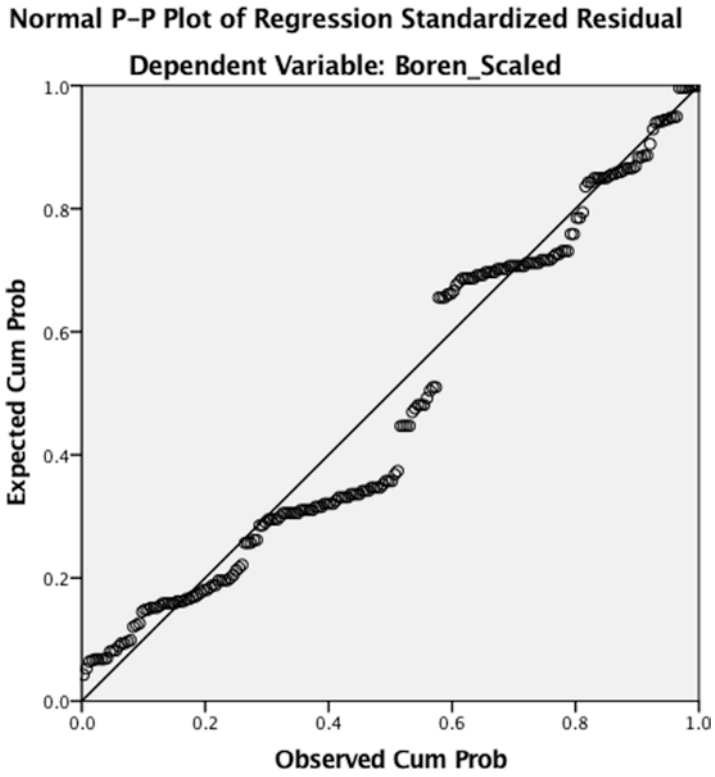
ACTFL OPIc level 5: I can engage in all informal and formal discussions on issues related to personal, general or professional interests. I can deal with these issues abstractly, support my opinion, and construct hypotheses to explore alternatives. I am able to elaborate at length and in detail on most topics with a high level of accuracy and a wide range of precise vocabulary.

	Can-do statements	ACTFL Levels	Mode
❑	I can interview for a job or service opportunity related to my field of expertise.	*AL*	*IC*
❑	I present an explanation for a social or community project or policy.	*AL*	*PS*
❑	I can present reasons for or against a position on a political social issue.	*AL*	*PS*
❑	I can give a clear and detailed story about childhood memories, such as what happened during vacations or memorable events and answer questions about my story.	*AM*	*IC*
❑	I can exchange general information about my community, such as demographic information and points of interests.	*AM*	*IC*
❑	I can exchange factual information about social and environmental questions, such as retirement, recycling, or pollution.	*AM*	*IC*
❑	I can usually defend my views in a debate.	*AH*	*IC*
❑	I can exchange complex information about my academic studies, such as why I chose the field, course requirements, projects, internship opportunities, and new advances in my field.	*AH*	*IC*
❑	I can provide a balance of explanations and examples on a complex topic.	*S*	*PS*
❑	I can explain participate actively and react to others appropriately in academic debates, providing some facts and rationales to back up my statements.	*S*	*IC*

Appendix 2: Plots for Checking Assumptions

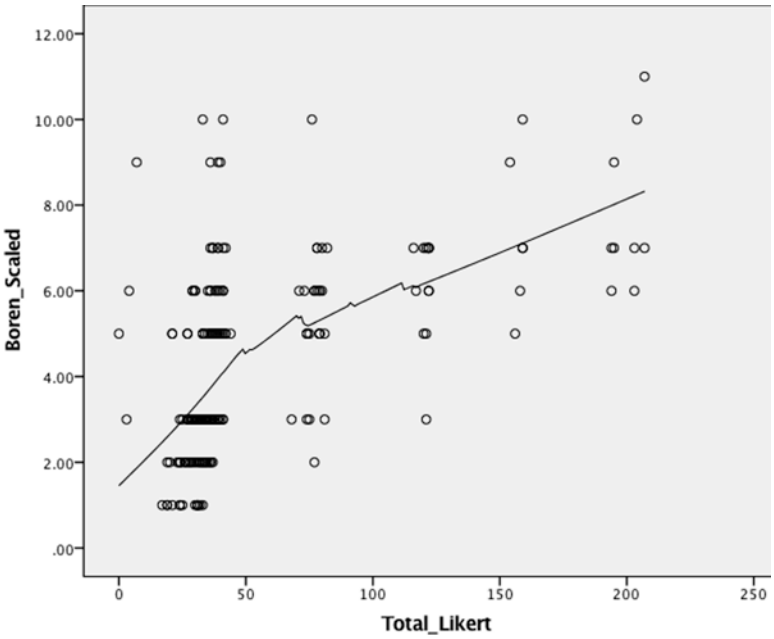
1. Scatter plots of the dependent and independent variables (linearity):

Absence of a straight line suggests that the data are non-linear.



2. Scatter plots of the standardized residuals (normality of error distribution):

   The bulges at 6.0 and 8.0 actually suggest that the data are bimodal.



References

ACTFL. (2012). *ACTFL proficiency guidelines – speaking*. Retrieved from http://www.actfl.org

ACTFL. (2015). *NCSSFL-ACTFL can-do statements*. Retrieved from http://www.actfl.org/global_statements

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*, 149–164. https://doi.org/10.1017/S0272263100007282

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F., & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal, 70*, 380–390. https://doi.org/10.1111/j.1540-4781.1986.tb05294.x

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System, 34*(1), 15–35. https://doi.org/10.1016/j.system.2005.08.004

Brecht, D., Davidson, D., & Ginsberg, B. (1995). Predictors of foreign language gain during study abroad. In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 37–66). Philadelphia, PA: John Benjamins.

Brown, N. A., Dewey, D. P., & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals, 47*(2), 261–285. https://doi.org/10.1111/flan.12082

Butler, Y. G. (2016). Self-assessment of and for young learners' foreign language learning. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 291–315). New York, NY: Springer International Publishing.

Byrnes, H., & Ortega, L. (2008). *The longitudinal study of advanced L2 capacities*. New York, NY: Routledge.

Chalhoub–Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics, 19*, 273–299. https://doi.org/10.1017/S0267190599190147

Chen, Y. M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12*(2), 235–262. https://doi.org/10.1177/1362168807086293

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals, 23*(1), 11–21. https://doi.org/10.1111/j.1944-9720.1990.tb00330.x

Davidson, D. (2010). Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals, 43*(1), 6–26. https://doi.org/10.1111/j.1944-9720.2010.01057.x

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Glover, P. (2011). Using CEFR level descriptors to raise university students' awareness of their speaking skills. *Language Awareness, 20*(2), 121–133. https://doi.org/10.1080/09658416.2011.555556

Green, A. (2014). *Exploring language assessment and testing*. New York, NY: Routledge.

Kenyon, D. M., & Malabonga, V. M. (2001). Comparing examinees' attitudes toward a computerized oral proficiency assessment. *Language Learning & Technology, 5*, 60–83. Available at http://llt.msu.edu/vol5num2/pdf/kenyon.pdf

Kenyon, D. M., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal, 84*(1), 85–101. https://doi.org/10.1111/0026-7902.00054

Laerd Statistics. (2013). *Ordinal regression using SPSS Statistics*. Available from https://statistics.laerd.com/spss-tutorials/ordinal-regression-using-spss-statistics.php

Lange, D. L., & Lowe, P. (1987). Grading reading passages according to the ACTFL/ETS/ILR reading proficiency standard: Can it be learned? *Selected papers from the 1986 Language Testing Research Colloquium* (pp. 111–127). Monterey, CA: Defense Language Institute. Available at https://archive.org/details/ERIC_ED287291

Lee, I. (2016). Putting students at the centre of classroom L2 writing assessment. *Canadian Modern Language Review, 72*(2), 258–280. https://doi.org/10.3138/cmlr.2802

Lowe, P. (1985). The ILR proficiency scale as a synthesizing research principle: The view from the mountain. In J. J. Charles (Ed.), *Foreign language proficiency in the classroom and beyond* (pp. 9–54). Lincolnwood, IL: National Textbook Company. Available at https://eric.ed.gov/?id=ED253104

Malabonga, V. M., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing, 22*(1), 59–92. https://doi.org/10.1191/0265532205lt297oa

Malone, M., & Montee, M. (2010). Oral proficiency assessment: Current approaches and applications for post-secondary foreign language programs. *Language and Linguistics Compass, 4*(10), 972–986. https://doi.org/10.1111/j.1749-818X.2010.00246.x

Mason, L., Powers, C., & Donnelly, S. (2015). *The Boren awards: A report of oral language proficiency gains during academic study abroad*. New York: Institute of International Education. Available at https://www.iie.org/Research-and-Insights/Publications/The-Boren-Awards-A-Report-Of-Oral-Language-Proficiency-Gains

Meredith, R. A. (1990). The oral proficiency interview in real life: Sharpening the scale. *The Modern Language Journal, 74*(3), 288–296. https://doi.org/10.1111/j.1540-4781.1990.tb01065.x

National Security Education Program. (2016). *The language flagship*. Retrieved from http://www.nsep.gov/content/language-flagship

Nikolov, M. (2016). A framework for young EFL learners' diagnostic assessment: 'Can do statements' and task types. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 65–92). New York, NY: Springer International Publishing.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Purpura, J. E., & Turner, C. E. (2014) *A learning-oriented assessment approach to understanding the complexities of classroom-based language assessment*. Teachers College, Columbia University Roundtable in Second Language Studies: Roundtable on Learning-Oriented Assessment in Language Classrooms and Large Scale Assessment Contexts. Teachers College, Columbia University, New York, NY. Retrieved from http://www.tc.columbia.edu/tccrisls/

Purpura, J. E., & Turner, C. E. (2015). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255–272). Boston, MA: De Gruyter Mouton.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1–20. https://doi.org/10.1177/026553229801500101

Surface, E., Poncheri, R., & Bhavsar, K. (2008). *Two studies investigating the reliability and validity of the English ACTFL OPIc with Korean test takers: The ACTFL OPIc validation project technical report*. Retrieved from http://www.languagetesting.com/wp-content/uploads/2013/08/ACTFL-OPIc-English-Validation-2008.pdf

Tigchelaar, M., Bowles, R., Winke, P., & Gass, S. (2017). Assessing the validity of ACTFL can-do statements for spoken proficiency: A Rasch analysis. *Foreign Language Annals, 50*(3), 379–403.

Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals, 49*(1), 75–92. https://doi.org/10.1111/flan.12178

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2014). Flawed self-assessment: Investigating self-and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*(1), 1–19. https://doi.org/10.1017/S1366728914000832

Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals, 49*, 201–223. https://doi.org/10.1111/flan.12198

Tschirner, E., Bärenfänger, O., & Wanner, I. (2012). *Assessing evidence of validity of assigning CEFR rating to the ACTFL oral proficiency interview (OPI) and oral proficiency interview by computer (OPIc).* (Technical Report 2012-US-PUB-1). Retrieved from Language Testing International: http://www.languagetesting.com/wp-content/uploads/2014/02/OPIc-CEFR-Study-Final-Report.pdf

Vande Berg, M., Connor-Linton, J., & Paige, J. M. (2009). The Georgetown Consortium Project: Interventions for student learning abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad, 18*, 1–75. Available at http://files.eric.ed.gov/fulltext/EJ883690.pdf

VanPatten, B., Trego, D., & Hopkins, W. (2015). In-class vs. online testing in university-level language courses: A research report. *Foreign Language Annals, 48*(1), 659–668. https://doi.org/10.1111/flan.12160

Magda Tigchelaar is an Assistant Professor of TESOL in the Department of Special Education and Literacy Studies at Western Michigan University. Her research interests include second language proficiency testing, self-assessment, and second language writing. Her current research centers on language learners' use of self-assessment materials for evaluating their own language proficiency and the development of language proficiency standards.

Where Am I? Where Am I Going, and How Do I Get There?: Increasing Learner Agency Through Large-Scale Self Assessment in Language Learning

Gabriela Sweet, Sara Mack, and Anna Olivero-Agney

Abstract This chapter explores the efficacy of Basic Outcomes Student Self Assessment (BOSSA), a fully integrated standardized second language self-assessment protocol. Designed for large-scale, sustainable use across languages, levels, and modalities, BOSSA supports learner awareness as a path to agency and empowerment.

BOSSA shifts the focus from the traditional teacher as center of knowledge (the only one who evaluates) to a learner-centered space where the students work in community to actively support and develop their language skills. The collaboratively created protocol was validated through piloting over several semesters, operationalizing self assessment at the University of Minnesota and transforming the language classroom experience for more than 10,000 students in ten languages.

Incorporating qualitative data from focus groups with students and instructors as well as quantitative data from student-reported benefit and self-assessment surveys, researchers found that a self-assessment protocol that pairs a proximal performance opportunity with training and practice with self assessment can successfully support learners, instructors, and language programs in large-scale contexts. In addition, it provides a workable response to the increasing calls for integrating research-driven

This research was supported in part by grants from the Center for Educational Innovation, University of Minnesota, and the Language Flagship Program Initiative of the National Security Education Program, U.S. Department of Defense.

G. Sweet (✉)
University of Minnesota, Minneapolis, MN, USA
e-mail: sweetga808@gmail.com

S. Mack
Department of Spanish & Portuguese Studies, University of Minnesota, Minneapolis, MN, USA
e-mail: mack@umn.edu

A. Olivero-Agney
University of Minnesota, Minneapolis, MN, USA

Italian Cultural Center, Minneapolis/St. Paul, MN, USA
e-mail: olive152@umn.edu

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_10

practice and transdisciplinary approaches as essential elements of second language teaching and learning.

Keywords Standardized learner-centered reflection protocol · Integrated performance tasks · Accuracy in self evaluation · Learner awareness · Transdisciplinary approach · Active learning · Empowerment · Cross-language applicability · Sustainable use

In higher education broadly, there are increasing calls to engage learners and improve student outcomes by integrating evidence-based instructional practices. Jankowski's report on behalf of the American Council on Education (2017) provides a list of five key areas where instruction and student outcomes intersect, but in which research-driven practices are not yet widely integrated. These areas include transparency, active learning, alignment, self-regulation, and more frequent assessment that occurs throughout the learning sequence, involving learners in the assessment process in an active way (pp. 6–7). At the same time, somewhat related field-specific discussions in Second Language Acquisition (SLA) have pointed to the need for adaptation and integration of evidence-based, transdisciplinary approaches in instructional practices. Notably, Atkinson et al. (2016), a working group of SLA scholars also known as the Douglas Fir Group, present a framework for transdisciplinarity in SLA, recognizing that learners in a multilingual world are best supported by a multidisciplinary theoretical and instructional approach; that is, one that addresses a wide range of questions, from the cognitive to the socioemotional, educational, and sociocultural (p. 39).

From this backdrop, some practical questions emerge: how do language educators begin to conceptualize teaching practice in this more complex instructional worldview, and how can we make the best use of limited resources to address these issues and integrate evidence-based practices? This chapter describes a collaborative project that situates self assessment as an approach to addressing many of these current concerns. The chapter begins by providing a background of self assessment and the Basic Outcomes Student Self Assessment (BOSSA) project, and continues with a detailed description of the components of BOSSA and its large-scale use to support language acquisition in higher education contexts. In keeping with the call to integrate and further develop evidence-based practice, we present data from several different aspects of the project, exploring issues of learner agency, awareness of the language learning process, and the levels of accuracy students reach in evaluating their language abilities. For the latter, we have collected data comparing student performance on American Council on the Teaching of Foreign Languages (ACTFL) tests with student self-assessed ratings. Some of these data presented here were collected through the Proficiency Assessment for Curricular Enhancement (PACE) project which began in Fall 2014 at the University of Minnesota, funded by a grant from The Language Flagship Program Initiative of the National Security Education Program, U.S. Department of Defense, and designed to maintain a culture of assessment (Vanpee & Soneson, this volume), self assessment, and curricular improvement

in the second language learning process, as well as a systematic program of proficiency assessment and professional development (Soneson & Tarone, this volume). Finally, we identify future directions for maximizing the benefit of self assessment for language learners in large-scale contexts, taking into account administrative challenges and examining the issue of long-term sustainability in terms of program accessibility, instructor training, and protocol delivery.

1 Literature Review

Research and interest in self assessment has grown exponentially in the past ten years. Since this type of alternative assessment began to be used in higher education contexts, the role of the instructor, at least in theory, has shifted from center of knowledge and power to facilitator or coach, with students taking a more active part in their own education (LeBlanc & Painchaud, 1985). This shift reflects an emphasis on learner autonomy and motivation, and the acknowledgment that they play a more important role than ever in learning (Fink, 2013). In addition, it has been clearly demonstrated that the use of self assessment as a pedagogical tool promotes both of these elements, in foreign and second (mostly TESL) language acquisition contexts (Lappin-Fortin & Rye, 2014; Masgoret & Gardner, 2003). The more invested students are in what they are learning, and the more they see that changes they make in learning practices will positively affect their ability to demonstrate what they are able to do, the more empowered students are to take charge of their learning experience.

Self-regulatory behavior (guided by students' beliefs about how they can control their learning) and increased awareness of the second language learning process are other demonstrated benefits of using self assessment in the second language classroom (Andrade & Valtcheva, 2009; Nielsen, 2014; Ziegler, 2014; Dolosic et al. 2016). Ziegler and Moeller (2012) document how regular use of self assessment in LinguaFolio®, a language-learner portfolio assessment tool, positively affects students' ability to self-regulate their learning behaviors. As Ziegler and Moeller note, students become more aware of what they can do as they document the steps they take in the learning process and make choices about which of their work selections they will submit for evaluation (p. 335). This approach, which includes support for repeated assessment over time, clearly demonstrates progress to all stakeholders.

Another important consideration for research on self assessment and second language acquisition is validity and reliability. In other words, to what extent are students' assessments of their ability to use the language correlated with a summative external measure? Work in this area shows a preponderance of data examining gains in proficiency or performance during and after study abroad, and shows that learners can become highly accurate self raters and are able to track their own progress over time (Stansfield, Gao, & Rivers, 2010; Dewey, Bown, & Eggett, 2012; Brown, Dewey, & Cox, 2014).

Taken together, the results of recent research in the above areas strongly suggest that self assessment is a logical response to Jankowski's (2017) call for increased support for learning, grounded in research-driven practice. Self assessment provides a path to transparency (clear communication and shared understanding of goals, criteria for evaluation, and indicators of success), lends itself to a pedagogical approach that allows for active involvement and engagement with course content and goals, provides student-centered assessment of learning, and supports learners to build reflection and self-regulation skills. In addition, these elements together present a transdisciplinary method that aligns with Atkinson et al. (2016).

Given the many documented benefits of self assessment, it behooves higher educational institutions to incorporate it in a systematic way into language programs. However, systematic integration on a large scale, across multiple levels and languages, adds additional aspects that have not been as thoroughly documented and analyzed. As an initial step for understanding more accurately how self assessment can work on a large scale, it is important to examine how large-scale use facilitates learner self-awareness, empowerment and autonomy, and to what extent self-assessment instruments and actual ratings correlate. It is these questions that we address.

2 Method

2.1 Background

The BOSSA protocol was created at the University of Minnesota as a collaborative initiative between the Department of Spanish & Portuguese Studies and the College of Liberal Arts Language Center. Originally developed for students in a fourth-semester Spanish course, it was quickly extended to students of French, German, Italian, and now also supports students of Arabic, Chinese, English, Hmong, Korean, Portuguese, and Russian. Altogether, BOSSA currently supports approximately 1300 learners each semester at a variety of instructional levels.

Instructors of the fourth-semester Spanish course in 2013 identified a need for a better understanding of students' abilities in speaking and writing in unrehearsed contexts, without access to electronic resources, at the beginning of the semester. Since the course in question serves 450–500 students in 18–20 sections per semester, large-scale adaptability of self assessment was one of the key goals of BOSSA from the outset. As the course is taught by 8–12 different instructors, and the instructor base varies from semester to semester, any solution for addressing this need had to be scalable for multiple sections; only in this way could the process successfully become standardized across multiple sections and instructors, and ensure it was manageable even for instructors assigned to the course a few days before the semester started.

To respond to this need, a series of language performance tasks was proposed, to be paired with self assessment. It was hypothesized that, by providing students with a proximal performance opportunity prior to self assessing, students would be better prepared and be able to self assess more accurately, as suggested by Butler and Lee (2006), who explored the efficacy of on-task self assessment among elementary school students learning English. The distinguishing feature of BOSSA is the integration of performance tasks and additional opportunities for reflection into a fully automated process, a series of actions or steps taken in order to achieve a particular end. In this way BOSSA is an efficient solution in the large-scale context, making self assessment meaningful to students through grounding in performance.

Furthermore, the early decision to create a standardized tool paved the way for subsequent use of the protocol in other languages and at other proficiency levels. The path from concept to pilot, and from pilot to full-scale use as described below, took place over several semesters and included a variety of methods to optimize the materials and validate the self-assessment measurement tools. This included an assessment of previous materials and case studies to inform the development of the self-assessment materials, linking with University, College of Liberal Arts, and Department student learning outcomes, as well as aligning course descriptors with national proficiency guidelines published by ACTFL (2012a).

2.2 Materials

The oral performance component is the Speaking Practice Task (SPT), a computer-delivered simulated oral proficiency interview that, like the ACTFL OPIc, targets the range of traits characteristic of specified oral proficiency levels. It provides students with a context, a simulated interlocutor (a conversation partner, Mai, who is a native speaker of the language the student is learning and who is studying English), and a communicative reason to use language (for example, you are writing a research paper on international college students and you decide to interview Mai to learn about students at her university). The SPT consists of three steps, each of which targets a different linguistic function and topic domain. Topics were framed so as to invite students to show a range of vocabulary. This was done by providing organizers within the task prompt to give students ideas for what to say related to the topic and to elicit speech at the sentence level. For example, when interviewing Mai about life at her university, learners were presented with a list of suggestions including programs of study, jobs, environment, and free time. Care was taken to avoid testing specific background knowledge or cultural knowledge. Students respond orally to prompts in the target language using headphones and microphones, and responses are recorded. There are now four SPTs, each calibrated to a different proficiency level, consisting of communicative tasks that vary by instructional level. For example, beginning students introduce themselves and talk about their likes and dislikes, while students at more advanced levels make explanations and provide solutions for situations that include an unexpected complication. Time

allotted for speaking on each task (or step) also varies according to target proficiency level, from one to three minutes.

The first SPT was designed as a classroom practice tool by a team of Spanish instructors and second language researchers in Spring 2013. The goal was to elicit speech at the Intermediate Mid oral proficiency level. This language performance activity, followed by students listening to their responses to prompts, provided a concrete referent for students to then rate how well they spoke, using criteria. Thus, the focus during development and piloting was on the instrument. Researchers were focused on evaluating the SPT's potential to provide students with a practice situation appropriate to the proficiency level they would be expected to demonstrate by the end of their course, rather than on the specific levels of oral proficiency students demonstrated.

Three parallel forms were developed targeting the Intermediate Mid level. Delivery used a web interface and Digital Language Lab (DiLL) software (developed by the Multimedia Learning Center at Northwestern University) to provide the audio and manage students' recordings. The three forms were piloted with 123 students of fourth-semester Spanish. Each form consisted of five tasks that elicited speech exhibiting the range of traits associated with the specific proficiency level and including probes for a higher level (e.g., personal description; plans for a future event; talk about events depicted in photographs [picture prompt]; talk about something that happened in the past; question-asking), and lasted 20 minutes. Students had 2 min to respond to each task. After completing the SPT, students self assessed how well they had completed each task.

Instructors and researchers reviewed students' recordings to determine the model's success in eliciting speech at the target level. A general analysis of the range of traits present showed that students' oral responses on the SPT could be analyzed using proficiency descriptors and that the sample was sufficient to be able to make a rating in a way that would likely be roughly comparable to a rating determined by the ACTFL OPI or OPIc.

In addition, a survey collected students' feedback to the design of the SPT, their comfort level with the experience, and their own performance. Overall, the response from students to the instrument was positive: 70% reported that they felt comfortable in a speaking assessment situation with the computer interface, and 80% said that they felt they were effective at accomplishing the tasks as described. Students' opinion about time allotted for speaking was almost unanimous; 98% felt there was enough time allotted for speaking.

From Fall 2013 through Spring 2014 instructors from the departments of French, German, Italian, and Spanish contributed to the review and refinement of the Intermediate Mid SPT while piloting the protocol in their classes. Fall 2014 marked the beginning of the implementation of BOSSA on a larger scale, adding a fifth language, Portuguese, and, through the development of SPTs at other proficiency levels, extending to four instructional levels. During the academic year 2015, instructors and researchers finalized development, resulting in the current version of the SPTs.

The self-assessment tool is the Build on Language Track (BoLT) questionnaire. It consists of criterion-referenced can-do statements aligned with specific student learning outcomes of University of Minnesota language programs and ACTFL proficiency level descriptors. The BOSSA statements diverge from the National Council of State Supervisors for Languages (NCSSFL)-ACTFL Can-Do Statements (ACTFL, 2012b) in prompting responses that situate students on a developmental trajectory. For example, one NCSSFL-ACTFL main indicator item at the Intermediate Mid level related to interpersonal speaking is "I can talk about my daily activities and personal preferences." Checkboxes follow, so that students can indicate content area, such as daily routine, interests and hobbies. The BOSSA parallel item for Intermediate Mid is "I can participate in conversations about my life and topics related to my world." Below the item appear content areas (for example, myself, my home, and my family; my studies; my daily activities or routine; etc.). For each content area, students respond using four-point Likert frequency scales (from "I can seldom do this, or I can't do it yet." to "I can do this almost always, or always."). This format is meant to promote awareness that although students may not be able to do this with consistency, at some point in the acquisition process, they will be able to do so. In addition, the BOSSA can-do items include content that instructors across languages deemed key for their level and course at the University of Minnesota, such as "I can make comparisons, such as what people are like in my hometown and where I live now, or the education system in this country and another one," or "I can clarify what I want to say even if I don't know a certain word or phrase, using strategies such as paraphrasing or describing."

The can-do statements are grouped into different sections: topic domain and pragmatics (contextualized language use integrated with linguistic functions) in instruments used for Novice High through Intermediate Mid learners. The self-assessment questionnaire targeting Advanced Low users includes a separate section isolating linguistic functions. BoLT self-assessment questionnaires are delivered online via Qualtrics, a software tool that collects and analyzes data, and has the capability to tabulate and deliver results through a web browser on-screen and via email.

Level cut scores for the BoLTs were set collaboratively by a team of University of Minnesota instructors from a variety of departments with experience teaching the levels targeted by each instrument. One was a certified ACTFL tester, and all instructors had familiarity with the ACTFL guidelines through yearly refresher training offered by the CLA Language Testing Program in conducting and rating the oral proficiency interviews that students typically complete at the end of the fourth semester. Several completed the weeklong ACTFL OPI rater training in either 2015 or 2016 or participated in an ACTFL proficiency workshop focused on written and oral proficiency in 2011. These level experts worked across languages using a modified standard-setting protocol, estimating how learners at the minimum proficiency level for given semester levels would respond to each item. Estimates were aggregated, averaged for each item, and then totaled to determine a range of scores related to ACTFL proficiency levels, from Novice High through Advanced Low (depending on the self-assessment instrument). Scale options for some items on instruments

targeting Novice High through Intermediate Mid were intentionally limited or capped to help students self assess realistically. For example, on the Intermediate Mid item for speaking noted earlier, "I can make comparisons, such as what people are like in my hometown and where I live now, or the education system in this country and another one", the highest option possible is the third level on the Likert scale "I can do this most of the time," as instructors estimated that students at this level could not do this all of the time. Thus, the development team informed the process of establishing a developmental hierarchy for linguistic functions related to particular content domains and contexts.

2.3 The BOSSA Protocol for Speaking

The BOSSA protocol for speaking consists of one or two 50-minute computer lab sessions (at mid-semester or at the beginning and end of the semester; exact schedule varies by program), automated feedback, and may also include online reflections completed throughout the learning sequence. All materials supporting the protocol are in English, allowing for consistency of use across languages. The BOSSA Session for Speaking comprises an articulated, guided sequence of six activities, integrated to provide learners the maximal benefit of self assessment.

First, students watch a short video (Regents of the University of Minnesota, 2016) that introduces them to self assessment and familiarizes them with criteria they will use later to evaluate their skills. Then students warm up in pairs with a short conversation activity, activating their second language schema. Next, students complete the SPT performance activity, after which they listen to their recordings and use a worksheet to reflect on how well they were able to accomplish the communicative language tasks. This experience of a proximal performance opportunity allows students to approach the question of "Where am I?" in a practical, concrete sense; they've just done the task and can assess it in a non-abstract, factual, evidence-focused way, situating their reflection in what just occurred. After working individually on the worksheet, pair work follows, giving additional opportunities for students to work together to process their thoughts about how language learning works (again, focusing on what they experienced and can notice from reflecting on the performance task they just completed). They also share notes about their strengths and challenges and set specific goals for improvement. Next, students lead a class discussion about proficiency in their own words while the instructor takes notes on the board; the board is photographed so that students have a record later. The instructor also facilitates the class discussion, as needed, to help students understand what is realistic in terms of proficiency goals per course learning outcomes, goals, and expectations. In this way, the performance task and subsequent discussion provide a structured and concrete base for reflecting together; learners engage with each other as part of the class community of practice, exploring within the context of group sociocultural norms, and processing together the questions "Where am I going?" and "How do I get there?"

In the final lab activity of a BOSSA Session for Speaking, students use the online self-assessment questionnaire to rate their speaking ability. This final step brings together all the practice and training in self assessment gained during the lab session: students have a specific idea of their skills in light of their actual speaking performance in the SPT, and they have new knowledge (from the discussion) that helps them assess those skills in terms of general language learning outcomes, goals specific to the course, and their own individual goals. Finally, after the students complete the online BoLT self-assessment questionnaire, they receive an automated email. The email includes a report of estimated proficiency level based on how students have self assessed. It also includes suggestions (collected from other language students like themselves) on how they can improve their language learning and information on other proficiency levels (not just the one at which self assessed).

2.4 Additional BOSSA Components

The BOSSA toolbox also includes the BOSSA Session for Writing, which has a similar structure to the Session for Speaking. It consists of a Writing Practice Task; a reflective worksheet where students evaluate their writing abilities first alone, and then with a partner; student-centered class discussion which promotes awareness of the writing process and what is realistic at varying levels of proficiency; and the Build on Language Track (BoLT). The BoLT for writing is a self-assessment questionnaire that focuses on the degree of support or resources writers need, and includes questions related to degree of detail students can provide to a variety of topic and linguistic function domains and attention to organizational aspects of writing.

Like the Speaking Practice Task, the Writing Practice Task provides a low-stakes opportunity for students to actively experience completing concrete course objectives: what they will be able to do in terms of communicative competence by the end of beginning, intermediate, and advanced courses. Students at each of the three Writing Practice Task levels are presented with several topic choices and clear expectations related to content, length, and organization as well as a timed period in which to complete the task. The Writing Practice Task is delivered online. The language input source can be adjusted to support students of non-roman alphabet languages, who also receive extra time to write.

Another component, the online Reflections, provides students with additional practice in self assessing. Programs may opt to use these customizable questionnaires/journals periodically throughout the semester, or just once at midpoint. Students can track their progress both related to specific course content (achievement) and for the larger proficiency lens; the Reflections allow students opportunities to look critically at their developing skills and thus become more familiar with the practice of self assessment. Students start by noting what they are currently working on, and specify in which skills they think they've made progress. Next, several can-do statements from the self-assessment instrument are recycled, to

support a focus on proficiency. Then students evaluate their language learning practices, keeping a record of what they do outside the classroom in support of their proficiency. Later in the Reflection, students focus on strengths and challenges, as well as progress toward the specific goals they set during the BOSSA Session for Speaking (or Writing).

Other BOSSA components were developed with support from Language Flagship funding for the Proficiency Assessment for Curricular Enhancement (PACE) project. Self assessment is a central element of the project, in actively engaging students in their own learning. Additional BOSSA components developed under the auspices of PACE include Build on Language Track Self Assessments for Reading and Listening. These online questionnaires consist of between eight and ten statements which, like the BoLT for Speaking described above, are aligned with specific student learning outcomes of University of Minnesota language programs and ACTFL proficiency level descriptors. Students respond to the statements again using Likert scales of frequency ("I can seldom do this, or I can't do it yet."–"I can do this almost always, or always.") or degree ("I can comprehend little/some/considerable detail").

The same procedure used to set level cut scores for the Speaking BoLT was followed for the Listening and Reading BoLTs. Each of these self-assessment questionnaires consists of three sections: students' listening and reading practices outside of class (what they choose to do on their own related to specific text types); strategies they use to help themselves comprehend written and auditory texts; and students' abilities given specific context, content, and linguistic function. For beginning learners, the section on practices is considered a warm-up and not scored. For intermediate and advanced learners, responses to items in the BoLT practices section are collected as part of the overall score for the instrument. Including those items underscores the importance of deliberate choices learners make about language use outside of class (agency) as a key factor in their growing proficiency (Duff, 2013).

2.5 Large-Scale Delivery

Initially, the SPT content was delivered as a web page, with Digital Language Lab (DiLL) software to deliver the audio and manage students' recordings. Instructors manually started and stopped students' recordings during the SPT and later made them accessible to students for the listening and reflection steps. They also had to direct students back to a web page to log into the self-assessment questionnaires at the end of the session. In other words, instructors play a key role in BOSSA, not only in communicating the importance of self assessment, but in the mechanical delivery of the session itself. Instructors are responsible for delivering the BOSSA Session for Speaking, and thus must understand how all components fit together to maximize the benefit of self assessment. Instructors must convey the instructional goals of the session, manage time so that all steps in the process are completed, and be able to control the technical infrastructure as well.

As the number of languages, courses, and sections implementing BOSSA grew, it became apparent there was a growing need for resources to support such a large-scale operation. First, developing a standardized training module for instructors was increasingly important. A specific training plan was created to assist departments to make sure that instructors had the training they needed so that the BOSSA session could be as stress-free and positive an experience as possible both for instructors and students. Second, a delivery method that would allow instructors to focus on pedagogical aspects of the process, with fewer technical responsibilities, was created. After usability sessions and piloting, delivery of the BOSSA materials using LiveCode software (1997) was adopted by many programs. It regulates the process and timing of the BOSSA articulated components through an integrated presentation of tasks and automatic recording and archival of students' responses to the SPT. The format provides a direct transition to the BOSSA self-assessment questionnaire through a link students access after class discussion. LiveCode generates students' recordings as mp3s and responses to the BOSSA reflective components (including class discussion) as text files; they are saved automatically to the student computer, allowing students to access the files immediately and preserve a copy for their own use (copy to a flash drive, save to a cloud service, send via email, etc.). Unlike DiLL, LiveCode does not archive students' recordings on a server for later access and review; once the student computers are reset at the end of the day, there is no copy of the recording saved on University-owned devices or servers.

2.6 Research Questions

To gain a richer understanding of how self assessment can effectively support second language learning, multiple methods were used to collect quantitative and qualitative data related to metacognitive awareness, accuracy, and agency. Three separate data streams were analyzed in order to respond to the following research questions:

1. Does the use of a proficiency-based self-assessment tool facilitate learner self-awareness in the second language classroom?
2. Does the use of a proficiency-based self-assessment tool facilitate learner empowerment and autonomy in the second language classroom?
3. Are there correlations between self-assessed ratings and ratings by an external measure?

2.7 Awareness as a Path to Agency: Research Questions 1 & 2

The first data stream is from two groups of learners in a fourth-semester Spanish course in Spring 2014 using the beginning (Round 1) and end of semester (Round 2) formats and formed the springboard for continued research.

All students completed a survey at the beginning of the semester, which asked about students' reasons for taking the class, their level of motivation to study Spanish, and collected demographic information. To determine the extent of the benefit of self assessment to learners (awareness and agency), a student self-efficacy survey after the BOSSA for Speaking Round 2 was administered to both groups. The independent variable was the use of the self-assessment questionnaire delivered in the language lab, and training in self assessment through periodic reflections. The test group consisted of 281 students who engaged in regular practice with self assessment via three reflections throughout the semester. The control group consisted of 86 students; these learners did not complete the self-assessment questionnaire or the reflections.

To answer the two research questions, items from the self-efficacy survey were grouped into data sets, or collections of related items. The first set of items focused on whether use of a proficiency-based self-assessment tool facilitates learner self-awareness in the second language classroom. Results of an independent samples *t*-test as shown in Table 1 below comparing means between the two groups responding using a five-point Likert scale to the question "I have a good idea of what my abilities are in this language (what I can DO with the language)" showed a statistically significant difference (p-value = .03; 95% CI [−.258, −.016]) between the test group (M = 4.05, SD = .49, N = 258) and control group (M = 3.91, SD = .48, N = 81). These results suggest that when students complete self-assessment activities to reflect on their language proficiency, increased awareness is one result.

The second set of items assessed whether use of the tool, in delineating clear expectations for student outcomes during the class and providing specific and timely feedback, empowered students to take a more active role in their acquisition of the target language. The key question, eliciting a simple Yes/No response, was "I made changes in my language learning (i.e. how I study, how I approach class and homework, what I do outside of class, etc.) as a result of doing these skills assessments." A majority of the control group reported not making changes (60.5% no *vs*. 39.5% yes; M = 1.58, SD = .49), while a majority of the test group reported that they did make changes in their language learning practices in response to using the protocol (58.1% yes *vs*. 41.9% no; M = 1.40, SD = .49). An N-1 two proportion test,

Table 1 Independent samples test showing benefit of self assessment: Learner self-awareness

Item text	Mean	SD	F	df	*t*-value	*p*-value	Mean difference	Std. Error difference
I have a good idea (test; n = 258)	4.05	.49	.48	337	−2.22	0.03*	−.14	.06
Equal variances assumed								
I have a good idea (control; n = 81)	3.91	.48						
Equal variances assumed								

Note: *Significant at the $p < 0.05$ level (2-tailed)

recommended by Campbell (2007), was conducted, comparing independent proportions for both small and large sample sizes. The test showed that the difference between the test and control groups (whether students changed their behavior as a result of using the self-assessment protocol) was statistically significant (p-value = 0.003).

Findings and analysis of the data gathered in the 2014 study (reported in Mack, Sweet, Olivero-Agney, Peltonen, & Rackowski, 2015) informed revisions to the BOSSA protocol and allowed for a widening of the lens to explore cross-language trends in self assessment, learner agency, and awareness. Analysis of this initial stage of the project, working with learners of one language and at one instructional level, was foundational for development and continued iterative trialing over two subsequent semesters and expansion to learners of eleven languages at seven instructional levels.

To determine the impact of self assessment on language learners' awareness and agency in a larger context, a second set of data from 1565 learners over three semesters (Fall 2015 through Fall 2016) of Arabic, Chinese, French, German, Hmong, Italian, Korean, Portuguese, Russian, and Spanish at multiple instructional levels was analyzed from the end-of-semester self-efficacy survey targeting students' attitudes toward using BOSSA tools to support language proficiency. This survey is administered only to students in those programs that opt for two BOSSA sessions each semester (at the beginning and end of the semester).

Students' mean responses to items on a five-point Likert scale (1 = Strongly disagree, through 5 = Strongly agree) suggest that trends noted in the 2014 study with one language at one instructional level are also applicable across languages and with students at a variety of levels. An analysis of the descriptive statistics from the multi-language group noted above shows that students valued the practice and training provided by the SPT and then reflecting (first alone, then in pairs and finally in the large group discussion) as helpful in preparing them to complete the final component of the session, the BoLT self-assessment questionnaire (mean = 3.75). Further, students connected increased awareness with having done the self-assessment activities ("I know what I can do", mean = 3.95; and "I could identify areas I need to work on", mean = 4.13).

A deeper look at students' awareness of the language learning process and their abilities is presented in Table 2 below. Analysis of the data in Table 2 shows the highest correlation between awareness of ability and identifying deficits (r = 0.55). Correlations among all the awareness items are moderate, and in each case there is a relationship of statistical significance.

Most importantly, the increased awareness noted above suggests that there is a path to agency as learners make a plan to address the gaps they noticed. An item on the self-efficacy survey targeting agency ("I made changes in my language learning") used a Likert scale of degree where 1 = No changes and 5 = A great number of changes. Aggregating responses 2 through 5 to determine the impact of self assessment on learner agency in this second stream shows that students reported making changes as a result of increased awareness stemming from the use of self

Table 2 Intercorrelations of student awareness

		I could identify areas I need to work on	Practice helped	I know what I can do
I could identify areas I need to work on	Pearson correlation	1.00	.45*	.55*
	Sig (2-tailed)		.00	.00
Practice helped	Pearson Correlation	.45*	1.00	.54*
	Sig (2-tailed)	.00		.00
I know what I can do	Pearson Correlation	.55*	.54*	1.00
	Sig (2-tailed)	.00	.00	

Note: *Correlation is significant at the $p < 0.01$ level (2-tailed)

assessment 92% of the time. This result (isolating negative responses while aggregating all positive responses) is similar to the 2014 study (in which students were given a binary choice).

2.8 Accuracy as a Measure of Awareness: Research Question 3

As shown above, learners report increased metacognitive awareness of the language learning process and of their speaking ability. The third research question further explores awareness, examining the levels of accuracy students can reach in self assessing in relation to a direct measure of language skills. Student performance data were collected for two semesters (Fall 2015 and Spring 2016) through American Council on the Teaching of Foreign Languages (ACTFL) language proficiency tests in speaking (the computerized oral proficiency test, or OPIc), reading (Reading Proficiency Test, or RPT), and listening (Listening Proficiency Test, or LPT). These tests were administered in selected second, fourth, sixth, and eighth semester classes of Arabic, French, German, Korean, Portuguese, Russian, and Spanish– classes who opted to participate in the PACE project at the University of Minnesota.

ACTFL proficiency ratings from the LPT, RPT, and OPIc were compared to students' self-assessed proficiency ratings to determine to what extent students' self evaluations of their skills match up with their rated performance. The ACTFL proficiency levels were equated with integers (e.g., 3 = Novice High, 4 = Intermediate Low, 5 = Intermediate Mid, 6 = Intermediate High). Data were analyzed first by aggregating per semester of instruction, and then by semester-level overall mean ACTFL-rated and self assessed (second-semester, fourth-semester, and sixth- through eighth-semester levels) and by-person ratings, for all of those for whom there were both ACTFL data and self-assessment data, as shown in Table 3. There were 58 students who received an ACTFL designation of Below Rating (BR) for reading and 65 who received the designation of BR for listening. (The BR designa-

Table 3 Levels of accuracy students reach in evaluating their language abilities (Fall 2015 & Spring 2016)

	ACTFL rating* (semester-level)	Self assessed* (semester-level)	Person-level Accuracy (at or within 1 sub-level)	Student N	Pearson Correlation	*p*-value	Effect size
Listening							
Semester 2	2.66	4.25	54%	126	.36	.00	.65
Semester 4	4.13	6.42	22%	218	.23	.00	.82
Semester 6–8	6.44	6.54	91%	149	.29	.00	.01
Mean listening	**4.41**	**5.74**	**51%**	**493**			
Reading							
Semester 2	3.32	4.25	75%	182	.34	.00	.39
Semester 4	5.22	6.41	56%	249	.28	.00	.42
Semester 6–8	6.80	6.44	89%	159	.35	.00	.11
Mean reading	**5.11**	**5.70**	**71%**	**590**			
Speaking							
Semester 2	4.01	3.96	93%	212	.37	.00	.00
Semester 4	4.88	4.62	92%	357	.18	.00	.07
Semester 6–8	6.09	6.07	91%	186	.61	.00	.00
Mean speaking	**4.99**	**4.88**	**92%**	**755**			

*Scale to compare ACTFL proficiency levels with BOSSA instruments proficiency levels: (1 = NL, 2 = NM, 3 = NH, 4 = IL, 5 = IM, 6 = IH, 7 = AL, 8 = AM, 9 = AH, 10 = S)

tion is used when student performance does not reach the lower limit of the range of ACTFL test used.) In addition, there were three students whose speaking performance was rated Unratable (U; used when the oral sample is too limited or obscured to be able to make a conclusive rating). These data are not included in Table 3.

Comparing means, learners provided with training and regular opportunities to rate their skills in the supported BOSSA session for Speaking (755 learners) self assessed more accurately (0.11 lower than ACTFL rated) than those who completed the self-assessment questionnaires as a stand-alone activity (for reading and listening). The margin of accuracy was wider for listening (493 learners, at 1.33) and reading (590 learners, at 0.59), with learners self assessing higher than they were rated for both modalities. Notwithstanding, these data suggest that awareness grows during

the trajectory of learning as each semester-level group self assesses their abilities progressively higher in each modality. This is similar to the trend documented by ACTFL ratings of abilities.

In addition, a test was run to calculate the Pearson product moment correlation between the criterion measure (ACTFL) and self-assessed ratings to determine the strength of the linear relationship between the two measures. Analysis of the data shows that the highest correlation of statistical significance is between the OPIc and the 6th–8th semester learners self-assessed speaking rating (.61). Interestingly, while all other correlations are weak to moderate, the lowest correlation (.18), also statistically significant, also falls under the category of oral skills, with fourth-semester learners significantly underestimating their abilities as compared to the criterion measure.

Looking at how individual students self assess their skills as compared to how they are rated by ACTFL, the data show a high degree of person-level accuracy for speaking at or within one sub-level on the proficiency scale (for example, self assessing at the Intermediate Low level and being rated Intermediate Low or Intermediate Mid) for all learners. There is less accuracy among individual students, in parallel to the semester-level average self- and ACTFL rating, for listening and reading. Fourth-semester learners in particular self assessed listening and reading abilities much higher than they were rated by the ACTFL standardized tests.

3 Discussion and Conclusions

The data streams, as shown in this project, provide documentation that large-scale self assessment successfully supports learners through facilitating self awareness, awareness of the language learning process, and learner agency. In addition, learners are able to evaluate their abilities with some accuracy; analysis suggests that the scaffolded support of the proximal performance opportunity (via the Speaking Practice Task) is an important factor in a realistic evaluation. Indeed, this unique element of BOSSA efficiently provides students with experience combined with the opportunity for reflection (Ash & Clayton, 2009), and this finding is logical given BOSSA's close alignment with other research-driven practices as outlined by Jankowski (2017, p. 6). Experience combined with the opportunity for reflection results in what Jankowski terms "deep learning" (p. 8), and is key in promoting learners as agents, and that combination is what BOSSA provides. Results over three years using the BOSSA protocol show a connection between students' awareness (what they can do with language in terms of communicative competence as well as the gaps they identify) and increasingly strong learner agency, as students make plans to address those self-perceived deficits. This moves students into the role of evaluator in relation to personal goals as well as specific course outcomes.

By self assessing and reflecting on how they learn, students are able to articulate the steps they will take to address their gaps and translate their reflections into actions. BOSSA is the point of departure for making changes based on a new awareness

students develop about their language learning process. This trend is consistent across a variety of languages and a variety of instructional levels, suggesting, from an evidence-based standpoint, that the large-scale application of this learning tool can effectively support language acquisition through promoting student engagement. Further, since program-level student learning outcomes form the building blocks for both the BOSSA performance opportunity and the self-assessment questionnaires, learners have a clearer understanding of what is expected of them (where they are going, by the end of the course) and have regular practice at measuring their abilities according to a course yardstick, using the criteria by which they will be assessed. Instructors also gain a clearer understanding of student learning outcomes in action. The integrated protocol generates evidence through a report that instructors can use to reflect on program objectives. The report includes aggregated results of the self-assessment questionnaire: a self-described language-learner profile comprised of challenges that students have identified.

BOSSA seems to encourage a paradigm shift at the class section level, one that aligns with notions of teaching and learning that place the learner at the center (Fink, 2013). This shift, from having the teacher as the center of knowledge (the only one who evaluates), to a learner-centered space, is initiated as students complete the performance task and then are guided as they began to understand how they themselves can consider their skills in an objective way, and then work in community to actively assess, support, and make a plan for developing their language skills.

As the protocol became increasingly integrated into the curriculum of language programs, extending to nearly 1300 students on average each semester, it became clear that a deeper exploration of scalable use was necessary. On a practical level, administration is now handled by a web application that manages BOSSA data, including scheduling BOSSA sessions for Speaking or Writing as well as instructor training sessions, making all tools accessible (self-assessment questionnaires, Reflections, worksheets, and instructions), and automating communications such as session reminders and requests. Also on a practical level, and as mentioned above, BOSSA materials were transferred to a LiveCode (1997) delivery option, with the goal of minimizing technical demands on instructors and streamlining student progression through the BOSSA materials. Instructors who choose this option report that it allows them to more easily manage the technical demands of a session.

From the research-driven practice perspective, we have conducted a mixed-methods analysis to examine differences between conducting the full-scale session of BOSSA in the computer lab during class time versus a mixed format in which learners complete some BOSSA elements outside of class (thereby freeing up computer lab resources). Overall, findings correspond to previous work documenting that self assessment in language learning increases learner agency and promotes awareness of the language learning process. However, the effect is stronger for learners who engage in the original 50-minute computer lab class session version of the BOSSA speaking protocol, and learners who receive training and opportunities to rate their skills in that format self assess more accurately than those who engage in the process in a mixed at-home and in-lab session. Full results are reported in Sweet, Mack, and Olivero-Agney (2017).

Over the years, the implementation of BOSSA has presented several challenges. First and foremost is the question of large-scale training for both students and instructors. From the student perspective, data from class discussions, focus groups, and comments from the final self-efficacy survey reveal that some students understand BOSSA as one of the many tests they take over the course of the semester, rather than a class activity. There is a need to continue to review communication with students about BOSSA protocol, refining the messaging and clarifying the different nature of the BOSSA lab session, where students start to take control of their learning and become agents. From the instructor standpoint, it's clear that they play a fundamental role in how the message is conveyed, serving as intermediary between BOSSA and students. To this end they need to have access to training opportunities and available tools both at the beginning of each semester and on-demand.

Another challenge reported by some students is talking about topics in unrehearsed and spontaneous situations. While the BOSSA protocol trains students to address unexpected communicative tasks, regular practice is essential to manage anxiety and the fear of making mistakes. As there is no grade associated with BOSSA and there are no consequences involved, the experience in the lab with the rest of the class provides a learning environment conducive to taking risks.

Future investigations should also look more closely at the impact of learner-specified changes in their practices on proficiency. Findings of this research point to increased learner agency in response to using self assessment, but don't explore specifically how learners put those changes into practice. In the words of one student, "It's what you choose to do outside the classroom that makes all the difference." Development of a stronger treatment of BOSSA's goal-setting component, along with more systematically integrated work on goals throughout the semester, would address this issue.

An additional challenge to address is the issue of self-efficacy for all learners. For example, through comments on the final survey, we also found out that some students don't perceive BOSSA as having a big impact on their learning process. They don't feel sufficiently competent in evaluating themselves and they pair the word "evaluation" exclusively with the instructor. Practice and training in self assessing can help students to overcome the lack of confidence in becoming self evaluators. This suggests that self assessment should start at early levels of instruction, allowing students to gain familiarity and practice and thus providing them with a fruitful experience.

A related area for future exploration is how sociocultural factors interact with self-efficacy and self-regulated learning overall, and if the opportunity of reflecting on a proximal performance task provided in the BOSSA protocol interacts with these factors, too. For example, as one reviewer of this chapter noted, it would be beneficial to take gender into account in the analysis of self-assessment accuracy. Might there be overestimation or underestimation of skills that covaries with self-reported gender? What other sociocultural factors might be relevant? As Grant and Zwier (2011) note, there is a clear need to conceptualize and analyze educational questions through multiple identity axes. As researchers and practitioners, we must acknowledge the role that intersectionality plays in student outcomes, and conduct

analyses that contribute to a better understanding of it. Furthermore, assuming that these differences in how learners experience self assessment exist, how can we use those predictive data to adapt BOSSA to better serve our diverse population of learners? For now, these considerations remain fruitful directions for future analyses of the data presented here.

Students may benefit from the proximal performance opportunity (provided in the BOSSA sessions for Writing and Speaking) if the self-assessment questionnaires for listening and reading were also paired with similar concrete performance tasks. The data from these questionnaires show that beginning and intermediate students consistently overestimated their abilities in these modalities as compared to ACTFL performance ratings, while they actually underestimate speaking ability. The grounding provided by the performance opportunity could help students to self assess more realistically for listening and reading. Efforts are underway to create a listening performance task using the BOSSA for Speaking model; programs would plug level-appropriate authentic listening passages (audio, video, or both) into a template that scaffolds activities to promote awareness of listening as a process.

It should be noted that any project of this scale necessarily entails reiteration, revision, and a long-term plan for adapting to the needs of the institution. Parallel to what we ask our learners to do via self assessment in language learning, we suggest that researchers and practitioners use data from proximal experiences to identify project strengths and weaknesses, set goals based on desired outcomes, and make a concrete plan to achieve future goals (and solve problems). Indeed, this process can be embraced as a path to providing the highest quality student experience, and is essential for long-term sustainability. Our experience to date has shown that a self-assessment protocol that pairs a proximal performance opportunity with training and practice with self assessment can successfully support learners, instructors, and language programs in large-scale contexts. In addition, it provides a workable response to the increasing calls for integrating research-driven practice and transdisciplinary approaches as essential elements of second language teaching and learning.

Acknowledgments The authors would like to thank the many instructors, students, and staff of the University of Minnesota who contributed to this project. We are also grateful to the reviewers of this chapter for their valuable comments and suggestions.

References

ACTFL. (2012a). *ACTFL proficiency guidelines 2012*. Retrieved February 28, 2017, from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

ACTFL. (2012b). *NCSSFL-ACTFL Can-Do statements: Progress indicators for language learners*. Alexandria, VA: Author. Retrieved April 22, 2017, from https://www.actfl.org/publications/guidelines-and-manuals/ncssfl-actfl-can-do-statements

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(1), 12–19.

Ash, S., & Clayton, P. (2009). Generating, deepening, and documenting learning: The power of critical reflection in applied learning. *Journal of Applied Learning in Higher Education, 1*(1), 25–48.

Atkinson, D., Byrnes, H., Doran, M., Duff, P., Ellis, N. C., Hall, J. K., … Norton, B. (2016). A transdisciplinary framework for SLA in a multilingual world. *The Modern Language Journal, 100*(S1), 19–47.

Brown, N. A., Dewey, D. P., & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals, 47*(2), 261–285.

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessments among Korean elementary school students studying English. *The Modern Language Journal, 90*(4), 506–518.

Campbell, I. (2007). Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine, 26*(19), 3661–3675.

Dewey, D., Bown, J., & Dennis, E. (2012). Japanese language proficiency, social networking, and language use during study abroad: Learners' perspectives. *Canadian Modern Language Review, 68*(2), 111–137.

Dolosic, H. N., Brantmeier, C., Strube, M., & Hogrebe, M. C. (2016). Living Language: Self-Assessment, Oral Production, and Domestic Immersion. *Foreign Language Annals, 49*(2), 302–316.

Duff, P. (2013). Identity, agency, and second language acquisition. In *The Routledge handbook of second language acquisition* (pp. 428–444). Routledge.

Fink, L. D. (2013). *Creating significant learning experiences: An integrated approach to designing college courses*. San Francisco, CA: Wiley.

Grant, C., & Zwier, E. (2011). Intersectionality and student outcomes: Sharpening the struggle against racism, sexism, classism, ableism, heterosexism, nationalism, and linguistic, religious, and geographical discrimination in teaching and learning. *Multicultural perspectives, 13*(4), 181–188.

Jankowski, N. A. (2017). Unpacking relationships: Instruction and student outcomes. *American Council on Education.* Retrieved February 3, 2017, from http://www.acenet.edu/news-room/Pages/Unpacking-Relationships-Instruction-and-Student-Outcomes.aspx

Lappin-Fortin, K., & Rye, B. J. (2014). The use of pre-/posttest and self-assessment tools in a French pronunciation course. *Foreign Language Annals, 47*(2), 300–320.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19*(4), 673–687.

Livecode Ltd. (1997). *LiveCode [Computer software]*. Edinburgh, UK.

Mack, S., Sweet, G., Olivero-Agney, A., Peltonen, J., & Rackowski, D. (2015). "*Yes, you can!": Self-assessment in the hybrid romance language classroom*. Presentation made at the Modern Language Association International Meeting, Vancouver, BC.

Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta–analysis of studies conducted by Gardner and associates. *Language learning, 53*(1), 123–163.

Multimedia Learning Center, Judd, A., & Marjorie. (2005–2008). *DiLL the digital language lab [Computer software]*. Evanston, IL: Weinberg College of Arts and Sciences at Northwestern University.

Nielsen, K. (2014). Self-assessment methods in writing instruction: A conceptual framework, successful practices and essential strategies. *Journal of Research in Reading, 37*(1), 1–16.

Regents of the University of Minnesota. (2016). *BOSSA: Welcome* [video]. Available at http://languagecenter.cla.umn.edu/boss/welcome.php.

Stansfield, C. W., Gao, J., & Rivers, W. P. (2010). A Concurrent Validity Study of Self-Assessments and the Federal Interagency Language Roundtable Oral Proficiency Interview. *Russian Language Journal/Русский язык, 60*, 299–315.

Sweet, G., Mack, S., & Olivero-Agney, A. (2017). Self assessment in language courses: Does in-class support make a difference? In I. Alexander & R. Poch (Eds.), *Innovative learning*

and teaching: Experiments across the disciplines (pp. 94–103). Minneapolis, Minnesota: University of Minnesota.
Ziegler, N. A. (2014). Fostering self-regulated learning through the European language portfolio: An embedded mixed methods study. *The Modern Language Journal, 98*(4), 921–936.
Ziegler, N. A., & Moeller, A. J. (2012). Increasing self-regulated learning through the LinguaFolio. *Foreign Language Annals, 45*(3), 330–348.

Gabriela Sweet works with the University of Minnesota College of Liberal Arts Career Readiness initiative on the new self-assessment model and tool, R.A.T.E. (Reflect, Articulate, Translate, Evaluate), used college-wide. Her research explores the connections between self-regulatory learning and performance, through increased metacognitive awareness and learner empowerment. She has worked in assessment development at the Center for Advanced Research on Language Acquisition, at Second Language Testing, Inc., and with the University of Minnesota Language Testing program.

Sara Mack Ph.D., is Senior Lecturer at the University of Minnesota, USA. Her work centers on research-driven approaches to second language acquisition in multi-section classroom contexts. Her research interests include multicultural inclusive and internationalized teaching and learning, sociophonetics, and learning and memory. She is also co-creator, with Gabriela Sweet, Anna Olivero-Agney, Joanne Peltonen, and Diane Rackowski, of the Basic Outcomes Student Self Assessment (BOSSA).

Anna Olivero-Agney has been a Teaching Specialist at the Department of French & Italian and Assessment Specialist at the Language Center of the College of Liberal Arts at the University of Minnesota. Her research interests include self assessment, theme-based instruction, and reading comprehension and strategies in a second language. She has worked in assessment development at the University of Minnesota.

Arabic Proficiency Improvement Through a Culture of Assessment

Katrien Vanpee and Dan Soneson

Abstract In this chapter, we demonstrate how systematic implementation of proficiency testing, student self-assessment, and instructor professional development contributed to large gains in student proficiency in Arabic within two years. ACTFL assessments of speaking and reading in Arabic at three levels of the curriculum conducted over the course of two years showed dramatic improvements at each level between testing sessions. Results are interpreted within the context of systemic changes introduced in a post-secondary Arabic Program at all levels. These changes included

- Incorporating external, proficiency-based assessments at all program levels and creating a culture of student resilience around proficiency testing to establish a culture of assessment;
- Supporting this assessment culture with student self-assessment and reflection on the learning process;
- Establishing and maintaining a culture of continual instructor professional development and teamwork to reinforce and support student proficiency development;
- Close collaboration with support units dedicated to excellence in foreign language teaching, such as the Language Center.

Keywords Proficiency · Assessment · Arabic · Resilience · Self-assessment · Collaboration · Curriculum design

K. Vanpee (✉)
Department of Asian Languages & Literatures, University of Minnesota, Minneapolis, MN, USA
e-mail: kvanpee@umn.edu

D. Soneson
University of Minnesota Language Center, Minneapolis, MN, USA

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_11

1 Introduction

With its solicitation of proposals for the Proficiency Initiative in 2014, the Language Flagship sought partnerships with established language programs to instill a culture of assessment in an effort to improve learner proficiency: "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" (Flagship, 2014). The University of Minnesota is a participant in this initiative, using proficiency tests to establish a culture of assessment among language programs in order to promote the development of language proficiency. The Proficiency Assessment for Curricular Enhancement (PACE) project includes seven language programs: Arabic, French, German, Korean, Portuguese, Russian, and Spanish. This chapter presents the experience of the University of Minnesota Arabic program with the PACE project and outlines how the implementation of proficiency assessment has both supported curricular enhancement and documented its effect. In Fall 2014 the Arabic program began the process of curricular change, with a renewed emphasis on language proficiency as well as cultural awareness. Test results of Arabic learners participating in the PACE proficiency assessments in Spring 2015 served as a baseline of overall proficiency and provided a basis for comparison as the curriculum developed. The results of Arabic testing over the first two years of the grant show a marked increase in proficiency levels at all stages of the curriculum. In this chapter, we present these results and discuss specific innovations in the Arabic program that may have contributed to them.

2 Introduction of Program and Reform

Since 2009, the Arabic Language Program at the University of Minnesota has been housed in the Department of Asian Languages and Literatures (ALL), where it is one of six language programs: Arabic, Chinese, Hindi/Urdu, Hmong, Japanese, and Korean. The Arabic program contributes to the overall mission of ALL, which offers courses on the literatures and cultures of the Arab world, in addition to its courses in East and South Asian cultures and literatures. Learners of Arabic can earn a minor or major in the department, with Arabic as their focus. Language programs in the department are relatively independent from each other and are managed by a Director of Language Instruction (DLI) at the rank of Lecturer. Language courses are taught by a cohort of full- or part-time Lecturers and Teaching Specialists. In the Arabic program, four full-time teaching staff work with roughly 170 students each fall and 130–140 students each spring.

Pursuant to recommendations of an external review of the Arabic language program undertaken in 2011, ALL conducted a search for a DLI for Arabic. The first

author of this chapter, Vanpee, joined the department as Arabic DLI in Fall 2014. Since this change, the Arabic program has been undergoing significant restructuring. In 2014–2015, an entirely new instructional team, consisting of three full-time instructors in addition to Vanpee, was brought on board. Mainstream, proficiency-oriented textbooks were adopted and a large amount of curricular materials created; new placement tests were developed; regular cultural programming was established; and class contact hours and course credits were added for the advanced level. The team implemented new instructional methods to promote a challenging and interactive learning environment. They also designed and implemented new courses, including a two-semester sequence of Egyptian Colloquial Arabic that was offered for the first time in 2015–2016. The Beginner and Intermediate-level courses involve five contact hours, the Advanced course four, and the Egyptian Colloquial course three hours per week. A three-credit post-advanced Arabic course was launched in Fall 2017, and a Jordanian Colloquial Arabic course will be offered for the first time in Spring 2019. At the time of writing, the Arabic language program offers five sections of Beginner Modern Standard Arabic (MSA) I and II; three sections of Intermediate MSA I and II; one section of Advanced MSA I and II; one section of post-advanced MSA; and Egyptian Colloquial Arabic I.

Students in the College of Liberal Arts (CLA) at the University of Minnesota must complete the fourth-semester course of a language or demonstrate proficiency in all four modalities at the intermediate level in order to fulfill the college graduation requirement. Prior to 2014, college proficiency expectations for a language such as Arabic corresponded to ACTFL Intermediate Low for speaking and writing and Intermediate Mid for listening and reading. Beyond this level, which is required for graduation from CLA, proficiency expectations had not been established for other levels of the curriculum.

At the outset of the restructuring process begun in Fall 2014, concrete proficiency target levels were established for each year in the program sequence. Table 1 lists these expectations.

The target levels for the first two years correspond to those of students learning "commonly taught languages" in the college, such as French, German, and Spanish. For the third year, Arabic target levels are as high as expectations for majors in the more commonly taught languages.

Table 1 Proficiency target levels for students of Arabic for years one through three

Course Nr	Semester	ACTFL proficiency target
ARAB 1102:	2nd semester	Intermediate Low
ARAB 3102:	4th semester	Intermediate High
ARAB 5102:	6th semester	Advanced Low to Advanced Mid

3 Proficiency Assessments for Arabic

As participants in the PACE project, beginning in Spring 2015 Arabic students have completed ACTFL proficiency tests at the end of each academic year. In Spring 2015, six of eight Arabic sections were selected to participate in PACE testing: two sections of Beginner Arabic, all three sections of Intermediate, and one section of Advanced, for a total of 105 students. Simultaneously, all eight sections of Arabic participated in the BOSSA speaking self-assessment protocol (see Chapter "Where Am I? Where Am I Going, and How Do I Get There?: Increasing Learner Agency Through Large-Scale Self Assessment in Language Learning", this volume). The second-semester sections took this protocol twice each Spring semester, while the second- and third-year classes participated in one session each Fall and each Spring. For PACE testing, instruments used in 2015 were the ACTFL OPIc for speaking and computer-adaptive reading and listening proficiency tests designed by Brigham Young University (BYU), similar to those described by Clifford and Cox (Clifford & Cox, 2013; Cox & Clifford, 2014).

The richness of the results produced by the 2015 PACE tests and the availability of funding for a second year of testing led to the decision to include all sections of Arabic in testing the following year. In Spring 2016, all 121 students enrolled in all nine sections of Arabic were tested: five sections of second semester, three of fourth semester, and one sixth-semester section. The program continued to use the ACTFL OPIc for speaking and began using a newly available testing instrument for reading proficiency designed by ACTFL, the Reading Proficiency Test (RPT). Because the preferred BYU computer-adaptive listening proficiency test was unavailable in 2016, Arabic students were not tested for listening that year. In 2015 and 2016, Arabic students took the proficiency tests in the second and third weeks of April, i.e. twelve to thirteen weeks into the semester.

4 ACTFL Proficiency Ratings

Students receive a rating for each test they take, based on the ACTFL scale (ACTFL, 2012). In compiling the results of the testing, ratings were converted into numbers in order to determine the mean level of proficiency for all students in each course. Individual sections of each course were combined to reach an aggregate mean for all students enrolled in the curriculum at each level. The numeric scale was weighted based on the large difference between the Mid sublevel and the High sublevel. In this process we are following the model of the Center for Applied Linguistics (CAL), who compiled the results of the first year of the PACE project (2015). Table 2 presents the scale used to calculate mean scores.

In addition to Arabic, the PACE project included testing of students at the end of the first, second, third, and fourth year of the curriculum for all students in representative sections of courses at these stages in seven languages: Arabic, French, German, Korean, Portuguese, Russian, and Spanish. As a means for comparison,

Table 2 Numeric conversions of ACTFL proficiency ratings

ACTFL rating	NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Numeric equivalent	0.1	0.3	0.8	1.1	1.3	1.8	2.1	2.3	2.8	3.0

Table 3 Aggregate proficiency results for all language programs 2014–2016

	Listening	Reading	Speaking	N
Year 1	NH (0.67)	NH (0.88)	IL (1.08)	321
Year 2	IL (1.07)	IM (1.36)	IM (1.28)	654
Year 3	IH (1.80)	AL (1.98)	IH (1.60)	246
Year 4	AL (2.04)	AL (2.15)	IH/AL (1.97)	148

Table 3 presents average ratings for all language programs in each modality, aggregated from the outset of the grant period in 2014 through Fall 2016.

5 Arabic Proficiency Ratings

Figures 1 and 2 show the results of the proficiency tests for Arabic over two years. Figure 1 shows speaking proficiency for each course and year of testing as measured by the OPIc, and Fig. 2 shows reading proficiency for each course and year. As mentioned, in 2015 reading proficiency was measured by performance on the ACTFL/BYU reading proficiency test, and in 2016 reading proficiency was measured by performance on the ACTFL Reading Proficiency Test. For each course and year, the figures illustrate the number of students in the course rated at each proficiency level, the average rating for each course as measured by the scale above, and the number of students tested (N).

For the sake of completeness, listening proficiency ratings for tests conducted only in Spring 2015, as measured by the ACTFL/BYU listening proficiency test, are presented in Fig. 3.

Analysis of the Spring 2015 test results in Arabic indicates that after two semesters (1102), students who started with Arabic in Fall 2014 when new instructional and assessment methods and a new curriculum were introduced, generally outperformed those in fourth-semester Arabic (3102) on the speaking test. On the reading test, their average performance closely approximated the results of fourth-semester students. Average results in speaking for students in 1102 were roughly equivalent to average results for students testing in second-semester French, Russian, and Spanish at the University of Minnesota. Reading results were lower than the average, while listening results were above the average for all four languages (see Table 3). In comparison with second-semester students in French and Spanish in a national study involving reading and listening, Arabic 1102 students testing in 2015 scored below the average for reading (NH for French and Spanish) and above average for listening (NM for French and Spanish) (Tschirner, 2016).

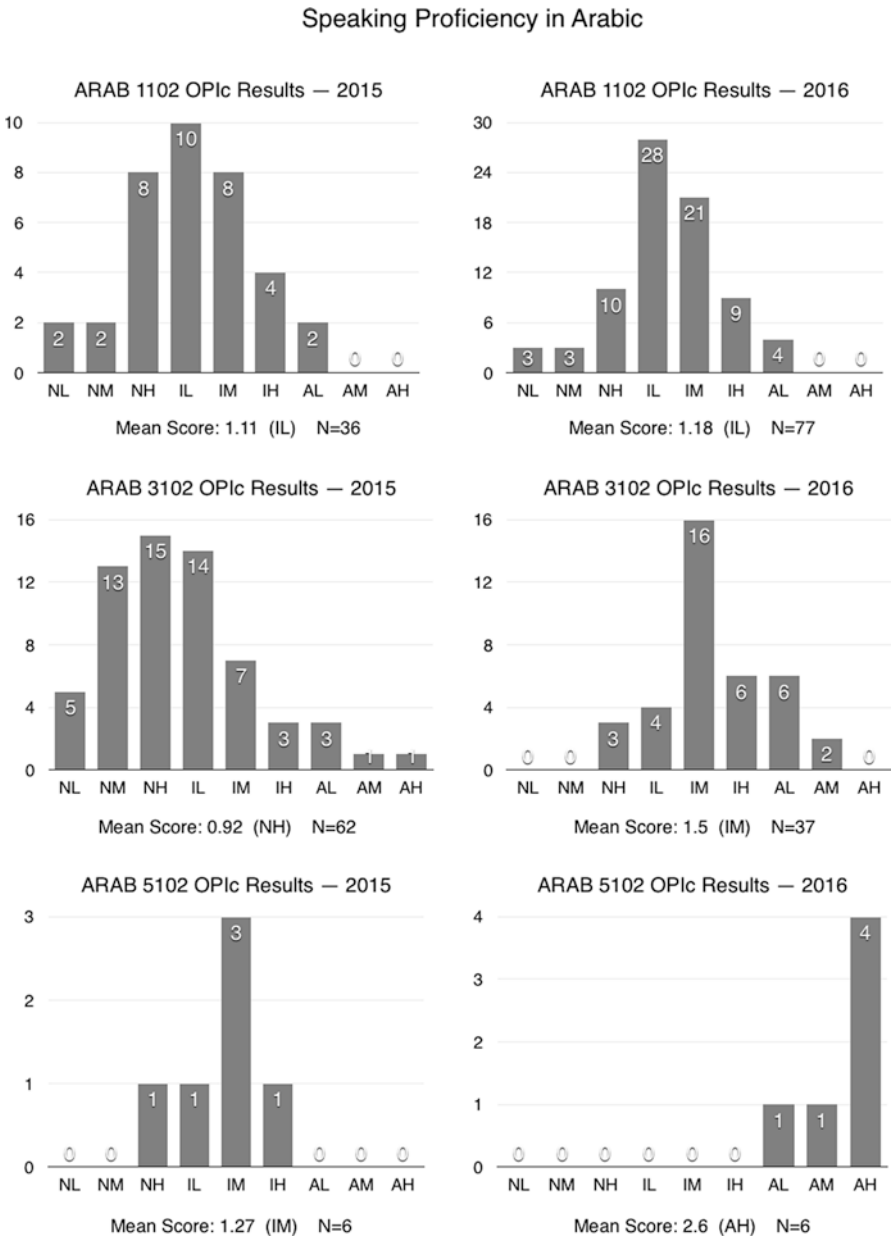


Fig. 1 Speaking proficiency ratings by course and test year

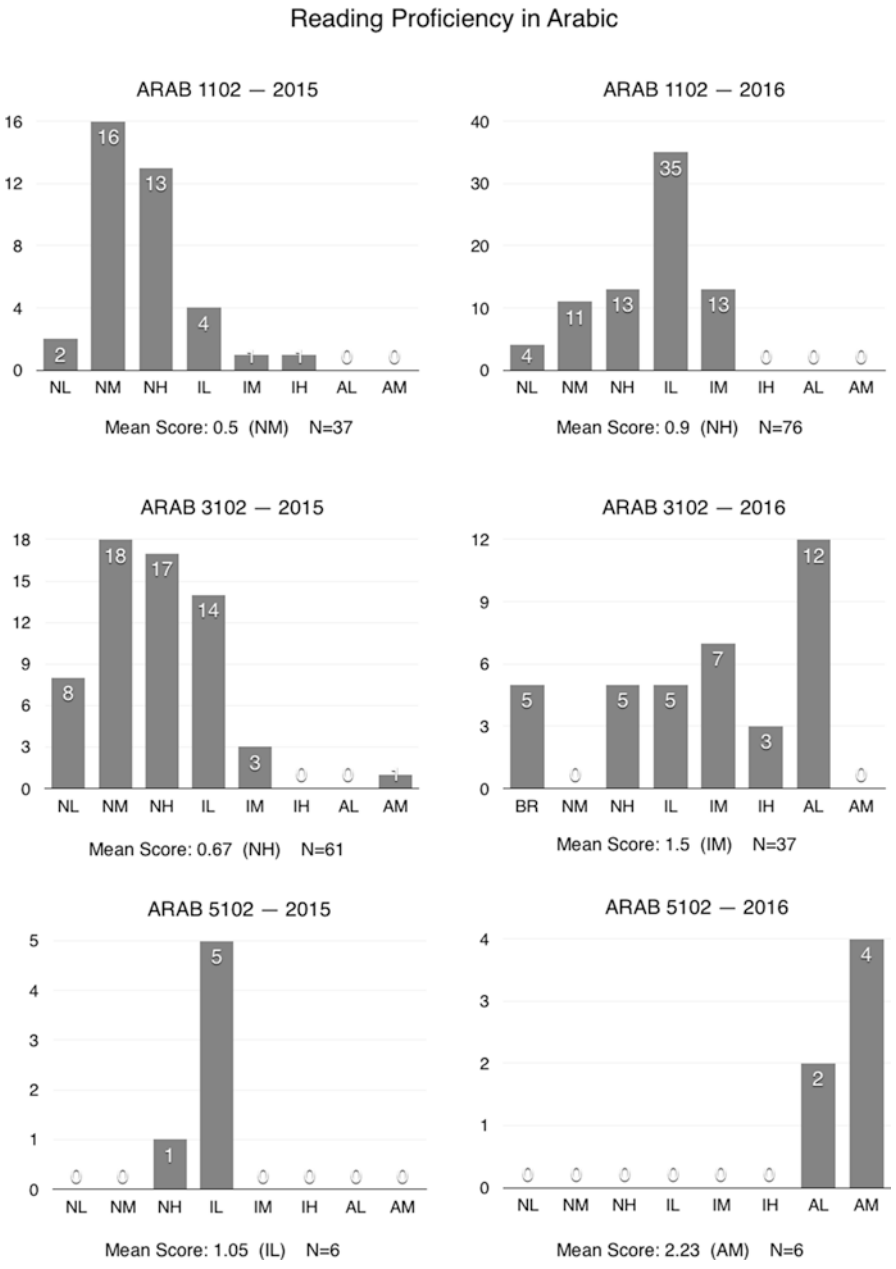


Fig. 2 Reading proficiency ratings by course and test year

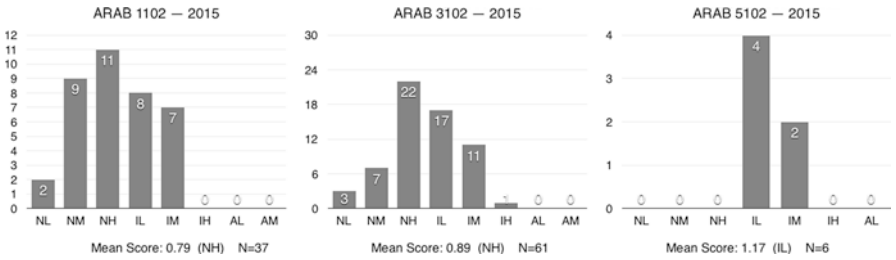


Fig. 3 Listening proficiency ratings by course, Spring 2015

Average performance of fourth-semester students (3102) in 2015 hovered around the Novice High level in all three modalities. Students were scoring three levels below the newly established target level (Intermediate High), and well below average reading ratings for all other languages tested in the PACE project after four semesters (French, German, Portuguese, Russian, Spanish). A similar discrepancy between target level and average performance can be observed for the sixth-semester students, whose average performance at the Intermediate Low level in all three modalities lagged behind the target level of Advanced Low to Advanced Mid by three to four levels, and well below average reading proficiency after three years in all PACE languages (Table 3).

In its report on the Spring 2015 test results of this project, CAL concludes that "there was no demonstration of marked improvement between Beginning and Intermediate level students" of Arabic (CAL, 2015, p. 57). This assessment underscores that students who had taken two semesters following the new curriculum could largely perform the same tasks in Arabic as their fourth-semester peers.

The revised curriculum extended to Intermediate and Advanced courses (ARAB 3101-2, ARAB 5101-2) during the second year of the project. For the second round of testing in Spring 2016, it was hoped that speaking results for first year would remain stable and reading and listening would improve, while those students who had proceeded to second year would reach higher test results overall after experiencing the revised curriculum for four semesters.

Figure 2 demonstrates that average second-semester (1102) reading proficiency scores improved from the Novice Mid level in 2015 to Novice High in 2016, verging on the Intermediate Low level. While the average rating is closer to Novice High than to Intermediate Low, 48 of 76, or 63% of the students were rated at Intermediate Low or above. This improvement brings Arabic in line with national averages for French and Spanish (Tschirner, 2016, p. 212) and slightly above the average for all second semester students tested at the University of Minnesota (Table 3). A stronger increase in reading proficiency scores can be observed for the fourth-semester students in 3102, whose average performance improved from Novice High in Spring 2015 to Intermediate Mid in Spring 2016, with 23 of 37, or 62%, rated Intermediate Mid or above. This improvement brings the overall fourth-semester reading proficiency score closer to the Intermediate High target level, with 15 of 37, or 41%, rated at Intermediate High or above. The average score in reading of 1.5 for second-

year Arabic students (3102) places them slightly above the University of Minnesota average for all languages (Table 3), and one sublevel higher than the national average for French and Spanish, listed as IL+ and IL respectively (Tschirner, 2016, p. 212). Also, for the sixth-semester students a significant improvement in reading proficiency scores can be observed, with average reading scores increasing by four levels from the Intermediate Low to the Advanced Mid level.

Figure 1 shows that on the speaking assessment, second-semester students tested in Spring 2016 on average continued to meet the Intermediate Low target level. This suggests that the improvement in the Spring 2015 speaking results may be indicative of a budding trend, rather than being a one-time lucky strike. Simultaneously, average fourth-semester speaking results improved by two sub-levels, from Novice High/Intermediate Low to Intermediate Mid/Intermediate High. A good number of the fourth-semester students (14 of 37, or 38%) met the established target proficiency level of Intermediate High, with 16 more rated at Intermediate Mid, making 81% rated at Intermediate Mid or above. Average reading proficiency for fourth-semester Arabic students in 2016 surpassed the 1.28 average for all languages tested at the University at the end of two years (Table 3). Finally, average sixth-semester speaking results improved by three levels, jumping from Intermediate Mid to Advanced High, meeting the established target proficiency level.

While in 2015, overall second-semester (1102) reading proficiency scores lagged behind those for speaking, one year later reading scores have narrowed the gap to speaking scores at Intermediate Low. Although average reading scores for 1102 students in 2016 were in the Novice High range, by the end of the fourth semester in 2016, average reading scores are equal to average speaking scores, at Intermediate Mid. Sixth-semester students (5102) in 2016 average in the Advanced Mid or above range for both modalities.

6 Individual Student Progress

A small number of Arabic students participated in testing in both 2015 and 2016. Fourteen students who took the assessments while in second semester in 2015 were also tested while in fourth semester in 2016. Table 4 shows their respective ratings for each modality from 2015 and 2016.

Among this group of students, ten out of fourteen (= 71%) second-semester test-takers either met or exceeded the speaking proficiency target level (IL) in Spring 2015. While only four students met the higher speaking proficiency target level for fourth semester (IH) in 2016, improvement between second and fourth semester is observed for ten students by at least one sub-level. Three students improved by two to three proficiency levels, yet three students remained at the same speaking proficiency level between second and fourth semester, in the Intermediate range, while one student's rating went down one sublevel.

For reading, only three second-semester students met or exceeded the ARAB 1102 target level of Intermediate Low in 2015. In Spring 2016, one year later, five

Table 4 Individual student proficiency ratings in speaking and reading in 2015 and in 2016

	Speaking		Reading	
Student number	2015 1102 Sp	2016 3102 Sp	2015 1102 R	2016 3102 R
1	NM	NH	NM	IM
2	IM	IH	IM	IH
3	IL	IM	NH	IL
4	IH	IM	NH	AL
5	IM	IM	NH	IM
6	IL	IM	NH	BR
7	IH	AL	NM	IL
8	NM	NH	NM	BR
9	NL	IM	IL	AL
10	IM	IH	NM	AL
11	IL	IH	NM	IL
12	IM	IM	NH	IH
13	NH	IM	IL	BR
14	IL	IL	NM	NH

of fourteen students met or exceeded the fourth-semester target level of Intermediate High. Two students received "Below Rating" (BR) scores in the reading test in 2016, meaning they did not clearly demonstrate reading proficiency of NH or above, but no definitive rating could be given. The relatively low attainment rates of the fourth-semester target levels for speaking as well as reading underscore the ambitious nature of the target levels. That said, more than 50% of the students in this group improved multiple sub-levels between second and fourth semester for reading, with two students improving by three sub-levels, and individual students improving by four or even five sub-levels in their reading proficiency.

A comparison of the Arabic PACE test results, particularly in 2016, with the results of University of Minnesota students testing in other foreign languages and with the nationwide results in reading reported by Tschirner in 2016, underscores the positive nature of the overall outcome for Arabic. These results demonstrate that students can develop levels of proficiency in Arabic that are equivalent to the expectations for students in other, more commonly taught languages in a similar amount of time. This should underscore that perhaps the Foreign Service Institute ranking of Arabic as a Tier V language can be reconsidered.

7 Self-Assessment

Beginning in Spring 2015, the Arabic program has implemented the Basic Outcomes Student Self Assessment (BOSSA) protocol for speaking along with the PACE proficiency assessments. The BOSSA protocol, originally developed by a team from the Department of Spanish and Portuguese and the University of Minnesota

Language Center, provides an opportunity for students to complete speaking tasks and then reflect on and discuss their performance. This is followed by the completion of an online self-assessment questionnaire (see Sweet, Mack, & Olivero-Agney, this volume). Starting from third semester, students take the self-assessment protocol once each semester of the course sequence. They are asked to complete three speaking tasks, to listen to a voice-recording of their performance on these tasks and to assess their own performance. Based on their responses to the self-assessment questionnaire, they receive a message that identifies the proficiency level that corresponds most closely with their responses on the questionnaire. The program also sends them an email containing that same message. Accompanying their score, students receive an outline of the characteristics of the corresponding ACTFL proficiency level and some guidance on how to continue developing proficiency in speaking. Instructors receive an aggregated report for their course that outlines at what levels students assess their speaking proficiency. In the following, we present portions of these aggregated reports for third-, fourth- and fifth-semester students, discussing the percentage of students who roughly assess themselves at each proficiency level, and comparing these percentages from year to year. For third-, fourth-, and fifth-semester students, each year a greater percentage of students rate themselves closer to the proficiency goals for the course. Figure 4 shows responses for third-semester students in Fall 2015 and Fall 2016:

While the percentage of third-semester students who self-assessed at the Intermediate Low level for speaking remained more or less the same, an increase of 14.8% is observed in the number of third-semester students who self-assess their speaking skills at the Intermediate Mid level in Fall 2016.

For fourth-semester students, self-assessment data are available for three consecutive Spring semesters. Figure 5 illustrates percentages self-assessing at each ACTFL level.

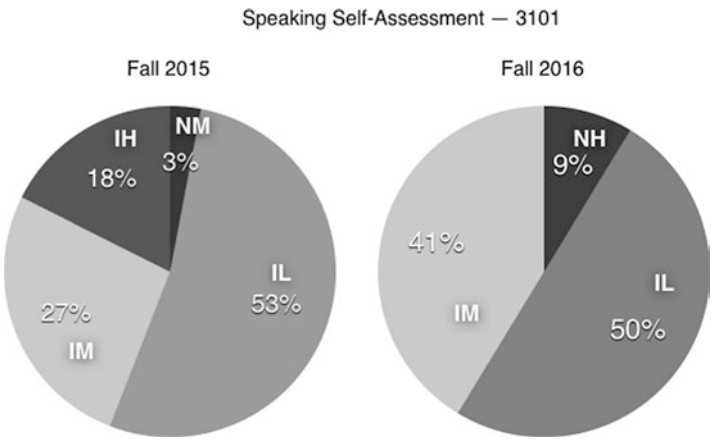


Fig. 4 Percentage of students self-assessing at each proficiency level for third-semester Arabic

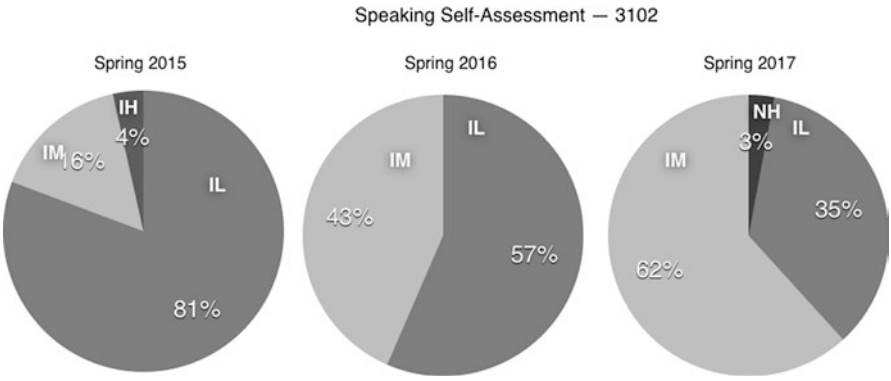


Fig. 5 Percentage of students self-assessing at each proficiency level for fourth-semester Arabic

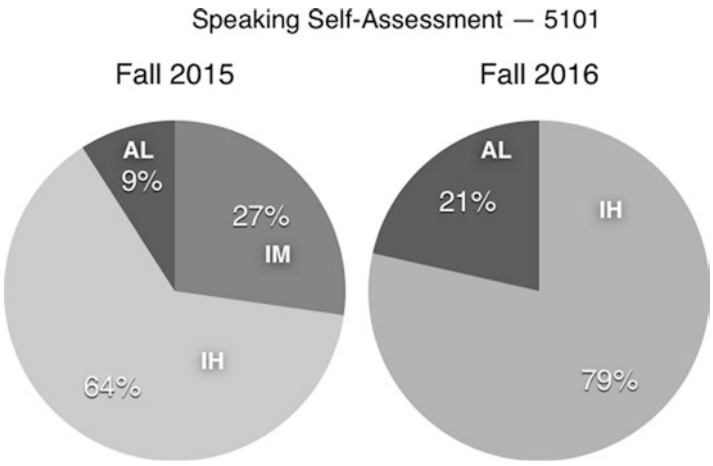


Fig. 6 Percentage of students self-assessing at each proficiency level for fifth-semester Arabic

Between Spring 2015 and Spring 2017, the number of fourth-semester students who self-assess their speaking skills at the Intermediate Low level steadily decreases, mirroring a significant increase in the number of students who self-assess their speaking skills at the Intermediate Mid level. This increase amounts to 46% of all fourth-semester students in a two-year time span. Whatever the accuracy of students' self-assessments may be, this palpable increase in speaking self-assessment level for the third- and fourth-semester students suggests a rise in confidence among those students about their Arabic speaking skills.

Finally, the speaking self-assessment results for fifth-semester Arabic students show a similar increase in the percentage of students self-assessing at higher levels. Figure 6 illustrates percentages self-assessing at each ACTFL level.

As was the case for the third- and fourth-semester students, fifth-semester students increasingly rate their speaking skills higher. An increase of 15% in self-assessments at the Intermediate High level can be observed, in addition to an increase of 12.3% of self-assessments at the Advanced Low level. No fifth-semester students rated their speaking skills at the Intermediate Mid level or below in Fall 2016. These responses may indicate an increase in upper-level student confidence in speaking.

8 Revising the Curriculum

How can we explain this dramatic improvement in a single academic year? In the following section, we discuss in some detail steps taken by the Arabic language program that may have contributed most to the improvement in student proficiency and self-assessment results.

8.1 Raising Expectations

Key to the success of the new program was the establishment of raised expectations not just for the students in the Arabic program, but also for instructors. The introduction of a new textbook for first-year classes, and one year later for second-year classes, as well as the development of entirely new course materials for third-year Arabic aided in establishing new expectations for student learning and performance in class and at home throughout the semester. First- and second-year Arabic adopted the *al-Kitaab* textbooks—a mainstream textbook series that is widely used across academic Arabic programs in the United States. The expressed aim of this series is to enable a large number of students to "reach solid intermediate-high proficiency using the ACTFL proficiency guidelines in all skills by the end of second year (after approximately four college semesters)" (Brustad, al-Batal, & al-Tonsi, 2013 xxiii). The widespread use of this textbook also ensured that students would share similar experiences with their colleagues in study abroad programs and facilitated adjustment to their study abroad courses.

Raising expectations of students also involved constructing the course syllabi for all levels of Arabic to create a climate of accountability around class attendance, participation, and homework submission. Syllabi now included a schedule of homework and instructor feedback on a daily rather than weekly basis, along with clear explanations of how tardiness and excessive class absences would affect students' course grades. Instructors received training in the use of the online course management platform Moodle where they were to register students' attendance, participation and homework scores. The purpose of this system was to reinforce continuous and systematic student work, as well as to create transparency between instructors and students around course requirements and the computation of students' final

grades. By the time the Arabic students took the PACE ACTFL proficiency assessments for the second time, all students were enrolled in an Arabic class in which they were held accountable for their attendance and participation in class, their submission of daily homework, and their preparation for tests and projects.

Raising expectations of instructors involved providing opportunities for them to improve their professional skill set, to grow as language educators, and to clearly understand standards for instruction by means of team meetings, professional development workshops, and class observations. Class observations were followed by detailed written feedback in which the strengths and working points of the instructor's performance were articulated. Instructors hired in 2014 and 2015 responded with enthusiasm to the established standards for instructor and student performance, to the opportunity of playing a vital role in a major curriculum revision, and to the prospect of strong student performance and proficiency results. Instructor buy-in and support has proven essential to bring about change in classroom management and student performance, and for a climate of accountability to develop.

Finally, raising expectations also involved implementing a system whereby all classes in the first four semesters physically meet in an interactive, communicative, student-centered instructor-led session. Rather than replacing one weekly contact hour with an online language program, the program held fast to five regular class meetings per week, deeming contact hours in which students use the language for interactive communication under the guidance of an instructor to be critical for promoting proficiency. To support this concept and to promote regular class attendance and participation, one Arabic course section was rescheduled to avoid conflict with the *jum'a* (Friday midday) prayer, so that all students could attend all course meetings.

At the advanced level, weekly contact hours and course credits were increased from three to four.

8.2 Instructor Involvement and Training

In order to increase buy-in and support from the instructors who teach the various course sections, the program set out to create a tight-knit instructional team that operates based on trust and mutual respect in an atmosphere of close and transparent cooperation. After initially scheduling Arabic team meetings every two weeks, the new team of instructors requested to increase the occurrence of those meetings to every week. Since not all instructors on the team had worked with the new textbook series before, it appeared useful to work together on selected items in the textbook and to exchange ideas and best practices for the teaching of specific grammar items, vocabulary lists, texts, and listening exercises. It was paramount that in these exchanges, instructors and program director work as equals, and that no one member of the team impose any teaching method on their colleagues. The expected outcome of activities and lessons drove the collaborative development of activities. In

addition, one instructor proposed the creation of a shared online activities bank, which has developed into a large repository of creative and effective activities.

Three years into the revised program, the Arabic team continues to meet every week. Instructors regularly select one or two items around which each team member designs activities – regardless of whether they teach the course in which these items are used or not. After exchanging updates on student performance and issues in the classroom during meetings, each person shares activities and the group critiques one another's work. Each instructor then has the opportunity to implement colleagues' activities in the classroom. In addition, on occasion a chapter of pertinent research has served as a basis for discussion during a meeting.

This in-depth work around class practice is supplemented by regular participation of all Arabic instructors in professional development workshops organized by other units such as the Language Center. Regardless of the content presented during those workshops, the practice of participating in them together and discussing them afterward as a group is in itself a team-building exercise. It also helps keep the group in touch with current research trends. Funding provided by the Flagship Proficiency Initiative PACE project brought to campus two external speakers, each of them experts in the field of Teaching Arabic as a Foreign Language, for a day-long workshop with the Arabic instructors. The first of those workshops revolved around the use of authentic texts in the Arabic classroom, while the second one focused on the incorporation of online corpora and concordances into vocabulary instruction. Ideas shared during those workshops have since been implemented in the classroom. Program instructors have also initiated meetings and activity exchanges with peer instructors from other language programs, which adds to the layers of inspiration from which all can draw. Finally, several instructors have enrolled in professional development workshops and training sessions outside of the University of Minnesota, regardless of the availability of university funds to finance these PD activities. Instructors' initiatives on this front, their positive contributions to weekly team meetings, constructive critique of one another's work, and cooperation with colleagues beyond the Arabic language program are placed front and center in their annual performance reviews.

8.3 Student Self-Assessment and Reflective Learning Practice

The PACE project was designed to involve not just externally rated proficiency assessments, but to develop a sustained program of student self-assessment practices throughout the foreign language course sequence. While student self-assessment has proven informative for the Arabic program teaching staff, the primary objective of having the Arabic students participate in these self-assessment sessions once or twice per semester was to plant the seeds of a reflective learning practice. In an effort to emphasize to students the benefit of self-assessment for them as language learners, these self-assessment sessions were never graded beyond participation. The protocol involved each student performing three

speaking tasks in front of a computer, listening to a voice-recording of their performance, and evaluating their performance, first individually, and then in discussion with a partner. A whole-group student-led discussion of strengths and challenges would follow, in which students were invited to share best practice study tips. Some instructors followed up on the self-assessment session with another 50-minute discussion session of learning strategies. During this discussion, students again took the lead in sharing study tips and articulating difficulties, with the instructor offering additional suggestions grounded in research on language learning strategies to vary and improve study habits, try out new study methods, and explore a variety of practice opportunities.

These practices laid the foundation for further measures aimed at increasing the reflective component of students' Arabic study. At the advanced level, surveys of students' study habits and perceived skill gaps formed the basis for curricular revisions that aimed to differentiate instruction at this level, tailoring it as much as possible to individual students' needs. Simultaneously, a reflective learning journal was added to the required course components for the advanced class. Students were expected to reflect on a weekly basis on successes and difficulties they had encountered with class activities and assignments; to set a limited number of short-term practice goals; and to evaluate their implementation of each week's goals. The instructor provided example reflection questions and practice goals. Following the first implementation of this project, which underscored the importance of close instructor follow-up and interaction with students' entries in their learning journals, the instructor increased the frequency of individualized feedback to every few weeks, engaging each student in an ongoing discussion of their study habits and goals. The journal project enhanced advanced students' competency as reflective learners and led to a variety of changes in student behavior in the classroom, study practice, and students' evaluation of their skill sets toward greater nuance and balance. All advanced students reported taking initiatives to improve their study habits and practice opportunities as a result of the journal project.

8.4 Establishing a Culture of Assessment

Students in the Arabic program were assessed on a more or less weekly basis through quizzes and tests, but quizzes and tests were not the main form of assessment in the classroom. Varied assessment practices aimed to provide opportunities for students with different skill sets and learning styles to demonstrate their skills in those ways in which they were most comfortable, *and* in ways that took them out of their comfort zone. Extensive attention was devoted to the assessment of students' presentational and interpersonal skills through frequent oral assessments at all levels of the curriculum. One such oral assessment consisted of a series of concise oral presentations followed by extensive Q&A with classmates, which required more time than the actual presentations. For these assessments, half of the grade points were assigned to confidence and pace in speaking, grammatical accuracy,

vocabulary range, and circumlocution skills, while the other half went to interaction with classmates during the Q&A. This assessment aided in developing good preparation habits for public speaking assignments, helped students confront anxiety about public speaking, and served as practice for the extensive final presentation students gave at the end of each semester. In addition, by placing emphasis on the Q&A component of the assessment, this assessment helped students practice preempting questions and responding in Arabic to unexpected inquiries. A second example of an interactive oral assessment used at all levels is the interview of a class guest. For this graded assessment, students are given a brief introduction to the guest the day prior to the interview and are assigned to prepare questions for the guest. During the first 15 minutes of class, instructor and students review their questions together, to allow students to feel confident about what they will be asking. The class is then given 40 minutes to interview the guest without intervention from the instructor. Students are awarded points for their participation in the interview and for the report they compose at home after class, in which they are expected to demonstrate not only what they have understood, but how well they were able to combine note-taking and listening during the interview.

Finally, at all levels, the program assesses students' presentational skills by having them produce short videos, either in response to discussion topics or, as more extensive projects, about an assigned topic. Video assessments enable students to record themselves multiple times, until they can deliver a product they are pleased with. Simultaneously, use of a recording platform like Flipgrid enables students to familiarize themselves with the format of speaking to a computer, which they encounter on the oral proficiency assessment administered through PACE. As students in previous years have reported discomfort with the format of the OPIc due to the artificial interlocutor, video assignments help students get accustomed to speaking to a computer.

In addition to frequent and varied testing, the Program also committed to the conscious nurturing of grit in the language classroom. Grit has been defined by Angela Duckworth as "passion and perseverance for very long-term goals" or "having stamina" —in other words, the ability to keep going despite challenges (2007, 2013). According to Duckworth, who researched the factors contributing to the success of individuals in a variety of learning and professional contexts, it is not primarily "social intelligence" or "IQ" that determines people's success in their long-term undertakings, but grit. In recent years, a body of research has developed around the concept of grit (Duckworth, Peterson, Matthews, & Kelly, 2007; Duckworth & Quinn, 2009; Maddi, Matthews, Kelly, Villarreal, & White, 2012; Wolters & Hussayn, 2015), with one academic institution adopting the concept as the foundation for an institution-wide experiment revolving around student success (Nutt, 2016). Other authors, meanwhile, have offered that "individualized pathways" for each student may be more important to enable success than an emphasis on grit alone (Rose, 2016). Attempting to strike a balance between these approaches, the Arabic program has favored an approach that combines experiments with differentiated, or individualized, instruction, with an active cultivation of grit. This cultivation involved minor but consistent daily practices, many of which align

closely with programmatic goals of maintaining high expectations. A few such practices are:

- insistence on the use of the target language from the earliest weeks of year one;
- consistent encouragement of students to give classroom contributions and responses multiple tries;
- having teachers model circumlocution skills instead of resorting to translation in any communicative situation in the Arabic classroom;
- creating extensive opportunities for students to practice Arabic outside of the classroom (through additional readings from the program's Arabic Library, listening exercises in the form of Arabic films, and partnerships with native speakers) and giving these opportunities high visibility through continuous advertising;
- challenging students on a daily basis to get involved in the classroom, rather than settling for volunteer contributions; and
- regular and explicit discussions with class groups and individual students, in the classroom, through the learning journals and in person during office hours about realistic expectations for progress; the importance of continuous review and practice; and incorporating rewards for sustained efforts. During these discussions with students, instructors who were themselves non-native speakers of Arabic regularly reflected on their own experience as learners of the language.

Ultimately, the Arabic program's mission is to support students' development toward skilled, confident users of Arabic capable of appropriately navigating the spectrum of Arabic language registers. With that mission in mind, training for confidence, even in challenging situations, becomes key. To allow such confidence and a "gritty" mindset to develop, program leadership modeled the classroom as a safe space for students to experiment; a space in which humor cuts through the pressure of performing with limited linguistic skill; and a space in which students are held accountable on a daily basis for their work and its quality. This did not involve awarding grade points beyond comments and corrections to students' responses on each daily homework assignment, to avoid shifting learners' attention from the learning process to the total course grade.

To retain the focus on the learning process, when the PACE proficiency assessments were added to students' regular achievement assessments, they were presented as a required course component, worth 10% of the total course grade, in which only participation and effort mattered. To integrate these proficiency assessments into the curriculum, they were scheduled during regular class periods. Instructor discourse around the PACE assessments was discussed at some length within the instructional team and focused consistently on the value of these assessments for the students, who would be able to add a strong rating to their resume and use it for graduate school and job applications. Not only the outcome, but the testing process itself was consistently presented as a valuable exercise for students intending to use Arabic for graduate studies or professionally. These students could expect to encounter similar proficiency exams in the future and familiarity with these tests would aid their future performance. This discourse was repeated consistently during

and after the testing process, particularly when students expressed having difficulty with the format of the OPIc.

9 Conclusion

Results of externally rated proficiency assessments conducted with Arabic program students show a dramatic improvement in proficiency results between the first and second year of testing. While during the first round of testing second-semester speaking results met expectations for students who had experienced the new curriculum, lagging proficiency results in fourth and sixth semester improved palpably during the second round of testing. Proficiency results for reading improved at all levels between the first and second year of testing. Simultaneously, self-assessment results for third-, fourth- and fifth-semester students suggest an increase in confidence about Arabic speaking skills. Among the changes made in the Arabic program from Fall 2014 onward, suggested as having contributed to the improvement in proficiency outcomes are the articulation and implementation of high expectations for students and instructors alike; intensive work within the instructional team on classroom management, effective teaching practices, materials development, and grading standards; the adoption of a self-assessment protocol followed by a more continuous practice of reflective learning; and the creation of a culture of assessment and resilience around proficiency testing through frequent and varied assessment practices and conscious nurturing of grit in the Arabic language classroom, along with the exploration of forms of differentiated instruction.

References

ACTFL Proficiency Guidelines. (2012). https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Brustad, K., al-Batal, M., & al-Tonsi, A. (2013). *al-Kitaab Fii Ta'allum al-'Arabiyya* (3rd ed., Part Two). Washington, DC: Georgetown University Press.

Center for Applied Linguistics. (2015). *Spring 2015 evaluation report*. Unpublished, delivered to the CLA Language Center at the University of Minnesota.

Clifford, R., & Cox, T. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals, 46*(1), 45–61. https://doi.org/10.1111/flan.12033

Cox, T., & Clifford, R. (2014). Empirical validation of listening proficiency guidelines. *Foreign Language Annals, 47*(3), 379–403. https://doi.org/10.1111/flan.12096

Duckworth, A. L. (2013). *Grit: The power of passion and perseverance*. TED Talk, https://www.ted.com/talks/angela_lee_duckworth_grit_the_power_of_passion_and_perseverance

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the short grit scale (Grit-S). *Journal of Personality Assessment, 91*(2), 166–174. https://doi.org/10.1080/00223890802634290

Maddi, S. R., Matthews, M. D., Kelly, D. R., Villarreal, B., & White, M. (2012). The role of hardiness and grit in predicting performance and retention of USMA cadets. *Military Psychology, 24*, 19–28. https://doi.org/10.1080/08995605.2012.639672

Nutt, L. A. (2016, June 14). *Best of times – Worst of times: How GRIT mindset changed a college.* http://www.pearsoned.com/education-blog/grit-mindset-changed-college/

Rose, L. T. (2016). *The end of average*. New York, NY: HarperOne.

The Language Flagship. (2014). *Request for proposal: The Language Flagship proficiency initiative application guidelines*. Available online at: https://www.thelanguageflagship.org/sites/default/files/Flagship%20Proficiency%20Initiative%20_%202014_0.pdf

Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals, 49*(2), 201–223. https://doi.org/10.1111/flan.12198

Wolters, C. A., & Hussain, M. (2015). Investigating grit and its relations with college students' self-regulated learning and academic achievement. *Metacognition and Learning, 10*(3), 293–311. https://doi.org/10.1007/s11409-014-9128-9

Katrien Vanpee is Director of the Arabic Language Program at the University of Minnesota, where she teaches Arabic language and literature. Her research interests center on classical and modern Arabic poetry, *nabaṭī* poetry, the cultural heritage of the Arabian Peninsula, Teaching Arabic as a Foreign Language, curriculum design, and program management. Her publications include *La Mafarr: Leermethode Arabisch* (Van Mol, Vanpee and Marogy; Peeters, 2007), and work that has appeared in *The Modern Language Journal* and *al-ʻArabiyya*.

Dan Soneson is Director of the Language Center in the College of Liberal Arts at the University of Minnesota Twin Cities and the Principal Investigator of the Proficiency Assessment for Curricular Enhancement (PACE) project, a four-year federal grant within the proficiency initiative sponsored by the Language Flagship. He has co-edited volumes on Language Teacher Education and Cultures and Languages Across the Curriculum, published as working papers by the Center for Advanced Research on Language Acquisition (CARLA) at the University of Minnesota, and has also published on language curriculum innovation and assessment. Other interests include self-assessment and implementation of technology in second language education.

A Cross-Linguistic and Cross-Skill Perspective on L2 Development in Study Abroad

Dan E. Davidson and Jane Robin Shaw

Abstract The present study reports on measured gains in L2 proficiencies in speaking, reading and listening of U.S. students (N = 308) who took part in year-long federally funded overseas immersion programs for Arabic, Chinese and Russian. Subjects were late adolescent and young adult learners of diverse social and economic backgrounds participating in year-long structured instructed immersion programs hosted in China, Kazakhstan, Moldova, Morocco and Russia. L2 gains in post-program proficiency levels from 4.76 to 7.74 standard deviations above pre-program measured levels are reported for both the early- and the late-stage learners: Mean post-program proficiency levels of ILR-2, CEFR-B2 are demonstrated by the early-stage learners across skills in all three target languages. The mean post-program proficiency levels of ILR-3, CEFR-C1 of the university subjects meets certification levels for language-designated positions in in most U.S. government and professional organizations. The study also examines skill gains across modalities: Advanced participants show concurrent gains across three skills: reading, listening, and speaking. Post-program reading and speaking are strongly correlated with pre-program listening at the advanced levels. Reading ability is strongly associated with gains in speaking and in listening skills, as the student progresses from novice through the professional level.

Keywords L2 gain · Immersion · Study abroad · Cross-skill correlations · Critical languages · SLA · Diversity abroad · Professional proficiency

D. E. Davidson (✉)
American Councils Research Center (ARC), American Councils for International Education, Washington, DC, USA

Myra T. Cooley Lectureship, Bryn Mawr College, Bryn Mawr, PA, USA
e-mail: davidson@americancouncils.org

J. R. Shaw
Department of Russian, Bryn Mawr College, Bryn Mawr, PA, USA
e-mail: jshaw@brynmawr.edu

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_12

At the request of the U.S. Congress, the American Academy of Arts and Sciences (AAAS) has released a major report on language learning in the United States, *America's Languages: Investing in Language Education for the 21st Century* (AAAS, 2017). Among five areas recommended for policy attention in the Academy Report was a call for expanded access to study abroad, for "students to travel, experience other cultures, and immerse themselves in languages as they are used in everyday interactions and across all segments of society" (AAAS, 2017, p. 27).

As noted in the Academy Report, language learning in the study abroad context has the capacity to produce significant linguistic and cultural gains, but overseas study is also costly, and substantial growth in language is by no means achieved by all those who go abroad (Freed, 1998; Mason, Powers, & Donnelly, 2015; Vande Berg, Connor-Linton, & Paige, 2009). As with any educational setting, program design, teacher preparation, student motivation, time-on-task, and an appropriately supportive environment for learning are critical components for successful language acquisition in the study abroad context. While the total number of U.S. students who study abroad has increased over the past two decades to 313,415 annually, most study currently takes place in English-speaking regions (*Open Doors*, 2016). Moreover, despite the well-documented benefits of longer-term immersion, only 2.5% of Americans studying abroad in 2014–2015 stayed a full academic year, reflecting an unfortunate decline in long-term study over the past twenty years (Dwyer, 2004; Kinginger, 2011; Pellegrino Aveni, 2005). In a recent large-scale comparison of summer and academic-year overseas study programs, language gains were compared for differing target languages, initial levels, and program durations: The greatest gains, regardless of starting point or target language, were associated with year-long programs (Davidson, 2015).

For those students who do undertake serious year-long language study, the structured, federally-sponsored programs initiated under the National Security Language Initiative of 2006 (see, for example, https://exchanges.state.gov/us/program/nsliy) have demonstrated a capacity over the past decade for producing advanced and superior-level speakers on the American Council on the Teaching of Foreign Languages (ACTFL) proficiency scale (see ACTFL, 2012). Alumni of the NSLI-Y (National Security Language Initiative Youth) high school senior secondary program (SP) are fully prepared to enter advanced-level university course work taught in the target language, and alumni of the undergraduate Flagship programs (UP; see https://thelanguageflagship.org/) regularly go on to join the U.S. workforce as bilingual professionals (Murphy & Evans-Romaine, 2015; Powell & Lowenkron, 2006; USED, 2008). Unfortunately, the number of Americans benefitting from these programs still falls far short of meeting the growing needs of government, business, international development, and society at large (Brecht, Rivers, Robinson, & Davidson, 2015; Damari, Rivers, Brecht, Gardner, & Robinson, 2017; Rivers, 2012).

1 Assessing Language Acquisition in the Study Abroad Context

The systematic study of language acquisition during study abroad is now a well-recognized subfield of second language acquisition scholarship (cf. special issue of *System,* 2017; Winke & Gass, 2018). Within that subfield, issues of variation in learning outcomes and ultimate attainment dominate much of the empirical research (Baker-Smemoe, Dewey, Bown, & Martinsen, 2014; Brecht, Davidson, & Ginsberg, 1995; Davidson, 2015; DeKeyser, 2007; Dewey, 2004; Freed, 1998; Mason et al., 2015; Watson, Siska, & Wolfel, 2013). Group-level analyses of standard measures of pre-program and post-program L2 proficiencies using the ILR (http://www.govtilr.org/) or ACTFL scales for speaking, reading, and listening can provide a validated and reliable cross-linguistic basis for comparing relatively robust numbers of student records, while controlling statistically for target language, modality, initial L2 level, duration of immersion, and other linguistic and learner background variables.

Proficiency-based measures are widely used today as a component of programs of formative and summative assessment as well as for participant certification purposes. The ACTFL and ILR proficiency scales, and the larger *World-readiness standards for learning languages* (NSFLEP, 2015) of which they are a part, are widely used today in K-12 (e.g., the Seal of Biliteracy in 25 states), college placement, study abroad, and teacher education programs (ACTFL, 2012; NSFLEP, 2015). Most importantly, the ILR scale, on which the ACTFL proficiency guidelines are based, is used virtually across all U.S. government agencies recruiting for language-designated positions (Herzog, n.d.; Interagency Language Roundtable, 2016; Liskin-Gasparro, 1984; Tschirner, 2011). As widely used as these standardized measures are, the authors acknowledged inherent limitations with these and other large-scale L2 proficiency models and testing scales currently in use, whether ILR/ACTFL, the Common European Framework of Reference, the TORFL (Russian), TOPIK (Korean), HSK (Mandarin), or others (see Bachman, 1988; Kramsch, 2014; North, 2006). While the current generation of proficiency tests do not capture the full dynamic range of linguistic and cultural repertoires of which the L2 user may be capable, the final proficiency rating, whether for speaking, writing, reading, or listening, is based on an individualized analysis of evidence produced by the candidate under controlled examination conditions: an L2 product (interpersonal communication, presentation, textual interpretation) evaluated in terms of its overall effectiveness and appropriateness for the intended audience.

2 Assessing L2 Across Modes and Modalities

Previous research on cross-skill gains in language proficiency in the study abroad environment has noted a relationship at the intermediate-to-advanced levels between study abroad participants' initial reading levels and their ultimate gains in listening comprehension, while strong structural control of the L2 has been consistently associated with gains across all skills (Brecht et al., 1995). Grammatical (structural) knowledge makes both visual and aural input comprehensible and allows the language learner to improve processing speed and build confidence in reading and listening (Krashen, 1985; Norris & Ortega, 2003; Ortega, 2009). Examining factors affecting L2 gain at the "superior" (ILR 3) level, Davidson (2010) observed that pre-program listening comprehension levels among advanced students of Russian were predictive of program-final oral proficiency outcomes. The higher the initial listening comprehension score, the more likely the candidate was to achieve a score of 3 ("superior") or higher in speaking by the end of the academic program. Noting the need for further study of cross-skill correlations with oral proficiency gains, the author hypothesized that strong listening comprehension appears to be critical for the L2 learner in detecting and acting on the spoken feedback of native interlocutors (e.g. re-tracings, recasts, paraphrasing) that make up a regular part of the students' extended interactions with local friends and contacts in the immersion context at that level (Davidson, 2010; Winke & Gass, 2018). Self-correction behavior, more generally, has also been identified as yet another predictor of L2 gain for young adults in the immersion environment (Golonka, 2000).

The American Academy report identifies several federally-sponsored programs as exemplifying best practices in effective overseas immersion language training (AAAS, 2017). Two of those federal programs, one open to senior secondary students (SP), the other primarily to undergraduate juniors and seniors (UP), monitor student progress through systematic pre- and post-program assessment of proficiency levels in speaking, reading, and listening. The two federal programs will serve as sources of performance-based data for the present comparative study of learning outcomes for American students of Arabic, Chinese, and Russian in the overseas immersion context. In addition to skill-specific reports for each of the target languages, cross-skill relations by skill-specific proficiency levels will be reported and compared here as well.

3 Study Participants

The present study includes data from late-adolescent and young adult participants in two major federal programs focused on an intensive in-country immersion study of Arabic, Chinese, and Russian (N = 308). They include year-long undergraduate students (UP) of Russian (N = 126) and pre-college participants in the Arabic academic-year program (N = 47), Chinese academic-year program (N = 78), and

Russian academic-year program (N = 57). The federal funding model for these programs was designed to encourage participation by students from a greater range of socioeconomic backgrounds than is normally possible for fee-based study abroad programs. The SP admits students on a competitive basis and without regard to their ability to pay, including students with no prior experience of learning the target language in question. The UP subsidizes a substantial portion of all program costs—under certain circumstances, all costs—and requires applicants to demonstrate advanced level (ILR-2) proficiency in speaking and at least one other skill and to test at no lower than a 1+ in the third skill. A writing proficiency test is now being added to the testing portfolio for the UP languages but is not included in the present analysis. (See Appendix 1 and 2 for full list of selection criteria for both programs.)

While it is impossible to control for pre-selection effects in the analysis of these two cohorts, the researchers believe, given the basis on which candidates were selected and funded, that the outcomes data included here may be regarded as generally representative of the impact of a year of overseas language immersion study on that segment of the U.S. student population who elect to apply for and accept positions in a federally-funded study abroad program, regardless of their socioeconomic and educational backgrounds.[1]

4 Data and Testing Instruments

Testing, conducted at the beginning and end of the programs, includes face-to-face or telephonic oral proficiency interviews (OPIs) and online proficiency-based reading and listening comprehension examinations based on the ACTFL and ILR scales (ACTFL, 2012; Interagency Language Roundtable Scale, 2016).

Anonymized participant score reports that included OPI pre- and post-program test scores were made available to the researchers and analyzed for a total of 308 year-long participants in Arabic, Chinese, and Russian. Reading and listening pre- and post-program scores were made available to the researchers for all advanced-level (UP) study subjects and for the Russian study subset of the early-stage (SP) learners. The overseas study programs in question took place between 2009 and 2014 in China, Jordan, Kazakhstan, Moldova, Morocco, Russia, and Taiwan. Small-group instruction, peer tutors, homestays, attendance of regular local classes, and integrated cultural enhancement programs were standard features of all programs. UP students also participated in internships. Detailed description of the overseas program designs and interventions in use over the 2009–2014, including

[1] Analysis of the distribution of K-12 foreign language enrollments across the U. S. indicates a correlation between socio-economic levels within a school district (as reflected in the 2010 U.S. Census) and the likelihood that the district will (or will not) offer a foreign language, defined as a language other than English, at the K-12 level. See discussion of estimate models, *National FL Enrollment Survey* (2017), https://www.americancouncils.org/sites/default/files/FLE-report-June17.pdf

the two selected for analysis in the present study (SP and UP), may be found in Davidson (2015).

The combined speaking, reading, and listening comprehension data (reflecting presentational, interpersonal, and interpretive modes of communication) are considered by the authors to provide a robust, cross-modal, multi-language array of aligned measures of L2 skills for use in assessing and comparing outcomes for the programs under study. Pre-program scores reflect the skill levels participants brought to their study-abroad experiences and serve as a baseline against which subsequent changes are measured.

5 Research Questions

Given the recognized value of overseas immersion for the acquisition of foreign languages at the advanced and professional levels, reliable information on learning outcomes across different target languages and with regard to specific skills should be widely available to teachers, advisors, and policymakers concerned with the preparation of a new generation of L2 users and professionals. For that reason, the present study poses the following research questions:

1. What are the mean gains in oral proficiency of students who participate in structured year-long study programs in Arabic, Chinese, and Russian, as measured by changes in levels of proficiency as well as in units of pre-program standard deviation?
2. To what extent do the choice of target language and the student's initial levels of proficiency affect gains in OPI?
3. To what extent are second language (L2) gains in reading, listening, and speaking correlated for students in the year-long study-abroad context? To what extent do specific pre-program skills account for post-program attainment across skills?

The growth in importance of study abroad and in access to study abroad has made these questions more broadly relevant today than was the case in years past. The current study hopes to advance understanding of linguistic factors that contribute to successful learning of three critical languages, Arabic, Chinese, and Russian, in the study abroad context.

6 Data Collection and Preparation

OPI testing was conducted by ACTFL-certified oral proficiency testers; post-program OPI testing of participants with higher initial proficiency levels was administered face-to-face, while pre-testing and lower-range tests were administered telephonically. Proficiency-based reading and listening comprehension tests were administered under proctored, computer-mediated conditions. Test specifications,

item development, and scoring protocols for the reading and listening proficiency tests are described in detail by Bazarova, Lekic, and Marshall (2009); statistical documentation of the reading and listening tests may be found at Wothke and Petersen (2017).

Testing data are reported using the ACTFL and/or the ILR scales, depending on the phase of study of the participant. For ease of statistical analysis, the researchers converted ILR-scaled scores to ACTFL scores using the following conversions:

0+ to novice-high,
1 to intermediate-mid,
1+ to intermediate-high,
2 to advanced,
2+ to advanced high,
3 to superior,
3+ to superior-high,
4 to distinguished.

To avoid introducing additional measurement error as a result of the necessary score conversion, and given that the ACTFL and ILR scales do not fully align, the authors also report ILR data in those cases where score conversions were undertaken. This procedure is consistent with other recent studies (Davidson, 2015; Davidson, Garas, & Lekic, 2016; Mason, Powers, & Donnelly, 2015).

Since proficiency scores represent ordinal values, pre- and post-program score columns in the data sets with numeric values were then created. Integers from 1 to 18 were assigned for each ACTFL rating from novice-low to distinguished, with novice-low as 1. An additional unit was added to the coding to account for threshold-level junctures on the proficiency scale (novice, intermediate, advanced, superior, and distinguished). Thus, novice-high to intermediate-low is marked by a move from 3 to 5, while intermediate-high to advanced-low is represented by a numerical shift from 7 to 9 on the linear scale, and so forth. While none of the participants received a final program score of "superior-low," as testers do not generally give this score, a space of 1 unit was left in the column for this rating in order to maintain consistent intervals across languages and proficiency levels. Values for all ACTFL ratings and for the intervening values are given below in Table 1:

Given the nature of the three-dimensional construct, the "inverted pyramid," employed for ILR and ACTFL proficiency assessment, L2 gains tend to post at a more rapid rate at lower levels of proficiency but require increasingly more time as the participant grows and advances to higher levels of proficiency (Brecht et al., 1995). A more nuanced mathematical model has yet to be developed and accepted within the foreign language assessment community to account statistically for the time-on-task differentials implicit in the successive levels of the ILR proficiency scale. (See also Tigchelaar, this volume, for additional information on this.) The use (above) of an additional numerical value (4, 8, 12, 16) at each threshold level along the scale is an entirely arbitrary but statistically helpful intervention both to mark the additional functional and expressive capacity represented by the next level up on the scale and to mitigate the effects of restriction of range within clusters of pre- and post- test scores.

Table 1 Numerical values by ACTFL (Ordinal) ratings

ACTFL rating	Value in database
Novice-low	1
Novice-mid	2
Novice-high	3
(Threshold)	4
Intermediate-low	5
Intermediate-mid	6
Intermediate-high	7
(Threshold)	8
Advanced-low	9
Advanced-mid	10
Advanced-high	11
(Threshold)	12
(Superior-low)	13
Superior-mid	14
Superior-high	15
(Threshold)	16
(Distinguished-low)	17
Distinguished	18

7 Analysis

To assess language specific and overall L2 gain within the immersion programs, initial distributions were run of OPI values for each language (Arabic, Chinese, and Russian) using data from all year-long SP and UP participants ($N = 459$). The distributions were categorized by pre-program OPI and post-program OPI; score changes ("delta" values), if any, were tabulated and included for each as well. The subjects were divided into three groups for analytic purposes: those who began the program at the "novice" proficiency level, those who began at the "intermediate" level, and those who began at the "advanced" level.

To test for relationships across modalities, multivariate pairwise correlations were run and univariate simple statistics were recorded using data from SP and UP Russian participants. (Reading and listening data were available only for the Russian subset of SP but for all participants in UP.) Both Pearson and Spearman correlations were generated. For each participant grouping ("novice," "intermediate," and "advanced"), a set of correlations among delta (reading), delta (listening), and delta (OPI) was generated (with "delta" signifying change within scores from pre-test to post-test); a set of correlations across all participant levels was also run.

Multivariate pairwise correlations, with corresponding univariate simple statistics, were also generated to test for relationships among initial and post-program levels in reading and listening. Sets of correlations were run across all levels and for each participant grouping ("novice," "intermediate," and "advanced"); pre-program

reading values, pre-program listening values, post-program reading values, and post-program listening values were correlated.

Both cross-skill and same-skill correlations (e.g. pre-/post-reading; pre-/post-listening, pre-/post-speaking) were performed throughout to verify the overall homogeneity of the data and to check, in particular, for any significant differential effects that might influence the analysis related to participant gender, age, heritage background, program year, and program site. No significant external or programmatic effects were found (Shaw, 2017).

A third set of multivariate pairwise correlations and univariate simple statistics was generated to test for relationships among initial skills in reading and listening and post-program OPI attainment. Sets of correlations were run for each language and across all proficiency levels ("novice," "intermediate," and "advanced"); pre-program reading values, pre-program listening values, and post-program OPI values were tested.

In order to clarify further the relationship of language gains across skills (speaking, reading, and listening), distributions of gains with participants categorized, as previously, by pre-program OPI levels were run across modalities. Mean delta (skill) values for "novice," "intermediate," and "advanced" Russian academic-year SP and UP participants were considered and compared. While previous distributions focusing on OPI results included participants without reading and listening data, for this test, only participants with delta values in every modality and all three levels, the Russian-only data set, were considered.

Pre- and post-program reading and listening data and their respective relationships to OPI gains were run for each modality (reading and listening), adjusted for pre-program values for that modality. For one set of distributions, participants were grouped by novice, intermediate, and advanced pre-program reading values; mean pre-program reading values, post-program reading values, delta (reading), post-program OPI values, and delta (OPI) were considered. For a second set of distributions, participants were grouped by novice, intermediate, and advanced pre-program listening values; mean pre-program listening values, post-program listening values, delta (listening), post-program OPI values, and delta (OPI) were considered.

Regression analyses using fit models were performed on all year-long participant data to test statistical relationships between pre-program skills and ultimate OPI attainment as measured by post-program OPI values. For fit model type, standard least squares with emphasis on effect leverage were chosen. The results were represented as leverage plots, and corresponding statistics were generated.

Post-program OPI values represented the dependent variable. Plots with pre-program reading, pre-program listening, and pre-program OPI values as independent variables were generated. The plots and accompanying statistics were then examined to determine which independent variable had least effect and whether any variables had negative effects; new sets of leverage plots were then generated as applicable using the remaining variables.

8 Results

The present study has addressed the measurement of L2 gain across languages with respect to the student's initial level of proficiency and choice of target language; delta (OPI) and delta (OPI) in units of pre-program standard deviation were also calculated. The subject population, as noted above, comprised late-adolescent and young-adult learners of the critical languages. For all distributions, the duration of the immersion program (intervention) was one academic year (9 calendar months).

The effect of the immersion intervention on the cohort ($N = 77$) beginning the programs in Arabic, Chinese, and Russian at the novice level is highly significant, ranging from 6.36 (Arabic) to 6.91 (Chinese) to 7.30 (Russian) standard deviations above the measured pre-program means. For those beginning the program in the same three languages at the intermediate level ($N = 53$), the effect is again highly significant, but slightly weaker: Arabic (6.93), Chinese (4.76), and Russian (5.74). For those beginning the program at the advanced level ($N = 112$), the mean gain deltas are 7.74 for all participants.

Reviewing the three language-specific cohorts across programs, one notes that the proficiency gains (deltas) are comparable across all proficiency levels (as are the standard deviations), with gains at the intermediate level slightly more modest than those posted by the novices and advanced students. The latter is particularly significant in light of the expected effect of the measurement artifact, noted above.

9 Pre-/Post-program L2 Gain Levels and Gain Amounts (Deltas) by Modality

Multivariate analyses were conducted for participants across proficiency levels and at each specific level comparing delta (skill) values across modalities. For Russian academic-year participants (the only group for which speaking, reading, and listening proficiency scores were available for all levels of study), Pearson correlations showed a moderate, statistically significant correlation between gains in reading and in listening over the period of study, noted here as "delta (R)" and "delta (L)" ($r = 0.3338$, $p = 0.0010$). Spearman correlations showed a moderate, statistically significant positive correlation between delta (R) and delta (L) (rho = 0.4002, $p < 0.0001$).

For novice Russian academic-year participants, Spearman correlations showed a strong, statistically significant positive correlation between delta (R) and delta (L) values (rho = 0.6825, $p = 0.0207$). For intermediate Russian academic-year participants, no correlations met the probability threshold for statistical significance. For advanced Russian academic-year participants, Pearson and Spearman correlations showed moderate, statistically significant positive correlations for all pairings. Spearman correlations also showed statistically significant positive correlations for all pairings, slightly stronger than Pearson but still moderate.

Academic year L2 Russian participant data ($N = 183$), from the cohort for which pre-and post-program reading and listening data were available for entering novice-level participants, as well as data for those who entered study at the intermediate and advanced proficiency levels) demonstrated strongly correlated, statistically significant relationships across all proficiency levels.

9.1 Russian Academic-Year Participants, All Levels: Reading to Listening

			Pearson correlations		Spearman correlations	
			Correlation	Sign. Prob.	Spearman rho	Prob. > |rho|
Pre-R	–	Post-L	0.8882	<.0001*	0.6023	<.0001*
Pre-L	–	Post-R	0.8914	<.0001*	0.7344	<.0001*

Pre-program reading levels were strongly correlated with post-program listening outcomes, and, conversely, pre-program listening levels also predicted post-program reading attainment.

9.2 Intermediate Participants: Reading to Listening

			Pearson correlations		Spearman correlations	
			Correlation	Sign. Prob.	Spearman rho	Prob. > |rho|
Pre-R	–	Post-L	0.9412	<.0001*	0.9396	<.0001*
Pre-L	–	Post-R	0.8969	<.0001*	0.8930	<.0001*

The finding is consistent with Brecht et al. (1995), which focused exclusively on semester-length overseas Russian immersion, and Davidson (2010), which compared summer, semester, and academic year outcomes for overseas Russian. In both studies, reading proficiency was strongly correlated, in turn, with target-language grammatical/structural control, and both reading and grammar served as more or less equivalent predictors of ultimate oral proficiency gain at the advanced level.

9.3 Pre-reading/Pre-listening to Post-OPI

Multivariate analyses for pre-reading (pre-R) and pre-listening (pre-L) scores with post—program OPI results for *all levels* of study showed the following results:

			Pearson correlations		Spearman correlations	
			Correlation	Sign. Prob.	Spearman rho	Prob. > |rho|
Pre-R	–	Post-OPI	0.7896	<.0001*	0.5956	<.0001*
Pre-L	–	Post-OPI	0.8041	<.0001*	0.6700	<.0001*

Pearson correlations were slightly stronger for pre-program listening and post-program OPI ($r = 0.8041$, $p < 0.0001$) than pre-program reading and post-program OPI ($r = 0.7896$, $p < 0.0001$) for the cohort as a whole.

10 Comparison of Mean AY Skill Gains (R, L, S) by Initial OPI Proficiency Level

Based on existing program data, for which academic-year immersion data are available at all three levels, the distribution of skill-specific gains categorized by the participant's initial (pre-program) speaking proficiency presents the following results.

	Delta (Reading)		Delta (Listening)		Delta (OPI)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Novice	3.20	2.10	3.50	1.65	5.70	1.70
Intermediate	4.23	1.79	4.00	1.52	4.54	1.27
Advanced	3.72	1.76	3.86	1.50	3.99	1.13

Novice (N = 10), Intermediate (N = 13), Advanced (N = 69)

As seen in the earlier distribution of Russian academic-year participants by level (Table 3.2), the mean gain delta for the cohort of advanced-level students is slightly smaller numerically than for those who began the program with speaking levels at the novice or intermediate level, most likely an effect of the measurement artifact discussed above. Looking at all three levels, however, it is clear that *the immersion experience for early stage learners is accompanied by relatively rapid rates of gain in speaking*. As the learners' speaking skills improve, the data show a more evenly distributed range of skill gains (both means and the size of standard deviations). Increased ability and opportunities for self-expression and interactions with locals also multiply the need for cross-skill and multi-modal forms of communication at the intermediate and advanced levels.

The results of analyses of pre-and post-program reading levels and their respective relationships to OPI gains are presented in the following distributions:

10.1 Distributions of all AY Scores Based on Initial Levels of Reading Comprehension

		Novice pre-program reading	Intermediate pre-program reading	Advanced pre-program reading
Pre-program reading	Mean	1.80	5.44	10.41
	Std. Dev.	0.42	0.73	0.50
Post-program reading	Mean	6.00	9.00	14.57
	Std. Dev.	2.49	1.87	1.27
Delta (Reading)	Mean	4.20	3.56	4.16
	Std. Dev.	2.35	1.67	1.39
Post-program OPI	Mean	8.50	9.44	14.13
	Std. Dev.	1.35	1.51	1.36
Delta (OPI)	Mean	5.00	4.78	4.16
	Std. Dev.	1.49	1.72	1.27

Novice pre-program reading (N = 10), Intermediate pre-program reading (N = 9), Advanced pre-program reading (N = 63)

Novice-level readers showed the greatest gains in both reading and in speaking, followed by those who began the program as advanced-level readers. Gains in reading were notable for each group, ranging from 3.56 to 4.20 mean delta (reading). When delta (OPI) values were compared based on participants' pre-program reading levels, mean delta (OPI) decreased only very slightly from 5.00 (novice) to 4.78 (intermediate) to 4.16 (advanced), differences most likely resulting from the effects of the measurement artifact itself.

The results of analyses of pre-and post-program listening levels and their respective relationships to OPI gains are presented in the following distributions:

10.2 Distributions of all AY Scores Based on Initial Levels of Listening Comprehension

		Novice pre-program listening	Intermediate pre-program listening	Advanced pre-program listening
Pre-program listening	Mean	2.46	6.50	10.46
	Std. Dev.	0.52	0.71	0.50
Post-program listening	Mean	6.15	11.60	14.37
	Std. Dev.	1.68	2.72	0.89
Delta (Listening)	Mean	3.69	5.10	3.91
	Std. Dev.	1.65	2.13	0.95
Post-program OPI	Mean	8.46	11.30	14.22
	Std. Dev.	1.45	2.00	1.22
Delta (OPI)	Mean	4.92	4.50	4.17
	Std. Dev.	1.75	0.97	1.21

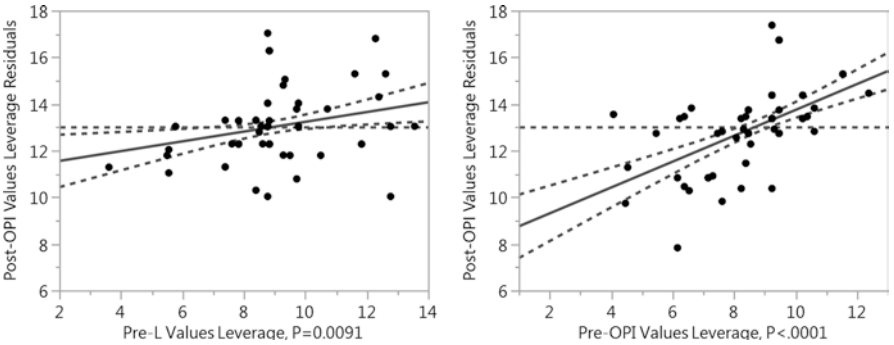
Novice pre-program listening (N = 13), Intermediate pre-program listening (N = 10), Advanced pre-program listening (N = 65)

Unlike for speaking and reading, participants who began the immersion program with intermediate levels of pre-program listening showed the greatest gains over the course of the immersion year. That said, listening gains were substantial for all groups, ranging from 3.69 to 5.10 mean delta (listening). When delta (OPI) values were compared based on participants' pre-program listening levels, mean delta (OPI) were observed to decrease very slightly from novice to intermediate to advanced listeners. This mirrors the patterns for delta (OPI) gains by level previously noted for speaking and reading. In fact, no matter which pre-program skill was selected as the independent variable, very similar patterns in delta (OPI) appeared, trends which are programmatically significant precisely because they are so small in this case, given the well-documented effects of the measurement artifact itself.

11 Fit Model Analysis: Pre-program Levels as Predictors of Post-program OPI

When the effects of the three pre-program variables were tested, parameter estimates were 0.02162 for pre-program reading, 0.1924 for pre-program listening, and 0.5487 for pre-program OPI. Pre-program reading had a slight positive effect but did not meet the threshold for statistical significance. Pre-program listening had a positive effect on post-program OPI; while it did not meet the threshold for statistical significance, it came much closer than did pre-program reading. Pre-program OPI had a substantial, statistically significant positive effect on post-program OPI.

However, when pre-program reading, the independent variable with the least leverage on post-program OPI, was removed and the relative effects of pre-program listening and pre-program OPI and analyzed again, the results were notable:



Parameter estimates were 0.2094 for pre-program listening and 0.5544 for pre-program OPI; both independent variables had positive, statistically significant effects on post-program OPI. Pre-program listening was observed to contribute more than 20% of the variation in post-program OPI values present in the model.

12 Findings and Discussion

Overall, the immersion intervention effects for the early-stage SP learners and for the advanced-level SP and UP subjects were highly significant: 7.30 and 7.74 standard deviations respectively. The linguistic and cultural impact on both groups is significant, permitting graduates of SP to enter college-level courses at sophomore and junior levels and UP graduates to move directly into government and private sector positions requiring professional levels (ILR-3) of linguistic and intercultural competence and above.

Mean delta (OPI) was 5.81 for novice participants, 4.36 for intermediate participants, and 4.19 for advanced participants; gains in units of pre-program standard deviation were 7.30 for novice participants, 5.58 for intermediate participants, and 7.74 for advanced participants. Proficiency gains were comparable across levels, with gains at the intermediate level slightly more modest than those posted by the novices and advanced students. The latter is noteworthy in that advanced-level gains are relatively more difficult to achieve, due to the effects of the measurement artifact (the inverted pyramid), which assumes considerably greater effort and time-on-task to move from Level 2 to Level 3 than from 1 to 2 or from 0 to 1 (ACTFL, 2012; Brecht et al., 1995). As noted above, the expected decline in mean delta (OPI) values as proficiency levels rise was, in fact, very gradual and barely observable. These findings relate to presumed cognitive, academic, and socio-emotional effects of the immersion intervention at more advanced levels of acquisition on the learning process.

A set of distributions was generated later in the study comparing delta (OPI), delta (reading), and delta (listening) values among Russian academic-year participants; these distributions included only those participants for whom delta (skill) values in all three modalities were available. Consideration of the mean delta (OPI) data for this participant subset allows the examination of patterns in gains by level with slightly different selection criteria in place. As noted above, given the well-documented "artifact effect" of the proficiency measurement model (the inverted pyramid), mean deltas would be expected to decrease as the student progresses in learning along the proficiency scale from one proficiency threshold to the next highest. For these cohorts, however, the delta values were still relatively robust for every level and modality: 5.70 for novice participants, 4.54 for intermediate participants, and 3.99 for advanced participants. As with the previous set of distributions for Russian academic-year participants, the change in mean delta (OPI) with increasing level was indeed observable, but limited. Cross-testing of other participant groups within the larger database did not substantially alter this pattern. The consistency of these results may be seen to further attest to the value of overseas language immersion as a facilitator of language gain at all levels, and to its particular value at the upper-intermediate and advanced levels, when comparable gains are more difficult to achieve in the domestic learning context.

13 Cross-Modality Patterns for Study-Abroad Participants

To assess the relationships among pre-program skill levels and post-program outcomes within Russian academic-year program, several sets of multivariate pairwise correlations were run. When gains across all levels were examined, Pearson correlations showed a moderate, statistically significant correlation between delta (reading) and delta (listening) ($r = 0.3338$, $p = 0.0010$). Spearman correlations showed a moderate, statistically highly significant positive correlation between delta (reading) and delta (listening) ($rho = 0.4002$, $p < 0.0001$) and a moderately weak but

statistically significant positive correlation between delta (reading) and delta (OPI) (rho = 0.2361, p = 0.0220). Reading and listening gains increased together across levels, as did reading and OPI gains.

When limiting the examination of skill gains to novice-level students within the Russian academic-year programs, reading and listening gains were found to increase together at the novice level. For advanced academic-year participants, Pearson correlations showed moderate, statistically significant positive correlations for delta (reading) and delta (OPI), r = 0.2423 and p = 0.0449; and for delta (listening) and delta (OPI), r = 0.2592 and p = 0.0315. Whether examined via parametric or non-parametric correlations, delta (listening) and delta (OPI) showed a slightly stronger relationship than delta (reading) and delta (OPI) among advanced participants. Earlier research has noted a relationship between listening and OPI among students of Russian at the advanced level and above (Davidson, 2010).

As a further exploration of the relationships among gains in different modalities, a set of distributions was generated comparing delta (OPI), delta (reading), and delta (listening) values among academic-year participants. In a comparison of different levels relative to one another, for delta (reading) and delta (listening), data for the intermediate-level cohort showed the greatest gains.

When the delta values for different skills of participants at a given level were considered, certain trends appeared: for the novices, delta (OPI) was a great deal higher than delta (reading) and delta (listening). For the intermediates and advanced, delta (OPI) was only slightly higher. For the advanced group, all the delta values were relatively similar; this close correspondence among delta values seems to mirror the consistent pattern of positive delta (skill) correlations seen among advanced participants in the overseas immersion program setting, a tendency towards the equalization of skill differentials in the context of the full immersion, acquisition-rich environment.

In addition to the examination of relationships among delta (skill) values, pre-program and post-program values across modalities were also analyzed. Possible cross-modal patterns in reading and listening were investigated via correlations of pre-program reading, pre-program listening, post-program reading, and post-program listening values among Russian academic-year participants. Of particular note, pre-program reading had a notably strong positive relationship with post-program listening ($r = 0.9412$ and $p < 0.0001$ with Pearson correlations, $rho = 0.9396$ and $p < 0.0001$ with Spearman correlations). Similarly, pre-program listening had a strong positive relationship with post-program reading ($r = 0.8969$ and $p < 0.0001$ with Pearson correlations, $rho = 0.8930$ and $p < 0.0001$ with Spearman correlations). The relationship between pre-program reading and post-program listening was slightly stronger than the relationship between pre-program listening and post-program reading.

For novice Russian academic-year participants, pre-program reading to post-program listening and pre-program listening to post-program reading were both positively correlated; pre-program reading to post-program listening was slightly more strongly correlated. For intermediate Russian academic-year participants, all categories were positively correlated. As with novices, pre-program reading to

post-program listening was slightly stronger than pre-program listening to post-program reading.

For advanced participants, in contrast, pre-program listening to post-program reading was positively correlated, while pre-program reading to post-program listening was not. As has been previously noted, "novice," "intermediate," and "advanced" participant categories have been delineated by pre-program OPI for testing purposes. While OPI levels serve as a good measurement of participants' overall L2 proficiency level, certain participants enter programs with relatively greater differences in a skill other than speaking, and, thus, may be seen to straddle category borders from a cross-modal testing perspective. Upper-level academic-year participants, who normally represent a greater period of previous study of the target language, have presumably experienced a broader range of instructional styles and a more diverse array of language-learning approaches by skill. As observed above in the analysis of delta (skill) values, advanced academic-year program participants show relatively similar and proportionate degrees of gain across all modalities while enrolled overseas.

A final series of analyses was conducted to examine the relationship of pre-program reading and pre-program listening values to post-program attainment as represented by post-program OPI values. For academic-year participants across all levels, both pre-program reading and pre-program listening were strongly correlated with post-program OPI whether examined via parametric or non-parametric correlations. For pre-program reading and post-program OPI, $r = 0.7896$ and $p < 0.0001$ with Pearson correlations, while for pre-program listening and post-program OPI, $r = 0.8041$ and $p < 0.0001$. Of the two pre-program skills in question, pre-program listening showed a modestly stronger correlation to post-program OPI.

For both novices and intermediates, pre-program listening and post-program OPI were highly correlated. Pre-program reading and post-program OPI were also correlated but fell short of the threshold for significance for either participant group.

From the point of view of the foreign language teacher or supervisor, the practical conclusion that flows from the relationship between pre-program listening and post-program OPI may be to recognize the importance of developing listening comprehension at the earliest stages of study. The observation of a correlation between listening comprehension and OPI gain at the intermediate level has not previously been reported in the literature.

14 Distributions as a Measurement of L2 Gain Across Modalities in Russian

As part of a consideration of gains in modalities beyond OPI, a set of distributions of Russian academic-year participants grouped by their pre-program reading levels was run. These included pre-program reading, post-program reading, and delta (reading) values for each skill-specific participant level as well as delta (OPI) and

post-program OPI. A second set of distributions was run with participants grouped by their pre-program listening levels; contents included pre-program listening, post-program listening, and delta (listening) values for each level as well as delta (OPI) and post-program OPI.

Fit group model analyses were conducted to examine the relationship of language gains across modalities. Pre-program values in all skills (reading, listening, and OPI) were leveraged to see how much they each accounted for gains as represented by post-program OPI results. When all three pre-program variables were examined jointly, parameter estimates were 0.02162 for pre-program reading ($p = 0.8376$), 0.1926 for pre-program listening ($p = 0.0970$), and 0.5487 for pre-program OPI ($p < 0.0001$). As expected, effects within the same modality were pronounced: pre-program OPI had the greatest effect on post-program OPI results, representing more than 50% of the variable portion explained by the model. Of the two cross-modal categories, pre-program listening approached the threshold for statistical significance and contributed a notable amount of the variable portion of the model. In contrast, pre-program reading did not contribute meaningfully to the overall effect. To further explore the strength of the effect of pre-program listening on post-program OPI results and tighten the model, pre-program reading was removed.

When the test was rerun with pre-program listening and pre-program OPI as the two independent variables, parameter estimates were 0.2094 for pre-program listening ($p = 0.0091$) and 0.5544 for pre-program OPI ($p < 0.0001$). Pre-program listening accounted for 21% of the variance in post-program OPI results, thus demonstrating a strong cross-modality effect.

15 Conclusions

The present study reports on L2 outcomes (measured changes in L2 proficiency levels in speaking, reading, and listening) of U.S. students ($N = 308$) who took part in year-long federally funded overseas immersion programs for Arabic, Chinese, and Russian. The subjects of the study were late adolescent and young adult learners, selected through a competitive process for participation in a group of well-resourced and carefully monitored year-long structured immersion programs at established host-country institutions in China, Kazakhstan, Moldova, Morocco, Russia, and Taiwan. The target languages in question represent a group of languages deemed "critical" for U.S. national security and economic interests by the U.S. government and considered typologically "difficult" (linguistically and in terms of time-on-task learning requirements) for English base-language learners (Thompson, 2014) in comparison to more commonly taught foreign languages, such as French, German, or Spanish.

The authors make no claim regarding the generalizability of these findings for study abroad programs, other than for those year-long models which have provided data for the present study. However, the notably high levels of language gain (rang-

ing from 4.1 to 7.1 standard deviations above the measured pre-program proficiency levels) reported here for both the early- and the late-stage students of critical languages have both policy and practical implications for the modern language profession and for all those concerned with preparing a new generation of graduates for a workforce in which professional-level language and intercultural skills are increasingly in demand (Brecht et al., 2015; Rivers, 2015).

The mean post-program proficiency levels (ACTFL/Advanced, CEFR-B2) demonstrated by the early-stage learners (SP) across skills are sufficient to ensure those students successful placement into advanced-level target-language courses offered at most U.S. universities (American Councils for International Education, 2017a; Bärenfänger & Tschirner, 2012). The mean post-program proficiency levels (ACTFL/Superior, CEFR-C1, C2) of the UP graduates represented in the study correspond to the professional language competencies required of those seeking employment in language-designated positions in many government agencies, as well as for those who expect to make use of their language skills in academia, business, research, international development, or domestic social services. Participants in both the early-stage (SP) and the advanced-level (UP) cohorts registered similar threshold-level L2 gains, regardless of the choice of critical language. In this context, it should be noted that while UP participants were required to meet an ILR-2 (ACTFL/Advanced) qualifying level in at least two skills at the time of application to the program, while early-stage learners were accepted at both the intermediate and novice levels of proficiency. Indeed, approximately one third of the entering students in SP reported no knowledge of the L2 prior to participation in the overseas programs.

Participant language gains are well-correlated across modalities. Advanced participants show concurrent gains across three skills: reading, listening, and speaking. Post-program reading and listening are strongly correlated, in turn, with pre-program listening skills. Initial levels of listening comprehension (pre-listening score) are positively correlated with growth in speaking skills at the intermediate and advanced levels, while reading ability, which functions as a proxy measure for more general levels of L2 structural and lexical control, is strongly associated with gains in speaking and in listening abilities, as the student progresses from novice to intermediate and to the advanced levels.

Of further note in the present study is empirical evidence of a process of *cross-skill equalization* as learners progress to the advanced and superior levels, despite notable early-stage skill gaps at the novice and intermediate levels among these groups. (Heritage learners are not included in the present study.)

Established practice within the foreign language field has focused on the value of study abroad for American L2 students who have completed one to three years of prior formal study, either in school or at the university level. The practice is understandable if study abroad is viewed as a one-time, relatively expensive intervention in (or enhancement of) the student's domestic undergraduate learning career. However, the latest survey/census of K-16 foreign language enrollments in the United States unfortunately confirms that no more than 20% of pupils currently have access to foreign language classes in U.S. school districts, while fewer than 7% of those who attend college enroll in a foreign language course (Brecht et al., 2013;

Brecht et al., 2015; American Councils for International Education, 2017b; Goldberg, Looney, & Lusin, 2015). Hence, requirements for prior study of the language as a prerequisite for study abroad exclude far too large a segment of the U.S. population to meet minimal standards of fairness and equal opportunity, even when issues of cost are put aside. The present study provides evidence of the notable language learning success that U.S. students of all backgrounds and with little or no prior study of an L2 can achieve in the overseas structured immersion context. Within the course of one year, students acquire levels of functional proficiency that can be put to immediate use in academia, service sectors, internships, and in their future careers.

Study abroad is a recognized "high-impact" practice in U.S. higher education (Kuh, 2012, 2016), and language-empowered study abroad can produce substantial linguistic gains for late-adolescent and young-adult learners across modalities, as demonstrated here, gains not typical in most domestic settings (Carroll, 1967; Kuh, 2012, 2016; Tschirner, 2011). In light of the declining rates of U.S. undergraduate participation in longer-term, language-focused study abroad (noted above), the present study offers further empirical support for the recent call by the American Academy of Arts and Sciences to foreign language departments, study abroad advisors, and institutional leaders to expand opportunities for language study at all levels in the context of institutionally approved education abroad activities, supported as well by the major federal initiatives aimed at preparing a new generation of linguistically and culturally competent U.S. professionals.

Appendices

Appendix 1 (SP)

NSLI for Youth Eligibility Requirements

www.nsliforyouth.org

NSLI-Y programs offer intensive language immersion in a variety of locations around the world. Scholarships are available for students to learn the following languages: Arabic, Bahasa Indonesia Chinese (Mandarin), Hindi, Korean, Persian (Tajiki), Russian, and Turkish.

Programs may take place in the following locations: China, Estonia, India, Indonesia, Jordan, Korea, Latvia, Moldova, Morocco, Russia, Taiwan, Tajikistan, Turkey and other locations around the world.

Eligibility Requirements

- U.S. citizen
- Grade point average (GPA) of 2.5 or higher on a 4.0 scale, or the equivalent

- 15–18 years of age at start of program (birthdate between July 10, 1999 and June 10, 2003 for summer programs; birthdate between September 20, 1999 and June 30, 2003 for academic year programs)
- Enrolled in high school (including home school)
- Not an immediate family member of an employee of the U.S. Department of State who works in the Youth Programs Division of the Bureau of Educational and Cultural Affairs or an employee at a NSLI-Y administering organization whose duties involve the NSLI-Y program
- Have not previously traveled outside the U.S. on a long-term (more than eight weeks) program sponsored by the Bureau of Educational and Cultural Affairs, Department of State
- Previous NSLI-Y *summer* program participants or participants of ECA-funded short-term programs are only eligible to apply for a NSLI-Y academic year program.

Previous language study is not a requirement. Students of all levels of language ability are encouraged to apply.

The NSLI-Y program seeks applicants who represent the diversity of the United States. Students of all racial, ethnic, religious, gender identities, sexual orientations, and socio-economic backgrounds are welcome to apply, as are students with disabilities.

Appendix 2 (UP)

The Language Flagship Capstone Program

www.thelanguageflagship.org

The Flagship Capstone full-year immersion is open to all Domestic Flagship undergraduate students who are committed to attaining professional or superior-level language proficiency through an intensive language training program tailored to their professional interests and academic specialization. It may occur during the third, fourth, or fifth year of a student's undergraduate program. The model also assumes and encourages that, in addition to full-year study, students will complete an additional period of immersion overseas to accelerate their language learning.

Applicants should have a strong academic record, a demonstrated interest in advancing their Arabic, Russian, Persian, Chinese, Hindi/Urdu, Korean, Portuguese, and Turkish skills and using these languages in their future career, and a desire to share their understanding of this language and culture within the larger community.

Undergraduate Applicants

All students who are enrolled at one of the Domestic Flagship Programs and reach the required proficiency level ILR-2 in their language on an Oral Proficiency Interview (OPI) and at least on one of the online modalities (reading, listening, writing), while scoring no lower than level ILR-1+ on the remaining two online modalities, are accepted to the Overseas Program, upon recommendation of the Overseas Project Directors.

Russian Overseas Flagship Post-BA or "At-Large" Applicants

The Russian Overseas Flagship Program accepts qualified applicants who did not participate in a Domestic Flagship Program and already have a bachelor's degree. The participants are selected on the basis of their language skills, academic merits, previous experience of study abroad, and ability to demonstrate how advanced Russian skills are going to help their career plans. At-large applicants to the Russian Overseas Flagship Program must either possess a B.A. degree or expect to receive one before starting the program. Successful applicants who are not heritage speakers must have completed at least three years of language at a college level and must have participated in a language study program in a Russian-speaking country for at least six weeks.

References

ACTFL. (2012). *ACTFL proficiency guidelines 2012*. Alexandria, VA. Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

American Academy of Arts and Sciences. (2017). *America's languages: Investing in language education for the 21st century*. Cambridge, MA: Author. Retrieved from https://www.amacad.org/content/Research/researchproject.aspx?i=21896

American Councils for International Education. (2017a). *National examination in world languages (NEWL)®*. Proficiency score levels and placement recommendations: https://www.americancouncils.org/services/testing-and-assessment/NEWL

American Councils for International Education. (2017b). *National K-16 foreign language enrollment survey.* N. Garas, Project Director. Retrieved from https://www.americancouncils.org/sites/default/files/FLE-report-June17.pdf

Bachman, L. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*(2), 149–164. https://doi.org/10.1017/S0272263100007282

Baker-Smemoe, D., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Variables affecting L2 gains during study abroad. *Foreign Language Annals, 47*(3), 464–486. https://doi.org/10.1111/flan.12093

Bärenfänger, O., & Tschirner, E. (2012). *Assessing Evidence of Validity of Assigning CEFR Ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by computer (OPIc) (Technical Report 2012-US-PUB-1)*. Leipzig, Germany: Institute for Test Research and Test Development.

Bazarova, S., Lekic, M. D., & Marshall, C. (2009). The online proficiency-based reading, listening, and integrated writing external assessment program for Russian: A report to the field. *Russian Language Journal, 59*, 59–78.

Brecht, R., Davidson, D., & Ginsberg, R. (1995). Predictors of foreign language gain during study abroad. In B. F. Freed (Ed.), Second language acquisition in a study abroad context (pp. 37–66). Amsterdam, Netherlands: John Benjamins.

Brecht, R., Abbott, M., Davidson, D., Rivers, W. P., Robinson, J., Slater, R., & Yoganathan, A. (2013). Languages for all? In *The Anglophone challenge*. College Park, MD: University of Maryland.

Brecht, R. D., Rivers, W. P., Robinson, J. R., & Davidson, D. E. (2015). Professional language skills: Unprecedented demand and supply. In T. Brown & J. Bown (Eds.), *To advanced language proficiency and beyond: Theory and methods for developing superior second-language ability* (pp. 171–184). Washington, DC: Georgetown University Press.

Carroll, J. B. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals, 1*, 131–151. https://doi.org/10.1111/j.1944-9720.1967.tb00127.x

Damari, R. R., Rivers, W. P., Brecht, R. D., Gardner, C. P., & Robinson, J. (2017, February 28). The demand for multilingual human capital in the U. S. labor market. *Foreign Language Annals, 50*, 13.

Davidson, D. E. (2010). Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals, 43*(2), 6–26.

Davidson, D. E. (2015). The development of L2 proficiency and literacy within the context of the federally supported overseas language training programs for Americans. In T. Brown & J. Bown (Eds.), *To advanced language proficiency and beyond: Theory and methods for developing superior second-language ability* (pp. 117–150). Washington, DC: Georgetown University Press.

Davidson, D. E., Garas, N., & Lekic, M. D. (2016). 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.

DeKeyser, R. M. (2007). Study abroad as foreign language practice. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 208–226). Cambridge, UK: Cambridge University Press.

Dewey, D. P. (2004). A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts. *Studies in Second Language Acquisition, 26*(2), 303–327. https://doi.org/10.1017/S0272263104262076

Dwyer, M. M. (2004). More is better: The impact of study abroad program duration. *Frontiers: The Interdisciplinary Journal of Study Abroad, 10*, 151–163.

Freed, B. F. (1998). An overview of issues and research in language learning in a study abroad setting. *Frontiers: The Interdisciplinary Journal of Study Abroad, 4*(2), 31–60.

Goldberg, D., Looney, D., & Lusin, N. (2015). *Enrollments in languages other than English at US institutions of higher education Fall 2013*. Modern Language Association. https://www.mla.org/content/download/31180/1452509/EMB_enrllmnts_nonEngl_2013.pdf

Golonka, E. M. (2000). *Identification of salient linguistic and metalinguistic variables in the prediction of oral proficiency gain at the advanced–level threshold among adult learners of Russian*. Unpublished dissertation, Bryn Mawr College.

Herzog, M. (n.d.). *An overview of the history of the ILR language proficiency skill level descriptions and scale*. Retrieved from http://govtilr.org/Skills/IRL%20Scale%20History.htm

Interagency Language Roundtable. (2016). *ILR skill level descriptions*. Retrieved from http://govtilr.org/

Kinginger, C. (2011). Enhancing language learning in study Abroad. *Annual Review of Applied Linguistics, 31*, 58–73.

Kramsch, C. (2014). Teaching foreign languages in an era of globalization: Introduction. *The Modern Language Journal, 98*(1), 296–311. https://doi.org/10.1111/j.1540-4781.2014.12057.x

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. London, UK: Longman.

Kuh, G. D. (2012). High-impact educational practices: What they are, who has access to them, and why they matter. *Peer Review, 14*(3), 29.

Kuh, G. D. (2016). *Study Abroad as a high-impact practice: Retrospective and prospective*. Retrieved from http://www.ifsa-butler.org/images/stories/pdf/advisors/Kuh_IFSA-Butler_6-18-2016.pdf

Liskin-Gasparro, J. E. (1984). The ACTFL proficiency guidelines: A historical perspective. In T. V. Higgs (Ed.), *Teaching for proficiency, the organizing principle* (pp. 11–42). Lincolnwood, IL: National Textbook.

Mason, L., Powers, C., & Donnelly, S. (2015). *The Boren awards: A report of oral language proficiency gains during academic study abroad*. New York, NY: Institute of International Education.

Murphy, D., & Evans-Romaine, K. (Eds.). (2015). *Exploring the US language flagship program. Professional competence in a second language by graduation*. Bristol, UK: Multilingual Matters.

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717–761). Malden, MA: Blackwell. https://doi.org/10.1002/9780470756492.ch21

North, B. (2006). *The common European framework of reference: Development, theoretical and practical issues*. Paper presented at the Japan-Europe International Symposium: A New Direction in Foreign Language Education: The Potential of the Common European Framework of Reference for Languages, Osaka University of Foreign Studies, Japan.

NSFLEP. (2015). *World-readiness standards for learning languages* (4th ed.). Alexandria, VA: The American Council on the Teaching of Foreign Languages (ACTFL).

Open Doors. (2016). *Report on international educational exchange online*. Retrieved from http://www.iie.org/Research-and-Publications/Open-Doors/Data

Ortega, L. (2009). *Understanding second language acquisition*. London, UK: Hodder Education.

Pellegrino Aveni, V. A. (2005). *Study abroad and second language use: Constructing the self*. Cambridge, UK: Cambridge University Press.

Powell, D., & Lowenkron, B. (2006). *National security language initiative.* Washington, DC: Office of the Spokesman. Retrieved from https://2001-2009.state.gov/r/pa/prs/ps/2006/58733.htm

Rivers, W. P. (2012). The unchanging American capacity in languages other than English: Speaking and learning languages other than English, 2000–2008. *Modern Language Journal, 96*(3), 369–379.

Rivers, W. P. (2015). *The contributions of language to the economic interests of the United States*. Prepared by the Joint National Committee for Languages (JNCL). https://www.amacad.org/multimedia/pdfs/TheContributionsofLanguagetotheEconomicInterestsoftheUnitedStates.pdf

Shaw, J. R. (2017). *Understanding language gain in the overseas immersion context: Multi-modal assessment of young adult learners of Arabic, Chinese, and Russian* (Doctoral dissertation). Bryn Mawr College, Bryn Mawr, PA.

Thompson, I. (2014). *Language learning difficulty*. Retrieved from http://aboutworldlanguages.com/language-difficulty

Tschirner, E. (2011). Reasonable expectations: Frameworks of reference, proficiency levels, educational standards. *Studies in Applied Linguistics* (1), 101–119. (Revised English version of Vernünftige Erwartungen: Referenzrahmen, Kompetenzniveaus, Bildungsstandards, 2008, *Zeitschrift für Fremdsprachenforschung, 19*(2), 187–208).

USED. (2008). *Enhancing foreign language proficiency in the United States: Preliminary results of the National Security Language Initiative*. Washington, DC: U.S. Department of Education. Available at https://nsep.gov/sites/default/files/nsli-preliminary-results.pdf

Vande Berg, M., Connor-Linton, U. J., & Paige, M. R. (2009). The Georgetown Consortium project: Interventions for student learning abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad, 18*, 1–75.

Watson, J. R., Siska, P., & Wolfel, R. L. (2013). Assessing gains in language proficiency, cross-cultural competence, and regional awareness during study abroad: A preliminary study. *Foreign Language Annals, 46*(1), 62–79.

Winke, P., & Gass, S. (2018). When some study abroad: How returning students align with the curriculum and impact learning. In C. Sanz & A. Morales-Front (Eds.), *Handbook of study Abroad research and practice* (pp. 527–543). New York, NY: Routledge.

Wothke, W. & Petersen, K. (2017). *Technical specifications. American Councils Language Assessment Systems (ACLASS)*. Available at https://www.americancouncils.org/technical-reports

Dan E. Davidson is Director of American Councils Research Center (ARC) in Washington, D. C. and Emeritus Professor of Russian and Second Language Acquisition at Bryn Mawr College, USA. He has written extensively in the fields of language, culture and educational development, including a 20-year longitudinal study of L2 acquisition during study abroad. His recent study, "Assessing Language Proficiency and Intercultural Development in the Overseas Immersion Context," appeared in *Exploring the US Language Flagship. Professional L2 Competence by Graduation* (Multilingual Matters).

Jane Robin Shaw is Lecturer within the Department of Russian and the Russian Language Institute of Bryn Mawr College, USA, where she defended her Ph.D. dissertation, "Understanding Language Gain in the Overseas Immersion Context: Multi-Modal Assessment of Young Adult Learners of Arabic, Chinese, and Russian," in 2017. Dr. Shaw's scholarly interests include late adolescent and young adult SLA in the immersion context, the articulation of domestic and overseas L2 study, the concept of "language difficulty," and the L2 assessment.

Part IV
Instructors and Learners

Language Instructors Learning Together: Using Lesson Study in Higher Education

Beth Dillard

Abstract The post 9/11 context brought a heightened awareness of the critical need to develop translingual and transcultural competence in language learners. This chapter takes up the question of what role—and what form—professional development for language instructors can take in the overall task of increasing students' language proficiency levels. It details a qualitative, interventionist study which examined how participation in an inquiry group mediated the conceptual development of three world language instructors in higher education. The study is framed by both activity theory, which informs an understanding of the inquiry group's situatedness in their sociocultural-historical context, and microinteractional analysis, which allows a view into how the turn-by-turn construction of meaning in the inquiry group created affordances for teacher inquiry. The findings of this study support the view that a combination of periodic workshops and sustained instructional inquiry groups can be particularly effective in promoting teacher conceptual development.

Keywords Professional development · Inquiry group · Lesson study · Developmental work research · Activity theory · Higher education · World language · Microinteractional analysis · Proficiency · Teacher learning

1 Introduction

In considering the task of building language learners' proficiency levels, a central concern, from my perspective as a language teacher educator, is the question of how to *continually* develop the pedagogical expertise of language *teachers* (MLA, 2007). The present study is situated broadly within the question of how professional

B. Dillard (✉)
Woodring College of Education, Western Washington University, Bellingham, WA, USA
e-mail: beth.dillard@wwu.edu

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_13

development can be leveraged to support the ongoing revitalization of language teachers, particularly in regards to their understanding of and ability to teach for proficiency.

In agreement with Rifkin's argument (this volume) that *The World-Readiness Standards for Language Learning* "help us as a field move away from an exclusive focus on the teaching of grammar, while providing instructors with a framework in which to purposefully construct lessons focused on using the target language…", I ask what role and form professional development might take to maximize teachers' productive use of that framework.

Professional development often takes the form of the one-shot workshop. Yet even the most intentionally designed workshop, characterized by a multidirectional flow of ideas and opportunities for practice, *can* be limited in long-term impact. I am not arguing that this is always the case, simply that it can often be the case. Instruction and inspiration, while crucial, are alone insufficient; the implementation of new pedagogies—like teaching for proficiency—must be scaffolded and supported over time if they are to become resilient elements of a teacher's practice.

The project I discuss in this chapter was borne out of my questions about how to design professional development in ways that might accomplish this goal: that of building new and resilient elements in a teacher's practice. In this qualitative, interventionist study, I used a combination of cultural-historical activity theory (CHAT) (Engeström, 2015) and a derivation of Developmental Work Research (a CHAT-inspired methodology) (Engeström, 2009) to make sense of how participation in an inquiry group mediated conceptual development for three world language instructors in higher education. Specifically, I asked: How do elements of a multilingual language instructor inquiry group serve to mediate language teacher conceptual development within the broader sociocultural context? Using both content and microinteractional analysis, I examined mediating means along a continuum between turn-by-turn construction of meaning and the surrounding sociocultural-historical context. In this chapter, I discuss several elements of this inquiry group that served to mediate language teacher conceptual development. These included: engagement with conflicting pedagogical concepts in discussions, structure and dynamics of those discussions, direct and indirect observation of each other's teaching, and meta-reflection mediated by transcripts of previous group meetings.

1.1 Cultural-Historical Activity Theory

To examine the various mediating means of language teacher conceptual development in this particular inquiry group, I drew on the theoretical framework of cultural-historical activity theory (CHAT). Rooted in the sociocultural tradition, CHAT describes a dialectical linking between individuals and society; CHAT examines how individual agency interacts with *seemingly* fixed socioeconomic and political structures (Engeström, 2009). CHAT, ultimately, provides a way of theorizing how the complex elements in an activity system afford and constrain the

goal-directed activity of individuals and groups (Cole & Engeström, 1993; Engeström, 2009; Sannino, Daniels, & Gutiérrez, 2009). These affordances and constraints include not only mediating means (both material and symbolic), but importantly the current and historical community context of the individual, the rules governing behavior (both spoken and unspoken), and the power structures functioning in the environment (Engeström 2009; Johnson & Golombek, 2011).

CHAT directly informed my methodological decisions in this study. I took an interventionist approach, using my own derivation of a CHAT-inspired methodology: Developmental Work Research (DWR) (Engeström, 2009). In DWR methodology the researcher first uncovers contradictions that exist in and between the various activity systems inhabited by participants. The researcher then mirrors those contradictions back to participants in order to stimulate a heightened awareness of the shared, culturally-mediated activity. This mirroring is called the "first stimulus." After mirroring these contradictions back to the participants, the researcher then introduces a new symbolic or concrete tool (the "second stimulus") into the system (Engeström, 2015). For participants, the second stimulus serves as a mediating means to help them address contradictions in their system(s). For activity theorists, the second stimulus allows mediation to be observed at the microgenetic level. In the case of this study, informal meetings between the instructors and myself functioned similarly to a first stimulus, and lesson study was utilized as a second stimulus. To my knowledge, only one other study (Tasker, 2014) has combined CHAT, DWR, and lesson study in the context of foreign language learning in higher education; in that study Tasker completed a lesson study cycle with three EFL teachers in the Czech Republic. Using grounded content analysis, Tasker (2014) identified five major findings: (1) "decision-makers" must be actively involved in professional development if there is to be institutional change, (2) outside experts must take on a more active, longer-term role, (3) EFL teacher professionalization should include participation in professional development activities, (4) lesson study can serve as a viable 'second stimulus' in DWR methodology, and (5) sociocultural theory provides a theoretical foundation for understanding how teachers learn through participation in lesson study (Tasker, 2014, p. iv). Methodologically, Tasker's study serves as an illustrative example of how DWR and lesson study can work synergistically to serve the needs of both teachers and theorists. What his work did not do, and what the present study aimed to accomplish, was document how this framework might also be useful in promoting and tracing teacher learning in diverse, multilingual groups of teachers who neither teach the same language nor even necessarily work within the same administrative structure.

1.2 Teacher Inquiry Through Lesson Study

Lesson study (*jugyou kenkyuu*) is a form of teacher inquiry originating in Japan over 100 years ago (Lewis, 2006; Lewis & Tsuchida, 1998; Stigler & Hiebert, 1999; Yoshida, 1999). This unique approach to teacher professional development became

popular in North America beginning in 1999; though taken up across disciplines and contexts, lesson study has been most enthusiastically received in elementary mathematics (Fernandez & Chokshi, 2002; Fernandez & Yoshida, 2004; Lewis, 2006). Lesson study brings teachers together to identify a problem of practice, collaboratively study that issue, and then create a "research lesson" applying ideas gleaned from that process. The research lesson is then taught to a live group of students as the other teachers observe. The process concludes with group reflections on *student* learning during the lesson (Yoshida, 1999; see also Lewis & Hurd, 2011; Stigler & Hiebert, 1999). The goal of lesson study is that the one lesson serves as a vehicle for teachers to explore their research goals (Fernandez & Yoshida, 2004, p. 7). In large part, this can be accomplished because lesson study requires a persistent focus on *student* learning (rather than *teacher* actions) throughout the process.

2 Methods

2.1 Context of Study

Over the course of one academic year, I worked with an inquiry group composed of myself and three female, non-tenure-track world language instructors from a research-intensive university in the Midwestern region of the United States; Hinata and Yukiko taught Japanese, Amina taught Arabic (all pseudonyms). All three women were native speakers of the language they taught, originally from countries speaking those languages. They had a wide range of experience, ranging from Hinata's four and half years of teaching, to Amina's eleven and Yukiko's twenty-four. Though all had entirely or primarily taught in higher education, all had also received K-12 training.

At the time of the study, there was a college-wide focus on building student language proficiency. This attention to proficiency spurred both renewal of existing professional development programs and creation of new opportunities. The women in this study were members of supportive programs and were already actively involved as learners and leaders in various professional development initiatives within their language programs and across the institution. They were active participants in college-wide workshops, and had also attended weeklong, intensive institutes organized by the university's Title VI National Language Resource Center. Finally, and concurrently with this study, both Yukiko and Amina took on leadership roles in an advisory board tasked with designing professional development for language instructors across the college. In sum, the participants in this study were already actively engaged in the development of their teaching practice before joining this study's inquiry group. With this in mind, I wondered how membership in an *ongoing* effort, like an inquiry group, might layer onto their existing participation.

Over the course of one academic year, the inquiry group in this study met seven times (see Table 1). We began meeting together in the Fall term to exchange ideas

Table 1 Overview of sessions

	Session	Date	Content
1st stimulus	1	10/29/2014	Informal meeting
	2	2/6/2015	Informal meeting
	3	2/25/2015	Informal meeting
2nd stimulus	4	4/1/2015	Formal beginning to modified-lesson study cycle
		4/6/2015	Video of Amina's lesson sent to group[a]
	5	4/10/2015	Debrief of Amina's lesson[a]
		4/13/2015	Observation of Hinata's class
	6	4/20/2015	Debrief of Hinata's lesson
	7	5/1/2015	Meta-reflection

[a]Not included in the current chapter

and provide collegial support to one another. The group came together in an organic way; it was the instructors, not the researcher, who invited each other. Early meetings were informal and unstructured. For example, at one meeting, Amina and Yukiko brought an Integrated Performance Assessment (IPA) that they were designing and asked the group for feedback. Over the course of these first three sessions, we got to know each other in ways that functioned similarly to the first stimulus in a DWR cycle. The organic nature of the group's origin was invaluable toward building trust within the group. Indeed, the genesis of this group is consistent with how Wenger describes the evolution of communities of practice in institutional settings: "Because communities of practice are organic, *designing them is more a matter of shepherding their evolution* than creating them from scratch. Design elements should be catalysts for a community's natural evolution" (Wenger, McDermott, & Snyder, 2002, p. 51 emphasis added). In this spirit, I inhabited the roles of researcher, facilitator, and "outside advisor;" the latter role especially was itself a mediating means central to both the DWR and lesson study frameworks.

The introduction and adapted use of lesson study served as the second stimulus in this DWR-derived intervention cycle. I introduced the idea of using a modified form of lesson study as a model for our work going forward, and we discussed how to modify it for our context. The most obvious challenge we anticipated was that the women did not share an instructional language or level. For this reason, the group decided that they would adapt lesson study and not collaboratively create a shared lesson; instead, they would focus their work on observing each other's teaching and together considering how to build and sustain student engagement.

Having uncovered various contradictions during the first three sessions, the last four sessions were devoted to using lesson study as a mediating means to explore some of the uncovered contradictions. The group completed two partial inquiry cycles, the first focused on Amina's teaching, and the second focused on Hinata's; this chapter is an examination of the second cycle. During this second cycle, we observed Hinata teach a 50-min lesson, gathered a week later to debrief her lesson, and finally, met 10 days after the debrief for a meta-reflection. The meta-reflection

was stimulated by participant reading and discussion of the transcript of the debrief of Hinata's lesson.

2.2 Data Analysis

I analyzed the data using content and microinteractional analysis. I focused first on the sociocultural and sociolinguistic context using a CHAT-informed content analysis (Miles, Huberman, & Saldaña, 2014) of interviews and field notes. Deductive coding was informed by the Activity System Observation Protocol (ASOP), an analytical tool informed by CHAT and designed to guide researchers looking at activity documented in their fieldnotes (Lewis & Scharber, 2012).

Having analyzed and described the broader context, I then focused my analysis on elements of the inquiry group serving as mediators of language teacher cognition. The data that informed this stage of analysis were: the videorecording of Hinata's lesson, audiorecording of the debrief session after that lesson, researcher notes, materials used in Hinata's lesson, Hinata's written reflection, and finally, a presentation that Hinata, Yukiko, and I had prepared at the conclusion of the inquiry group's work. I coded data deductively for moments of contradiction and mediation, and then inductively coded those moments in order to makes sense of what was happening in (and as a result of) those conversations. Finally, I used microinteractional analysis to examine how the *structure* of the group conversations, especially during these moments of contradiction, was itself a mediating means in teacher development. Using detailed transcriptions of salient moments (Jefferson, 2004), I examined turn-taking patterns, including cooperative interruptions (Liddicoat, 2011; Sacks, Schegloff & Jefferson, 1974; Schegloff, 2000).

3 Findings

3.1 Mirror Data: Uncovering Contradictions

Over the course of the three initial meetings, the group's conversations began to revolve around common tensions. Two fundamental contradictions emerged, one related to using the textbook as a tool, and the other related to gaining and keeping student engagement.

The instructors discovered that they shared a sense of dissonance between textbooks designed with no particular context in mind, and their own need to meet the specific learning needs of students in the context of their classroom. This contradiction between curricular design and implementation is widely shared by teachers of all disciplines across both K-12 and higher education. Specific to world language education, Guerrettaz and Johnston (2013) documented how an instructor creatively

leveraged a textbook in the "ecology" of the language classroom to support student learning in ways the textbook author could not have predicted. This research has been praised by language materials experts (Garton & Graves 2014). Just as in Guerrettaz and Johnston's (2013) research, the instructors in this study had both the space and knowledge to skillfully adapt the content of their textbooks to the ecology of their classrooms and the goals outlined in the ACTFL standards. They experienced this contradiction in an expansive way, empowered to make professional decisions about the implementation of their curriculum.

The second contradiction that emerged in these discussions focused on student engagement, which was referenced as an implicit criterion for decision-making in lesson planning and curricular choices. By "student engagement," the instructors seemed to picture students who were active, cheerful (as read through facial expressions, laughing), and diligent (studying outside of class). There was also the assumption that an engaged student would use the target language as much as possible during class. That these qualities and behaviors would lead to higher levels of language proficiency was the implied goal; however, to have students enjoy the classes and the process of learning a new language was the directly spoken goal. The women agreed that planning with student interest in mind went a long way toward the end of "student engagement." Concrete examples of this type of planning emerged during the initial sessions. For example, Hinata and Yukiko talked during an early meeting about how student interests had driven the design of their IPA unit. The contradiction in this case revolved around the question of *how* to leverage textbooks in curriculum design in ways that might increase student engagement and ultimately proficiency.

3.2 Lesson Study as a Mediating Artifact

In both DWR and lesson study, participants need to identify a problem space where they want to focus their energy. In the case of DWR, uncovered contradictions within and between activity systems inform the choice of this problem space; in lesson study, it is teachers' perceived *gaps in student learning* which guide the inquiry. The women in this study chose to focus on student engagement; in particular, they decided to interrogate *how to leverage their textbooks in curriculum design in ways that might increase student engagement.* This chapter focuses on how the group took up this salient problem space during the second teaching observation, the debrief of that observation, and the meta-reflection on the debrief.

An illustrative example of how the group took up work within this problem space can be seen in how they talked about engaging students in vocabulary learning. Below, I describe the salient mediating means utilized (implicitly and explicitly) by the instructors as they made sense of promoting vocabulary learning. I conceptualize these mediating means as falling into three overlapping categories which related to: the content of the conversations, the conversational structure (i.e., turn-taking and cooperative interruptions), and the methods of lesson study (i.e., observation of

others, both as disruptive to one's own experiences and pedagogical training, and as suggestive of new possibilities; meta-reflection mediated by transcripts of previous meetings).

Engagement with Conflicting Pedagogical Concepts At the debrief, an initial and powerful observation made about Hinata's teaching was the high percentage (90%+) of target language used by Hinata and her students during class. Sparked by this observation, the ensuing conversation centered on vocabulary learning in relation to authentic materials and target language usage. This makes sense; in order to use authentic materials and the 90%+ target language usage which ACTFL advocates, teachers must accept that *students can make sense* of language input *they haven't explicitly been taught.* In Excerpt 1 below, the women take up this dilemma.

Excerpt 1: "Using Target Language"

Source: Hinata's Debrief/Time stamp: 00:05:22-00:07:15

1	Amina:	But I I see the students also like using the target language|and [you said this is=
2	Hinata:	[ah::::::::::::::|
3	Amina:	=<u>first</u> class to teach this topic|[so I'm curious to know did you (.) like teach
4	Hinata:	[mm mm
5	Amina:	= the vocabulary be↑fore|or you give them a sheet to study at ↑home|or
6		[anything like that?|
7	Hinata:	[yeah:::::::::: so we have (.) vocabulary sheet right? ↑|And=
8	Amina:	=so they study at home the vocabulary and then they come ready for the topic?|
9	Hinata:	n:::::: (h) ((laughter)) [I would eh say::::: (.) not always.|You know like=
10	Yukiko:	[not always ((laughter))|
11	Amina:	=okay.|
12	Hinata:	=there some really serious students who do preparation at home=|
13	Amina:	=okay.|
14	Hinata:	and then they know already like [(.) what vocabulary they use in cla[ss|
15	Amina:	[okay| [okay|
16	Hinata:	but I I would say like maybe half of the students haven't prepared yet (.)|but you
17		know uh the the activity that I did was like just using I used my textbook,| and
18		the vocabulary is also in the textbook too,|
19	Amina:	okay|
20	Hinata:	and then the first, um=
21	Amina:	=but I mean, if the [vocabulary in the textbook, do they know the meaning?=|
22	Hinata:	[yea[h| mm::=
23	Amina:	= like, what [is the meaning?|
24	Hinata:	[yeah actually this (.) textbook has the:: >you know like
25		eh< English and Jap[anese.|
26	Amina:	[oh, okay, English and [Japanese.v
27	Hinata:	[Both on the same page|[so they=
28	Amina:	[okay okay
29	Hinata	=can you know, like look back.| And, <u>yeah</u>, although that this topic was first

introduced on that day that I was demonstrating, but, mmm, for their warm-up activity they are, they were not you know like uh you know like uh expected to use new vocabulary.

Beth: mm mm mm=

Hinata: =even though I was introducing like what's social network and then what does it mean to your life. And, but you- they can use like you know already learned vocabulary like so (.) yeah

At the beginning of the excerpt, Amina introduces a question for the group's consideration: *How is it that Hinata's students are able to use the target language on the first day of a new unit?* In this question, she tests her assumption that students would need to learn the vocabulary explicitly through the use of vocabulary sheets with direct English translations. In the lines that follow, Hinata confirms that her students did indeed receive vocabulary sheets, but adds complexity to this response, explaining that the target language usage the inquiry group had observed at the beginning of the class only required students to use known vocabulary. A few minutes later, and in response to my follow up question on how well she thought students understood her, Hinata adds further complexity to her description of what vocabulary teaching and learning look like in her classroom.

Excerpt 2: "For example, I say"

Source: Hinata's Debrief/Time stamp: 00:07:35-00:08:20

Hinata: right and then I'm not controlling my use of vocabulary. You know like I sometimes you know use obvious you know like the you know students the words that students might not kno[w, obvious[ly.

Beth: [mm mm mm [mm mm [mm

Hinata: [but I just you know anyway I £use £it.[1] But like, you know, um if students know the eh you know like important words. For example I say like "please listen" and then something. So, "listen carefully" and then like *chuui shite kiite kudasai* and then if that carefully part cannot be understood, but student might know that oh teacher want us to listen to,

Amina: I think it's like um like they get used to a routine, that's why they understand, yeah

Group: ((various sounds of agreement: "ah:::" "yeah yeah yeah" "right"))

In this excerpt, she explains that she does not "control (her) use of vocabulary" and sometimes uses "words that students might not know." She then gives a concrete example of the phrase "*chuui shite kite kudasai*" (literally, "listen carefully

[1] The symbol £ indicates laughter while talking (as distinct from laughter apart from words). Jefferson, G. (2004). Glossary of transcript symbols with an Introduction. In G. H. Lerner (Ed.), *Conversation Analysis: Studies from the first generation* (pp. 13-23). Philadelphia: John Benjamins.

please") to argue that, though students might not understand *chuui shite* (carefully), they could still grasp the more frequently used *kite* (listen) and *kudasai* (please). Through this example, she asserts that language learners do not need to understand every word that they hear or read.

Twenty minutes later, the question of how students could have made sense of *words that hadn't been explicitly taught* returns, this time in the context of discussing an authentic text Hinata had used in her lesson. The focus of the observed lesson had been on friendship and social networking. One goal Hinata had was to introduce students to the popular Japanese messaging service LINE (similar to Facebook Messenger or WhatsApp). She accomplished this by having students examine four charts displaying statistics related to the various social media (e.g. LINE, Twitter, Facebook) used by Japanese college students. Just before Excerpt 3 below, and after commenting that students had struggled with this activity, Hinata asked us if, "even with kind of limited ability to read, do you think it's still kind of effective?" Amina responded that it depended partially on the goal of the task, saying: "like what information they need to find or this graph is about." Hinata then translated for us exactly what the questions were asking. For example, she explained that the first question asked "What kind of social network Japanese college students used." At this point an individual in the group wondered aloud about Hinata's decision not to define new, potentially confusing vocabulary on the handout. Would doing so have made the activity, based around an authentic material, inauthentic? Excerpt 3 displays the conversation that followed.

Excerpt 3: "100% authentic versus modified version"

Source: Hinata Debrief/Time stamp: 00:29:23-00:31:60

1	Hinata:	well, yeah that's my kind of, the tension between using 100% authentic versus
2		modified version
3	Beth:	well, so, you and probably Yukiko as well could best understand what students
4		were saying. How do you feel based on what they were saying. How do you
5		sense what their comprehension was? Do you feel like this was something that
6		they mostly got? or were really confused about? or...and if confused, where did
7		you sense the barriers?
8	Hinata:	mmm. so first two graphs, those are simple, it's just like listing up, like
9	Yukiko:	in social networking
10	Hinata:	so these are simple, but the second and third one, it is actually asking like. this
11		one is how often do you use facebook? and these are kind of tricky—because it
12		says, I don't use it
13	( ):	mmm
14	Hinata:	yeah and then they don't know that word, so only Chinese students could
15		understand
16	Beth:	could understand it
17	Yukiko:	and also like eh LINE LINE LINE is like a some Japanese, mostly Asian know
18		probably, I don't know myself so the thing is like eh I think that Hinata just
19		present this one first and then explain what LINE is ( ) later and she was saying

20		I'm going to explain later.
21	Beth:	mmm
22	Yukiko:	I don't know, was it, probably it'll be better to talk about LINE first
23	Beth:	mmm
24	Yukiko:	because LINE use
25	Hinata:	ahh
26	Beth:	that's an idea
27	Yukiko:	I know like you want like eh critical thinking you know this thing they come up
28		with oh okay something like social networking and particularly like Japanese or
29		Asian populations. But I think it's too much probably, probably it's better to just
30		say, it's in Japan and there is one more thing, like listing up, I think there's
31		something like uh maybe have students what kind of social networking
32		resources

Hinata explicitly names the surfacing contradiction in line 1: "Well, yeah that's my kind of, the tension between using 100% authentic versus modified version." In doing this, Hinata opens up the dilemma for deeper inquiry.

In response to my prompting in lines 3–7, Hinata then goes on to describe the "trickier" elements of the charts that might have been barriers to learner comprehension. For example, in line 14 she points out a particular kanji that only Chinese students, able to use their knowledge of Chinese characters as context clues, would have been able to make sense of. In line 17, Yukiko also points out a possible area of confusion: the application "LINE" is likely unfamiliar to the non-Asian students in the class. Yukiko then transitions the conversation from a focus on identifying problems to suggesting changes. Between lines 17 and 20, she suggests that it would have been better to tell students from the beginning of the activity that "LINE" is a popular texting application in Japan. She asserts in line 29 that the inquiry-based approach that Hinata took, where students would discover this information through analyzing the charts, was "too much probably."

This tension (providing authentic input vs. scaffolding or modifying the input) surfaces again later in the interaction; Hinata responds (line 1) with the honest statement that she's not confident she strikes the right balance.

Excerpt 4: "I'm not 100% sure"

Source: Hinata Debrief/Time stamp: 00:38:50-00:40:50

1	Hinata:	oh yeah. (.) I'm still, as a teacher, I'm not 100% sure which one is better↑
2	Beth:	mm hmm
3	Hinata:	so we're doing integrated performance assessment and then for the IPA part
4		they, uh we don't put any assistance (.) you know, like
5	Group:	ah::::
6	Beth:	[mhm
7	Hinata:	[so I wanted to practice↑ and then get, [y'know, students used to [this
8	( ):	[ah:::: [oh:: kay

9	Hinata:	because in the real world they don't have
10	Beth:	right
11	Hinata:	like an £English word ((laughter)) They have cell phone to check out

Beginning in line 1, Hinata explicitly names the tension (just as in Excerpt 3, line 1: "I'm still, as a teacher, I'm not 100% sure which one is better↑" She goes on to explain her rationale (lines 23–31, with backchannels removed for ease of reading):

> So we're doing integrated performance assessment and then for the IPA part they, uh we don't put any assistance (.) you know, like, so I wanted to practice↑ and then get, y'know, students used to this, because in the real world they don't have like an £English word ((laughter))

In line 11, the conversation takes an unexpected turn when Hinata presents a counter argument to the claim she has just made (that students don't have access to English translations in the real world). In line 11, Hinata asserts that they *do* have that access in the real world, through use of their cell phones, introducing the interesting possibility that using digital technology to look up English translations *is* actually an authentic practice. Still, having students wrestle with texts in order to discover the meaning of new words takes more class time than either using a vocabulary sheet with predefined words or allowing device usage. The tension resonates with the group, and shortly thereafter in the conversation, there are multiple, overlapping affirmations.

The tension unresolved, only 5 min later the inquiry group goes back to the question: how are students making sense of *words they haven't been explicitly taught*? If students don't look up the meaning of the word, how is it that they figure out the meaning? In Excerpt 5 below, Hinata and Yukiko both provide examples of how they work with students through the target language to figure out the meaning of new kanji.

Excerpt 5: Building on known kanji

Source: Hinata Debrief/Time stamp: 00:45:30-00:47:29

1	Hinata:	[yeah they ask (.) for example, they don't
2		know this kanji↑ and then they ask the meaning of it, but I said, like "oh you
3		know this negative, so something about negative"
4	Beth:	mm::::::
5	Hinata:	and then this is, actually I £gave £them £an £answer £right £away. "this means
6		to [use" so they (don't [u-]
…		
7	Yukiko:	[cause they know like the kanji
8		for use. Yeah, they learned the kanji for use
… *One of the women wonders what language students use to ask questions in Hinata's class*		
9	Yukiko:	oh maybe we have to just go like uh first, we know this kanji, and we know this
10		kanji, and just [go through it, like [okay ( ) you end up getting authentic=

11	Beth:	[mm:::
12	Hinata:	[ah::
13	Yukiko:	=material, you can just uh you can recognize some kanji and grammatical
14		forms, you can go through with whole class as you say, uh:: (.) and £then
15	Beth:	yeah, see [where they go may[be

Hinata, in lines 1–3 describes how she scaffolds student understanding by helping them make sense of context clues in the sentence, and even *within* the unknown kanji by looking at radicals. Yukiko then provides another example, again explaining that she would talk students through each of the kanji that they *did* know, in order to try to guess the meaning of the unknown kanji through context.

In sum, these excerpts show Engeström's theory at work; moments of expansion and growth are stimulated by contradiction (2009), leading to shifts, at least in thinking, if not also in action. The content of the inquiry group's conversations shows that they wrestled with contradictory ideas and evidence about how students make sense of new, not explicitly taught language. These ideas and evidence came not only from the recent observation of Hinata's class, but certainly also from their wider sociocultural-historical experiences. For example, Amina's coursework in language pedagogy and Hinata's training in creating an IPA are evident in their comments. Put differently, observing Hinata's class formed a productive contradiction by introducing a new and disruptive mediating means into the instructors' existing, socioculturally-created system. The conversations that resulted from this disruption in the system show that the women tried to reconcile these contradictory ideas and evidence, leading to changed interpretations and understandings of their teaching practice.

Conversation Structure and Dynamics Microinteractional analysis revealed the salience of *how* the women engaged in conversation with each other over contradictory ideas and evidence. The *structure and dynamics* of their conversations were important mediators of the group's ability to productively wrestle with contradictory ideas and evidence. Let's revisit Excerpt 1 from above.

In Excerpt 1, and in particular in lines 1–36, the conversation is an active back and forth, complete with overlaps and interruptions, between Amina and Hinata. Amina, in particular, energetically pursues her question in a way that, at first reading, seems to cut off Hinata and not give her a chance to speak. Coding the excerpt for turn taking patterns,[2] however, shows that the overall trajectory of talk is preserved, and that Amina's interruptions function to clarify Hinata's meaning; thus the *interruptions are cooperative in nature*. More specifically, Hinata is giving Amina "conditional access to the turn;" that is, Hinata, sometimes in the middle of a turn of talk, yields her turn to Amina for the purpose of clarifying meaning. This is one of

[2] In this case, I coded for Turn Relevant Places (TRP). A TRP marks the place where it would be acceptable for a speaker change to take place. It is marked in the transcripts as |

the four categories of overlapping speech which Schegloff argued *does not need repair.* In contrast, *un*cooperative interruptions would be marked by shifts in the overall trajectory of talk that are felt as competitive; these, in Schegloff's argument, *would* need repair (Schegloff, 2000).

Notice how Amina talks past three TRPs before ceding the floor to Hinata at her TRP at the end of line 6. In line 7, Hinata anticipates the start of her turn and begins the utterance "yeah:::," but holds it until Amina completes her turn with "…anything like that?" In line 7, Hinata begins a response to Amina's question, stating that "we have (.) a vocabulary sheet…," however Amina seizes on Hinata's TRP and attempts to clarify what students do with that vocabulary sheet. In line 8 Amina says: "…so they study at home the vocabulary and then they come ready for the topic?" Here Amina seems to be testing an assumption that the students would need to memorize the vocabulary before being able to use it in the context of a class activity. Between lines 9 and 18, Hinata is able to elaborate on her explanation relatively uninterrupted; however, Amina plays an active role by adding in five "okay" continuers in lines 11, 13, 15 (2x), and 19. These continuers, as well as Amina's overlap and retaking of the floor in line 8, are primarily cooperative in nature; they function to continue the talk in the same direction the main interlocutor, Hinata, is taking it. This remains true, but takes on a different tone in line 21. Amina retakes the floor as Hinata pauses with an "um," saying: "*but I mean*, if the vocabulary in the textbook, do they know the meaning? Like, what is the meaning?" On the one hand, "but I mean" functions to redirect the conversation, ever so slightly, by implying that what Hinata is saying is *not* addressing Amina's question. On the other hand, it also functions to move the conversation as a whole to a deeper mutual understanding of different ways of teaching vocabulary; for this reason, the interruption is cooperative in the broader sense. In response to this clarifying question, Hinata states in lines 24–25 that the textbook has both English and Japanese. It is at this point, on line 26, that Amina finally seems satisfied with Hinata's response: "oh, okay, English and Japanese."

Conversations characterized by cooperative interruptions are mediating, because they facilitate the co-construction of meaning. In the prior section I concluded that the inquiry group's conversations showed that they wrestled with contradictory ideas and evidence; here I argue that *cooperative interruptions* help explain *how* they were able to productively discuss these contradictory ideas and evidence. A group of individuals cannot co-construct meaning if that group cannot maintain productive trajectories of talk; this is the case even, and especially, when there is confusion and/or disagreement. The conversation can mediate *deconstruction* of ideas, and crucially, it should, if it is to spur development and co-construction of new knowledge; however, conversational structure cannot *itself* degenerate and still be a mediating tool.

Methods of Lesson Study Perhaps the most powerful mediating means for instructor development in this study was observation of teaching, which plays a central role in lesson study. Through direct observation of one another's classes, the instructors' own training and teaching experiences came into contact with what they observed

their colleagues doing in the classroom. Observation of one's *own* teaching through videos is also a powerful mediator. Observation of others, whether direct or imagined, and observation of self through video thus had the power to deconstruct previously fixed ideas about teaching, as well as construct new ways of teaching. At the same time, observations of teaching carry the potential of providing inspiration as well. Observing Hinata's teaching not only served as a disrupting force, but also as inspiration, providing ideas for possible new ways of teaching.

One example of how observation opening up new possibilities came at the very beginning of the debrief, in a conversation primarily between Amina and Hinata about target language use in the classroom. Earlier I discussed this conversation (Excerpt 1) in light of how the structure of conversation, characterized by productive disagreement, served as a mediating means; here I revisit Excerpt 1 to examine how the teachers' observation of Hinata's lesson introduced new ideas about teaching into their conversation.

Amina opens this portion of the conversation with a clear statement of what she observed in Hinata's class: "I see the students also like using the target language and you said this is first class to teach this topic" (lines 1 and 3). Here the 'and' in line 1 functions more as a 'but,' as the illocutionary force of her statement is to put what she observed (target language use) into contrast with what Hinata said (first class of new unit). In the conversation that followed, Amina iteratively refined her question to find out whether students received English translations of vocabulary or not; Hinata responded to Amina's questions, eventually satisfying Amina with the information that, yes, students' textbooks did have English and Japanese.

What is significant here is how observing Hinata's class spurred this and other conversations in the first place. Though prior to observing Hinata's class the women had talked about the teaching and learning of vocabulary, these conversations had been theoretical in nature; observing Hinata and her students using the target language grounded the conversations in a real sense of what was possible.

4 Discussion

The transformative potential of these mediating means in this particular inquiry group can be best interpreted through the lens of Grossman, Wineburg, & Woolworth's (2001) distinction between a community and a pseudocommunity. In a pseudocommunity individuals "behave *as if* we all agree" (Grossman et al., 2001, p. 955). Indeed,

> the maintenance of pseudocommunity pivots on the suppression of conflict. Groups regulate face-to-face interactions with the tacit understanding that it is against the rules to challenge others or press too hard for clarification. This understanding paves the way for the *illusion of consensus*. (p. 955)

In contrast, a "mature *community* is [willing] to engage in critique in order to further collective understanding" (p.980, emphasis added).

The inquiry group in this study manifested the characteristics of a more mature community, as defined by Grossman et al., through the mediating means of discussing *conflicting pedagogical concepts.* In looking at the conversational content, it's evident that the women did not shy away from pedagogical questions that genuinely challenged each other's thinking. They discussed their differences of opinion about target language use, and about implicit versus explicit vocabulary learning. Even more importantly, they brought their different perspectives on these topics to bear on the discussion in productive ways. This ability to publicly disagree afforded the women opportunities for conceptual development by bringing new evidence into conflict with existing interpretations (Horn, 2010).

In addition to the *content* of the conversations, microinteractional analysis reveals how the *structure and dynamics* of the conversations mediated the group's ability to productively wrestle with such contradictory ideas and evidence. In particular, the inquiry group's conversations were characterized by cooperative interruptions. As indicators of active, engaged conversation, cooperative interruptions stand in contrast to "an interactional congeniality … maintained by a surface friendliness," which marks a pseudocommunity (Grossman et al., 2001, p. 955). The women in this study were comfortable critiquing ideas, introducing counterevidence, or persisting in calls for clarification. The individuals in the group pushed each other to speak in specifics rather than abstractions. Doing so in a conversational structure that was respectful, affirming and supportive allowed assumptions to be tested, differences in understanding to become apparent, and ultimately, contradictions—and opportunities for conceptual development—to bubble to the surface (Grossman et al., 2001).

Another critical mediating means for conceptual development was *direct and indirect observation of each other's teaching.* Although the prior experiences of the women influenced the ideas they brought to our group conversations, *observing* each other's teaching supported conceptual development by providing invaluable input of both confirming and contradictory evidence into the discussions. Directly observing one another afforded all members of the inquiry group "transparent access to colleagues' practices," a prerequisite to learning within a community of practice (Levine & Marcus, 2010, p. 396). Excerpt 1 is an illustrative example of this, as the observation of Hinata's use of target language encouraged Amina to inquire into Hinata's particular way of promoting vocabulary learning. In this case, direct observation of teaching practices served to disrupt thinking about teaching and contributed to the mediation of conceptual development.

The experience of being directly observed was a powerful variation of this third mediating means for Hinata. In the debrief conversation Hinata reexamined her teaching practice as she was asked to explain her rationale for certain pedagogical moves and make sense of feedback from her colleagues. Further, in her reflections on being observed, Hinata shared that the feedback she received—because it came from peers she *trusted*, was shared in a *comfortable environment*, and had the *concreteness* and *embeddedness* of a specific observation—enabled her to think deeply about her practice. Hinata's experience in the inquiry group stands in contrast to what is possible in one-shot workshops, which in their singular nature cannot

develop these types of long-term, trusting relationships. In sum, participating in direct observation (of self and others) was a productive tool for this group of women, in large part because the inquiry group was a mature community.

The present study introduced a mediating means in the form of *meta-reflection mediated by transcripts of previous group meetings.* I am not aware of any studies that have documented this particular use of transcripts. This kind of meta-reflection was shown to be a productive element to add to lesson study; reading and reflecting on transcripts of previous group meetings ('debriefs') proved to be useful mediators of conceptual development. Through reading and discussing the debrief transcript, the women were able to (re)view their own comments, "hearing" them as if they were outside parties to the conversation. DWR calls on researchers to act as "outside experts" to mirror back emergent contradictions to participants; the process of (re) viewing the transcripts served a similar function for the women. Hinata talks about how this process enabled her to gain additional insights from the initial debrief conversation than she had in the moment.

This study found that the sociocultural context, specifically the supportive environments of the women's programs, was the most salient mediating means of their conceptual development. In fact, all other mediating means in this study were predicated upon the women's membership in this sociocultural context *which afforded choice, experimentation, and innovation.* Hinata and Yukiko's work with Integrated Performance Assessment (IPA) units is one of the clearest examples of this; the freedom afforded them allowed Hinata to *choose* to attend a professional development seminar about Assessment (including the IPA), and to *experiment* with using IPA units in her course. These IPA units not only transformed Hinata and Yukiko's teaching, they inspired ideas of *what was possible* within the inquiry group.

5 Conclusion

The findings of this study support the view that a combination of periodic workshops and sustained instructional inquiry groups can be particularly effective in promoting teacher conceptual development. Hinata's comment below eloquently summarizes the synergistic relationship:

> So, from … [institutional workshops] I get knowledge. For example, last year we learned [exploratory practice] and IPA and I took IPA classes … last summer, so I got knowledge. And kinda like there I start thinking about how I can you know implement what I have learned into my own teaching. And in the small group like we have or more smaller, like smaller even smaller, with Yukiko, I kinda, those places are um like good good ones to kinda think more you know about how I can implement those like knowledge into your teaching. (Hinata's interview)

It would be tempting for me to argue that language teacher professional development should *only* take the form of small, long-term inquiry groups like the one in this study. The time that the group spent together was not only professionally fruitful, but personally rewarding. Yet this study has shown that the workshops and insti-

tutes which the women had attended were also integral to their conceptual development, because they introduced new ideas and pedagogies that the women could later explore and try out. For example, Hinata and Yukiko likely wouldn't have been experimenting with IPA units without the college-wide focus on building student language proficiency, and had they not previously attended summer institutes about assessment and content-based instruction. At the same time, though, Hinata's reflections suggest that without her partnership with Yukiko, and her involvement in the inquiry group, she might not have been able to implement the content of the summer institutes to the same extent. In sum, it was the synergistic pairing of workshops/institutes with sustained involvement in a small group of colleagues which afforded the teachers in this study a space to experiment with proficiency-based teaching in their classrooms. Thus, I would argue, in addition to *more* professional development in higher education (see Malone, this volume), we need that professional development to be respectful of and responsive to the rich variety of experiences that teachers contribute to their own learning, as well as that of their colleagues.

References

Cole, M., & Engeström, Y. (1993). A cultural-historical approach to distributed cognition. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 1–46). Cambridge, UK: Cambridge University Press.

Engeström, Y. (2009). Expansive learning: Toward an activity-theoretical reconceptualization. In K. Illeris (Ed.), *Contemporary theories of learning: Learning theorists in their own words* (pp. 53–73). Abingdon, UK: Routledge. https://doi.org/10.4324/9780203870426

Engeström, Y. (2015). *Learning by expanding: An activity-theoretical approach to developmental research* (2nd ed.). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9781139814744

Fernandez, C., & Chokshi, S. (2002). A practical guide to translating lesson study for a U.S. setting. *Phi Delta Kappan, 84*(2), 128–134. https://doi.org/10.1177/003172170208400208

Fernandez, C., & Yoshida, M. (2004). *Lesson study: A Japanese approach to improving mathematics teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Garton, S., & Graves, K. (2014). Identifying a research agenda for language teaching materials. *Modern Language Journal, 98*, 654–657. https://doi.org/10.1111/modl.12094

Grossman, P., Wineburg, S., & Woolworth, S. (2001). Toward a theory of teacher community. *Teachers College Record, 103*(6), 942–1012. https://doi.org/10.1111/0161-4681.00140

Guerrettaz, A., & Johnston, B. (2013). Materials in the classroom ecology. *The Modern Language Journal, 97*(3), 779–796. https://doi.org/10.1111/j.1540-4781.2013.12027.x

Horn, I. S. (2010). Teaching replays, teaching rehearsals, and re-visions of practice: Learning from colleagues in a mathematics teacher community. *Teachers College Record, 112*(1), 225–259.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp. 13–23). Philadelphia, PA: John Benjamins. https://doi.org/10.1075/pbns.125.02jef

Johnson, K. E., & Golombek, P. (2011). A Sociocultural theoretical perspective on teacher professional development. In K. E. Johnson & P. Golombek (Eds.), *Research on second language teacher education: A sociocultural perspective on professional development* (pp. 1–12). New York, NY: Routledge. https://doi.org/10.4324/9780203844991

Levine, T. H., & Marcus, A. S. (2010). How the structure and focus of teachers' collaborative activities facilitate and constrain teacher learning. *Teaching and Teacher Education, 26*, 389–398. https://doi.org/10.1016/j.tate.2009.03.001

Lewis, C. C. (2006). Lesson study in North America: Progress and challenges. In M. Matoba, K. A. Crawford, & M. R. Sarkar Arani (Eds.), *Lesson study international perspective on policy and practice*. Beijing, China: Educational Science Publishing House.

Lewis, C. C., & Hurd, J. (2011). *Lesson study step by step: How teacher learning communities improve instruction*. Portsmouth, NH: Heinemann.

Lewis, C., & Scharber, C. (2012). *Activity System Observation Protocol (ASOP)*. Bright Stars: Technology-mediated settings for urban youth as pathways for engaged learning. Research proposal submitted to the W.T. Grant Foundation (funded).

Lewis, C. C., & Tsuchida, I. (1998). A lesson is like a swiftly flowing river. *American Educator, 22*(4), 12–17, 50–52. Retrieved from: https://ncetm.org.uk/public/files/34863/swift+flowing+river.pdf.

Liddicoat, A. (2011). *An introduction to conversation analysis*. New York, NY: Continuum. https://doi.org/10.1111/j.1540-4781.2013.12024_4.x

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Thousand Oaks, CA: SAGE.

MLA ad hoc Committee on Foreign Languages. (2007). Foreign languages and higher education: New structures for a changed world. *Profession*, 234–245. https://doi.org/10.1632/prof.2007.2007.1.234.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735. https://doi.org/10.2307/412243

Sannino, A., Daniels, H., & Gutiérrez, K. (2009). *Learning and expanding with activity theory*. Cambridge, UK: Cambridge University Press.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society, 29*(01), 1–63. https://doi.org/10.1017/S0047404500001019

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap*. New York, NY: The Free Press.

Tasker, T. (2014). *Exploring EFL teacher professional development through lesson study: An activity theoretical approach*. Unpublished doctoral dissertation. Pennsylvania State University, Applied Linguistics.

Wenger, E., McDermott, R., & Snyder, W. M. (2002). *Cultivating communities of practice*. Boston, MA: Harvard Business School Press.

Yoshida, M. (1999). *Lesson study: A case study of a Japanese approach to improving instruction through school-based teacher development*. Unpublished doctoral dissertation. University of Chicago, Department of Education.

Beth Dillard is Assistant Professor of Second Language Acquisition at Western Washington University. She teaches courses in second language acquisition and pedagogy in the English Language Learning endorsement program. Her research interests include teacher learning, content and language integration, and the use of systemic functional linguistics to promote students' academic language development.

U.S. Foreign Language Student Digital Literacy Habits: Factors Affecting Engagement

Jeffrey Maloney

Abstract In today's academic contexts students are presented with a wide variety of digital technologies that present opportunities for authentic input and interactions with other speakers. Of special importance to consider is the cultivation of digital literacies and their connection with different factors (Guikema & Williams, 2014). Until now, there have been few attempts at linking extramural digital literacy practices with factors such as proficiency, study abroad experience or declared language majors. This study focuses on exploring students' digital literacy practices in the L2 and draws a connection with the level of daily practices, language proficiency, study abroad experience, and declared language major. A pre-test survey was created and given to roughly 600 American Spanish L2 students that elicited information about tech-use across two indices: technology for language learning (e.g., dictionaries, apps) and technology for entertainment (e.g., movies, social media) in the L2. Surveys were taken before completing ACTFL certified tests in reading, speaking and listening. Findings indicate significant correlations for language proficiency, declared language major and study abroad experience and reported levels of technology use in the L2. Findings are discussed in reference to how to improve student engagement via digital means.

Keywords CALL · Digital literacies · Study abroad · Proficiency · Spanish as a foreign language · ACTFL

1 Introduction

With the broad proliferation of technologies into everyday life, students today have access to multiple means to engage with the L2 for learning and entertainment. Recognizing this, SLA researchers have called for an increase of research on how these new technologies impact pedagogy and the language learning process

J. Maloney (✉)
Languages and Literature, Northeastern State University, Tahlequah, OK, USA
e-mail: maloneyj@nsuok.edu

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_14

(e.g., Chapelle, 2007, 2009; Garrett, 2009). Subsequent research has been dedicated to uncovering the benefits of specific technologies on language development (e.g., Lin, 2015), and there is plenty of work outlining the benefits of specific CALL tools (Álvarez Valencia, 2016; Bull & Wasson, 2016; Chen, 2013; Liu, Lu, & Lai, 2014; Peterson, 2016). Additionally, a growing body of research focuses on incorporating new technologies into language curricula (Burston, 2014; Celik, 2013; Chun, Kern, & Smith, 2016; Thorne & Reinhardt, 2008) and on how specific language skills benefit from the use of technology in the classroom (e.g. Grgurović, Chapelle, & Shelley, 2013).

While incorporating technology into the classroom and the impact of specific tools are primary foci of CALL research, Kern (2014) observed that new technologies also offer a means for students to continue language learning and socialization outside formal contexts. What has not been investigated is the relationship between reported levels of technology use and multiple learner variables such as language proficiency and study abroad experience. This chapter reports on the results of a larger study on student profiles and learning outcomes. I focus on three variables that may have a relationship with students' levels of reported technology use in informal contexts among Spanish learners at an American institution of higher education: language proficiency, study abroad experience and declared language major. Investigating which, if any, factors have a relationship with technology use in informal contexts can help inform pedagogical practice. Instructors can better leverage technologies so that language students can use them for language acquisition and socialization beyond the classroom.

2 Background

University language students in the United States have experience with many kinds of technologies and social media in their L1 (Lenhart, 2015; Pearson, 2015). With regards to the L2, however, researchers have expressed that foreign language classrooms do not easily encourage students to explore and engage with L2 technologies. Thorne, Black and Sykes (2009) observe that: "Despite the broad penetration of online tools, cultures, and literacies into many arenas of everyday life, L2 classrooms often remain bounded contexts providing limited opportunities for committed, consequential, and longer term communicative engagement afforded by new technologies" (p. 804). Some have argued that language pedagogies must shift to accommodate the changes taking place on the digital front (Lai & Gu, 2011). Researchers also note that new media require language educators to be proactive and responsive to the global changes new technologies prompt (Thorne & Reinhardt, 2008). There is a call for a language pedagogy that incorporates exposure and engagement with L2 culture and communities of speakers, and in many cases this can be achieved through digital means (Levy, 2009). While bringing technology into the classroom is a worthwhile pursuit, students' ability to effectively leverage new technologies in informal contexts for language acquisition and socialization is

important. Students only have a few hours a week during limited years of academic study to cultivate communicative competence in a L2. The researchers cited previously argue for a focus on enabling students to succeed with new technologies in the L2 in class, and to carry these skills into daily life. In this regard, there has been interest in the impact that technology-focused activities outside of class can have on language development, identity, and socialization. Major areas of interest include video games (Chik, 2014; Cornillie, Thorne, & Desmet, 2012; Peterson, 2012, 2016; Ryu, 2013; Sylvén & Sundqvist, 2012), social networks (Lin, Warschauer, & Blake, 2016; Reinhardt & Zander, 2011), and online interest communities or fan-fiction (Black, 2009; Thorne et al., 2009). Findings indicate that students have positive views of such activities, and research has found that information and communication technologies (ICTs) can be used to cultivate identities and experience socialization. This work is promising, but more remains to be done to understand which factors play a role in students' levels of engagement with different digital media and technologies in the L2. Researching different factors and their relationship with reported levels of use of digital technologies can improve pedagogical practice that can lead to learner autonomy and long-term engagement with the L2 (Godwin-Jones, 2011).

2.1 Digital Literacies and the L2

In my opinion, long-term communicative engagement and a "more sophisticated competence" is well served by cultivating digital literacies and engagement with digital practices in the L2. Lankshear and Knobel (2008) explained digital literacy as "a shorthand for the myriad social practices and conceptions of engaging in meaning making mediated by texts that are produced, received, distributed, exchanged, etc., via digital codification" (p. 5). Hafner, Chik, and Jones (2013) identified digital literacies as "the modes of reading, writing and communication made possible by digital media" (p. 1). Further, Meyers, Erickson and Small (2013) identified three research perspectives on digital literacies. The first is concerned with acquiring *information age* skills. The second is defined as *cultivating habits of mind*. The third, which is the focus of this chapter, is *engagement in digital cultures and practices*. While research that informs all three perspectives identified by Hafner et al. (2013) may be desirable, examining what factors have a relationship with students' levels of engagement in digital cultures and practices can provide a starting point into examining which factors have a relationship with technology adoption.

So far, I am familiar with one study that investigated different variables and their relationship with reported levels of technology use. White (2016) focused on student motivation and technology use outside of class and found that even highly motivated students did not engage with technologies in the L2. Beyond this, some research has focused on what students are doing with technology both inside and outside of the L2 classroom (Lai & Gu, 2011; Lai, Shum, & Tian, 2016; Sylvén &

Sundqvist, 2012). There have been survey studies focused on different factors impacting general technology use, such as age (Williams, Abraham, & Bostelmann, 2014), L1 practices (Trinder, 2016), and learner readiness (Winke & Goertler, 2008). Others have also reported on what students already do (Levy & Steel, 2015; Steel & Levy, 2013). These studies are discussed in more detail below.

2.2 Previous Survey-Driven Studies on Technology and L2

In a study on learner readiness for CALL instruction, Winke and Goertler (2008) surveyed over 900 foreign-language students to see how prepared they were for blended foreign-language instruction. Results showed that it should not be taken for granted that students will enter the language classroom prepared to use technologies typical of CALL courses, such as video or audio recording and uploading files. They also state that "regularly surveying the students will help teachers and administrators design appropriate tasks, harness new technologies students already use in their personal lives, and generate motivation for learning online" (p. 497). This observation highlights the need to understand what technologies students are using, which is a focus of this study.

Steel and Levy (2013) examined what technologies students were using inside and outside of class, comparing results from a 2011 survey with two studies from 2006. In Steel and Levy's survey, students rated what they felt were the most beneficial tools for language learning. At the top of the list were tools such as online dictionaries and translators, followed by online videos, social networks and devices dedicated to listening to music. Importantly, Steel and Levy examined whether students were using different technologies, not how often they used them or factors impacting the adoption of certain technologies. Beyond learner readiness and what students are already using, some research has focused on the impact of age on digital literacies development (see Lee, Yeung, & Ip, 2016).

Looking at factors affecting levels of technology use, Trinder (2016) surveyed 175 Austrian students to examine what influenced students' choices in using different technologies in English. She observed that students still preferred face-to-face (F2F) interactions but recognized the potential impact of different technologies on L2 learning. Students were more likely to utilize and find useful what she termed "traditional" forms of media like films and online videos. These technologies, she observed, were already integrated into students' daily lives, thus making it likely that students would use them in the L2. While Trinder's survey did account for frequency of use, her analysis focused on students' preferences, not the factors that could have affected levels of use and engagement with different technologies.

Most survey-driven studies focused on factors that impact the adoption of certain activities in support of formal learning. Thus, many of the studies have not investigated frequency of use or engagement in certain activities, or factors affecting frequency of engagement.

3 The Current Chapter

Motivations for using technology vary widely and may be difficult to quantify. Measuring the relationship that different variables have with levels of technology use, however, can provide insight into where classroom practice can help. As mentioned, I have focused on three different variables to begin with. The first, language proficiency, is an important measure when considering engagement with technologies and online communities. Proficiency could have important implications for students' motivations as well as perceived efficacy of different tools. Students with lower proficiency may not see much value in attempting to use different technologies outside of class, due to language barriers and inaccessibility of the materials. Higher proficiency students, however, may engage with technology practices in the L2 as they may have had more exposure to digital practices and media outlets in and outside of the classroom, thus encouraging their use. To my knowledge, Thorne and Reinhardt (2008) is one of the only studies that has examined technology use of advanced language learners. They provided a framework for activities that could be employed in class to encourage adoption and use of different technologies in the L2. Again, however, there is no work on what encourages use outside of class.

Having a declared major is also an important aspect to consider as this may be an indicator of motivation as well as the levels of input and socialization students receive throughout their college career (Sung & Tsai, 2014). It can be safely assumed that students who have declared a language major may be more motivated than non-majors to expose themselves to more outlets of input in the L2. Also, because of the declared major status students may have many more opportunities for exposure and interaction with resources and individuals that could promote technology use in both formal and informal contexts.

The third variable in this study, study abroad experience, has received more focus than the other two, but work has focused on how to leverage technology to enhance and improve student experience while abroad (e.g., Lee, 2011). In addition, much of the study abroad experience research (e.g., Kinginger, 2013) has not evaluated the effect that studying abroad had on students' technology habits overall.

This study is not concerned with competence in using individual tools, but with how often students utilize digital tools to engage in meaning making with texts and other forms of multimedia in the L2 outside of class. Thus, this chapter does not measure digital literacy per se, but is focused on different variables that influence students' engagement with activities that require literacy in digital, informal contexts regardless of the specific tool (Lea, 2016).

4 Methods

This study uses survey data, language proficiency ratings, and background information to examine the relationship of different variables with participants' reported levels of technology use in an L2, outside of the formal learning context. I focus on

the variables of language proficiency, study abroad experience and major status and their relationship with reported levels of usage of a wide variety of technologies and practices, as measured by a pre-proficiency exam survey.

I have formulated the following research questions to guide this study:

1. What are the levels of foreign language students' technology use across proficiency groups?
2. What relationship do study abroad experiences and declared language major have with the reported levels of technology use?

4.1 Participants

This study was conducted as part of an ongoing project at a large mid-western land-grant university in the United States beginning in 2015 and is still ongoing. For this chapter, I analyze the results of survey data and proficiency tests from Spring 2016. The participants for this study were drawn from the entire student population enrolled in Spanish foreign language courses at the university. Data were collected from students in the second semester through fourth year courses. Valid data from the survey resulted in a total of 617 different participants for the quantitative analysis, with 11 participants for semi-structured interviews (not reported on in this chapter). The participant pool consisted of 407 females and 198 males, with 12 identifying as *other* or not reporting. Compared to the student population at the institution (about 50% male and female institution-wide), the participant pool was disproportionately female.

4.2 Method

ACTFL Exams The university at which this study took place received a large grant from the Department of Defense to administer wide-scale proficiency exams that were developed and scored by the American Council on the Teaching of Foreign Languages (ACTFL). The three (speaking, listening and reading) tests were administered via computer in a large computer lab in which students could take the exams on a walk-in basis at their convenience. Each test took roughly fifty minutes to complete. Students could take the three exams in any order within a 1-month time span during the 2016 Spring semester.

The exam data are a part of a larger, ongoing project that includes multiple proficiency tests for reading, speaking and listening. Each of the participants received a rating on the ACTFL proficiency scale (ACTFL, 2012), which includes 11 total possible levels. For this chapter, I have collapsed all of the novice (novice-low – high), intermediate (intermediate-low – high) and advanced levels (advanced-low – high), in order to enable easier reporting and data analysis.

Survey I developed a set of two questions included in a larger pre-test survey that each participant took prior to the ACTFL proficiency exams:

1. How often are you currently engaged in the following activities in the target language (the language for which you are taking the test today) outside of class?
2. How often are you currently engaged in the following activities to assist you with the target language?

Each question provided a list of technologies/platforms and students were asked to report how often they used the listed items. We also asked students about where they learned about different technologies, such as from friends, instructors or whether they taught themselves. The survey also elicited information about the students' backgrounds, including heritage status, study abroad experience, year in the program and their reasons for studying the foreign language.

The survey items focused on students' technology use were rated on a 6-point Likert scale. I selected the technologies based on the body of previous work that examined students' technology use with input from other colleagues. As my survey items were included in a much larger survey, there was a limit to the number of items I could create and include. Based on the results from the survey questions, interviews were used to provide a better picture of some of the students' habits and attitudes.

Like Trinder (2016), I split the different technologies listed in the two questions into multiple categories. The first category, "Communication and Input/Content technologies," included platforms or devices that "facilitate one-to-one or one-to-many communication" (p. 89), such as social media or chat software, general web content, or stand-alone media generally used for entertainment or to look up information. The second category, which I have labeled "Discipline Specific Technologies" included activities focused on improving language proficiency, building vocabulary or using translation software. Table 1 shows how I have categorized the different questions for this analysis. The Likert scale assignments and their associated levels of frequency are listed in Table 2.

5 Results

What follows are the results from the survey questions, including descriptive and inferential statistics. The results were compared utilizing non-parametric tests as the data were not distributed normally. These quantitative results are supplemented by a brief discussion of the major themes that arose in eleven interviews with participants which are not reported on in this chapter.

For ease of reporting and analysis, I have chosen to focus on the results of the speaking exam only, although means for reported levels of technology use for all three of the exams are included and discussed in relation to overall findings. I focus on the speaking exam results because there were a larger number of valid exam scores for this modality compared to the other two.

Table 1 Survey items stratified across categories

Communication & input/content technologies	Discipline-specific (language learning) technologies
Communicating using technology	Using translation software
Using social media	Contacting other people via chat or text message
Discussion forums/sites	Visiting websites dedicated to L2 learning
Listening to music	Visiting online forums
Listening to news broadcasts or podcasts	Utilizing vocabulary building apps
Watching online videos (YouTube, Vlogs, Etc.)	Utilizing dictionary apps
Watching TV or movies	Utilizing general language learning apps
Playing video games	
Writing emails	
Using Social media	
Communicating with others using technology	
Visiting blogs	
Visiting general interest sites	
Visiting discussion forums (e.g. Reddit)	

Table 2 Likert scale rating & frequency assignment

Never	Once a month or less	A few times a month	Weekly	A few times a week	Daily
1	2	3	4	5	6

5.1 Instrument Reliability

The results of the questions in the Communication, Input/Content Technologies category are highly reliable, with a Cronbach's $\alpha = .93$ overall. The Discipline-specific technologies proved to be a bit less reliable, with a Cronbach's $\alpha =$ of .8. The question sets from the Communication Technologies and the Input/Content Technologies categories are considered "Excellent", while the Discipline-specific Technologies questions have a "Good" reliability (Kline, 2013).

5.2 Overall Results

The results of the proficiency exams are reported in Table 3. All participants did not complete each of the three of the exams, leading to numbers lower than the total of 617 participants for each test. The reading proficiency exam had the most evenly spread results for participants, with most rated in the intermediate range. For speaking, most students also received a rating within the intermediate range. The listening test had much fewer valid scores, and students tended to score in the novice range.

Table 3 Overall results of proficiency exam

Proficiency result	Reading	Listening	Speaking	Mean
Novice	142	231	177	183.33
Intermediate	258	217	329	268.00
Advanced	102	26	23	50.33
Total	502	474	529	

5.3 Research Question 1

To address the first question, I first examined the means for each of the question items, separated across the three different proficiency categories. Means for each group are reported in Table 4 for the Communication, Input and Content Technologies, and in Table 5 for the Discipline – Specific Technologies.

There was an increase in the reported levels of technology use across the different technologies included in the survey. In addition, a comparison of Tables 4 and 5 shows that overall, students were utilizing the Discipline-Specific technologies more often than the other category. The overall mean for the level of use for the Communication and Input/Content Technologies was 1.99 (1 meaning never, 6 meaning daily), meaning students reported using them less than once a month or almost never. For the Discipline-Specific Technologies, the reported average was 2.62. This means that on average, students were engaging with L2-learning technologies between once a month and a few times a month.

The Communication & Input/Content technology activity that was reported as most used by the participants was listening to music, with an average of 3.12, or just over a few times a month. The two that were reported as being the least used were visiting discussion forums and playing video games, with means of 1.49 and 1.42, respectively.

Due to the overall increase across the different technologies and proficiency exams, I ran Kruskall-Wallis (non-parametric ANOVA) tests for each of the questions to search for significant effects of language proficiency on the reported levels of tech use. The Kruskall-Wallis tests for each of the questions returned a significant result for the speaking test for all the Communication and Input/Content Technologies except for playing video games. This indicates that proficiency has a significant effect on the reported levels of engagement with Communication and Input/Content technologies for each technology included in the survey except video games. This shows a positive relationship: the higher the proficiency level, the higher the level of reported use of different technologies. The particular significant results are identified in Table 4.

For the Discipline Specific technologies, the results were more varied across the different proficiency groups. Specific results for the Speaking exam can be seen in Table 5. The Kruskall-Wallis tests returned significant results for each technology except for vocabulary apps and online dictionaries. It is interesting to note that for the use of translation software there is a negative relationship: students who received higher proficiency ratings reported less reliance on translation tools and software.

Table 4 Reported means of communications & input/content technology use

Technology	Reading	Listening	Speaking	Mean
Music				
Novice	2.33	2.50	2.29[ab]	2.37
Intermediate	3.00	3.25	3.05[bc]	3.10
Advanced	3.48	3.76	4.43[ac]	3.89
Podcasts/News				
Novice	1.52	1.58	1.45[ab]	1.52
Intermediate	1.75	1.73	1.72[c]	1.73
Advanced	2.02	2.76	2.48[c]	2.42
Online videos				
Novice	1.88	1.92	1.78[ab]	1.86
Intermediate	2.02	1.98	1.98[bc]	1.99
Advanced	2.25	3.00	3.62[ac]	2.96
TV/Movies				
Novice	1.93	1.98	1.89[ab]	1.93
Intermediate	2.20	2.27	2.17[bc]	2.21
Advanced	2.56	3.36	3.52[ac]	3.15
Video games				
Novice	1.39	1.28	1.29	1.32
Intermediate	1.30	1.31	1.29	1.30
Advanced	1.37	1.68	1.86	1.64
General interest sites				
Novice	1.39	1.39	1.39[c]	1.39
Intermediate	1.44	1.41	1.43[c]	1.43
Advanced	1.65	1.38	2.19[ac]	1.74
Blogs				
Novice	1.39	1.39	1.29[ab]	1.36
Intermediate	1.45	1.41	1.54[bc]	1.47
Advanced	1.68	2.16	2.00[ac]	1.95
Social Media				
Novice	1.73	1.85	1.68[ab]	1.75
Intermediate	1.98	2.02	2.00[bc]	2.00
Advanced	2.46	2.88	3.64[ac]	2.99
Discussion forums				
Novice	1.39	1.34	1.35[b]	1.36
Intermediate	1.38	1.38	1.40[b]	1.39
Advanced	1.51	1.73	1.91[a]	1.72
Communicating using technology				
Novice	1.83	1.96	1.73[ab]	1.84
Intermediate	2.31	2.33	2.28[bc]	2.31
Advanced	2.47	3.08	4.00[ac]	3.18

Writing emails				
Novice	1.51	1.62	1.42[ab]	1.52
Intermediate	1.80	1.83	1.97[bc]	1.87
Advanced	2.33	2.88	3.14[ac]	2.78

Note. x[a] indicates significant difference from intermediate. x[b] indicates a significant difference from advanced (novice to intermediate, intermediate to advanced). x[c] indicates a significant difference from novice

Table 5 Reported means of discipline-specific (Language Learning) technologies

	Reading	Listening	Speaking	Mean
Translator				
Novice	4.47	4.34	4.38[ab]	4.40
Intermediate	4.03	3.99	4.03[bc]	4.02
Advanced	3.95	3.69	3.09[ac]	3.58
Chat				
Novice	2.53	2.49	2.38[b]	2.47
Intermediate	2.53	2.53	2.57	2.54
Advanced	2.75	3.27	3.64[c]	3.22
Websites				
Novice	2.81	2.82	2.85[a]	2.83
Intermediate	3.13	3.20	3.20[c]	3.18
Advanced	3.36	3.42	3.00	3.26
Forums				
Novice	2.02	1.91	1.89	1.94
Intermediate	2.00	1.94	2.00	1.98
Advanced	2.02	2.35	2.36	2.24
Vocab				
Novice	2.67	2.61	2.57	2.62
Intermediate	2.58	2.55	2.54	2.56
Advanced	2.33	2.04	2.23	2.20
Dictionary				
Novice	3.15	3.13	3.04[a]	3.11
Intermediate	3.27	3.52	3.60[c]	3.46
Advanced	3.64	3.00	3.09	3.24
General apps				
Novice	2.60	2.62	2.51	2.58
Intermediate	2.46	2.26	2.35	2.36
Advanced	2.18	2.08	2.32	2.19

Note. x[a] indicates significant difference from intermediate. x[b] indicates a significant difference from advanced (novice to intermediate, intermediate to advanced). x[c] indicates a significant difference from novice

5.4 Research Question 2

The second research question focused on whether study abroad (SA) experience in a country where the language is spoken, or having declared a major in Spanish, had a significant relationship with the reported levels of technology use. Roughly 14% of all participants reported having declared a major in Spanish, while a total of 103 (18%) of the participants reported having studied abroad during their time as an undergraduate.

For the analysis of SA experience, I first separated the participants into SA and Non-SA groups. Descriptive statistics for the two groups are reported in Table 6**.** Participants who participated in SA report higher levels of engagement with each of the technologies in both categories, except for translation software/tools. The reported level of dictionary use is the only technology that had a lower reported level of frequency for the SA group and the non-SA group. In order to test for significant group differences, I ran Mann-Whitney tests for each question. The results of each test are reported in Table 6 along with the standardized test statistic *z,* effect size *r*, and significance *p*. Results of the Mann-Whitney test indicated that those participants who participated in SA report a significantly higher level of engagement with all the Communication and Input/Content technologies included in the survey. All technologies except communicating via text or chat for help and visiting language learning websites returned significant results. It is important to point out, however, that although the differences are significant, none of the different measures of technology use returned more than a small effect size.

To assess whether there would be differences between Spanish majors and non-majors, I again ran a Mann-Whitney test to check for differences between the groups for each question. For this analysis, I excluded those participants with study abroad experience to avoid potential confounds. This resulted in forty-five participants with a declared Spanish major and 455 participants that did not declare a Spanish major. The data were not normally distributed, so I ran a non-parametric Mann-Whitney test for significance. Results of the Mann-Whitney tests for majors and non-majors are contained in Table 7.[1] It can be seen that having a declared Spanish major has a significant effect on reported levels of engagement with different technologies outside of class.

[1]Year in school may confound reports on differences between language majors and non-language majors. Additional analyses were performed on the data controlling for those participants that were in their second year of classes. Results were still generally the same. However, a few of the reported levels of technology use were no longer significantly different between the two groups.

Table 6 Reported means of communications, content/input technologies and discipline specific technologies across SA groups along with Mann-Whitney test results

			Mann-Whitney test results			
	Mean	SD	*U*	*Z*	*r*	*p*
Communication, content/input technologies						
Listening to music						
Non-SA	2.79	1.54	32,021	4.52	0.18	>.001**
SA	3.57	1.61				
Listening to news or podcasts						
Non-SA	1.69	1.13	30,198	3.24	0.13	0.001**
SA	2.02	1.28				
Online videos						
Non-SA	2	1.24	29,821	3	0.12	0.003**
SA	2.38	1.41				
TV/Movies						
Non-SA	2.14	1.23	30,399	3.26	0.13	0.001**
SA	2.54	1.4				
Video games						
Non-SA	1.29	0.85	28,056	2.28	0.09	0.022*
SA	1.52	1.16				
Writing emails						
Non-SA	1.8	1.16	32,833	5.16	0.21	>.001**
SA	2.4	1.23				
Using social media						
Non-SA	1.9	1.38	31,974	4.48	0.18	>.001**
SA	2.49	1.57				
Communicate w/ technology						
Non-SA	2.12	1.4	29,635	2.83	0.12	0.005*
SA	2.57	1.57				
Visiting blogs						
Non-SA	1.44	0.98	30,493	4.04	0.16	>.001**
SA	1.87	1.33				
Visiting general interest sites						
Non-SA	1.41	0.95	30,402	3.89	0.16	>.001**
SA	1.84	1.3				
Visiting discussion forums						
Non-SA	1.35	0.87	29,918	3.67	0.15	>.001**
SA	1.73	1.22				
Discipline-specific technologies						
Translation tools						
Non-SA	4.14	1.3	24,029	−0.504	−0.02	0.614
SA	4.06	1.25				

Text or chat for help						
Non-SA	2.56	1.56	27,273	1.63	0.07	0.1
SA	2.82	1.55				
Language websites						
Non-SA	3.02	1.62	29,091	2.87	0.12	0.004**
SA	3.57	1.49				
Forums						
Non-SA	2	1.38	26,628	1.8	0.07	0.07
SA	2.23	1.41				
Vocabulary apps						
Non-SA	2.53	1.55	27,856	2.28	0.09	0.02*
SA	2.92	1.6				
Dictionary Apps						
Non-SA	3.25	1.73	31,416	4.54	0.19	>.001**
SA	4.14	1.58				
General LL apps						
SA	2.39	1.57	28,345	2.41	0.10	0.02*
Non-SA	2.78	1.54				

* = p < .05. ** = p < .01

6 Discussion & Implications

The first research question focused on the effect of proficiency on the reported levels of technology use outside of the formal learning context. Analyses showed that there is a significant effect for language proficiency on students' reported levels of use. This was the case for a wide variety of technologies across both categories of Communication and Content/Input and Discipline Specific Technology.

The positive trend notwithstanding, participants reported not engaging with different technologies more often than once every few weeks or so. They tended to report use of some forms of media such as music, television shows, or movies more often than others. This finding echoes what Trinder (2016) found in that students are more likely to incorporate the L2 into practices that they are already doing regularly in the L1, as may be the case here since most American teens are engaged in these activities (Lenhart, 2015). The most frequent activity that students reported engaging in was listening to music, followed by watching online videos. Of note is the fact that students reported using discipline-specific technologies more often than the other types. This is not surprising, as these software products, tools and platforms would be most easily accessible and available to them as language learners. Their use is also required by some instructors. More highly proficient learners reported

Table 7 Reported means of communication, content/input and discipline specific technologies across major & non-major groups

			Mann-Whitney test results			
	Mean	SD	*U*	*z*	*r*	*P*
Communication, content/input technologies						
Listening to music						
Non-major	2.68	1.49	14,593	4.81	0.22	>.001**
Spanish major	3.95	1.52				
Listening to news or podcasts						
Non-major	1.62	1.06	13,517	4.03	0.18	>.001**
Spanish major	2.46	1.55				
Online videos						
Non-major	1.92	1.19	13,457	3.93	0.18	>.001**
Spanish major	2.82	1.49				
TV/Movies						
Non-major	2.07	1.2	14,031	4.21	0.19	>.001**
Spanish major	2.87	1.34				
Video games						
Non-major	1.28	0.84	10,357	0.09	0.00	0.93
Spanish major	1.38	0.96				
Writing emails						
Non-major	1.7	1.102	14,904	5.61	0.25	>.001**
Spanish major	2.79	1.32				
Using social media						
Non-major	1.82	1.32	13,480	3.81	0.17	>.001**
Spanish major	2.72	1.75				
Communicate w/technology						
Non-major	2.05	1.37	13,659	3.88	0.17	>.001**
Spanish major	2.85	1.46				
Visiting blogs						
Non-major	1.42	0.96	10,970	1.72	0.08	0.09
Spanish major	1.69	1.13				
Visiting general interest sites						
Non-major	1.38	0.9	11,787	2.21	0.10	0.03*
Spanish major	1.79	1.34				
Visiting discussion forums						
Non-major	1.34	0.85	10,397	0.21	0.01	0.83
Spanish major	1.44	1				
Discipline-specific technologies						
Translation tools						
Non-major	4.18	1.28	8935	−0.807	−0.04	0.42
Spanish major	3.95	1.34				

Text or chat for help						
Non-major	2.5	1.56	12,359	3.17	0.14	0.002**
Spanish major	3.31	1.44				
Language websites						
Non-major	2.96	1.62	12,099	2.89	0.13	0.004**
Spanish major	3.74	1.29				
Forums						
Non-major	1.97	1.4	10,585	1.71	0.08	0.09
Spanish major	2.15	1.23				
Vocabulary apps						
Non-major	2.5	1.56	11,865	1.5	0.07	0.13
Spanish major	2.82	1.62				
Dictionary apps						
Non-major	3.17	1.7	12,632	3.48	0.16	0.001**
Spanish major	4.05	1.7				
General LL apps						
Spanish major	2.38	1.57	10,191	0.7	0.03	0.49
Non-major	2.51	1.57				

* = p < .05. ** = p < .01

using translation software/tools and vocabulary apps less than those participants who received lower proficiency ratings. This could reasonably be expected as students would need to rely less and less on these technologies as they improved their language capabilities, moving from translation tools/apps and dictionaries to relying more on their own skills, or they may not be using them in their courses. The participants' responses also showed a decrease in certain discipline-specific technologies, while all the other types showed a trend of increase. It could be argued that as students develop in the L2, they find more value in using different technologies outside of class. The finding that participants moved away from tools like translation software and vocabulary building apps also suggests a shift of focus: as students' progress, they rely less upon tools and technologies to assist with language production and comprehension and move more towards media consumption and tools more focused on communication. This could especially hold true for students that demonstrate higher levels of motivation to become socialized into the target language community. A second possibility is that students who used technology more frequently may be developing their proficiency more than those who did not. For example, although the participants in this study do not report playing many video games, research on the use of such games for L2 development has indicated that it can lead to language development (Sylvén & Sundqvist, 2012) and autonomous learning (Chik, 2014).

It is important to highlight here that even though use of technologies and technology habits did increase, it increased very little. Training programs dedicated to improving adoption and engagement with L2 technologies/platforms and subsequent measurements of L2 proficiency could answer the questions about the relationship between language proficiency and L2 ICT use. In addition, subsequent research why students used different technologies will help to shed more light onto ways instructors can encourage use.

Steel and Levy (2013) found that 65% of the students indicated using different technologies outside of class. What was not included is how often students used them. In this study, students indicated using different technologies roughly once a month, with most reporting below this rate. Students are engaging in each activity at least on some level, but none of the activities appeared to be integrated regularly into students' daily lives, even though they may be exposed to certain technologies and tools in class. This fact may bring into focus what some of the goals of foreign language curricula may be. Kramsch (2006) and Thorne and Reinhardt (2008) and others note that there is a need for language instruction to expose students to forms and genres that are not typically found within the traditional language classroom. Promoting long-lasting connections with the language and culture may be accomplished by encouraging students to engage regularly with certain digital literacy practices. For this study, the technologies included in the survey would be a reasonable place to start creating opportunities for students to become familiar with the offerings within the platforms in the L2. A question that should then be addressed is how often students should be engaged in these activities. Language students have limited time to receive formal instruction, with many university majors only requiring a few semesters of study. Equipping students with an understanding of how to operate digitally in the L2 may enable students to continue their development as speakers and individuals within the multiple communities that make our global society.

The second research question expands on the first. I examined two factors that could be related to the amount of time students engage with different technologies outside of classes. Participants who studied abroad in Spanish speaking countries report significantly higher levels of technology use than those who have not. What is still unclear is the nature of the relationship with study abroad and technology use. Are students who use more technology also more likely to study abroad? This certainly is a possibility and will be a focus of subsequent research.

Another possibility is that students who participated in study abroad became more acquainted with digital practices and forms of engagement. Research has found multiple benefits for participating in a study abroad experience (e.g. Tanaka & Ellis, 2003), and engagement with digital technologies may yet be another. Student engagement with the culture and the language would obviously benefit from spending meaningful time in a context where the language is spoken. Students who have the means and elect to participate in SA programs would (hopefully) have many more opportunities to make meaningful, personal connections -- promoting both the desire and capability to use technology to stay in touch with L2 speakers and the L2 community. I hope to uncover more about this with more longitudinal and qualitative data.

Something can also be said about the impact declaring a major in Spanish can have on levels of technology use outside of the formal learning context. It is certainly possible that engagement with technologies outside of class is impacted by the motivation to learn and to connect with others. Even when students that have studied abroad are controlled for, there is still a significant effect of having a declared Spanish major on reported technology use. An important question to address with this factor is similar to SA experience. Subsequent research will focus on uncovering the nature of the relationship between declaring a language major and engagement with technology in the L2.

A third research question (not reported on in this chapter) focused on what participants thought affected their technology usage habits in the L2. The reasons why different students do not engage with ICTs can be a combination of perceptions about their own L2 ability, their awareness of what is available, and different priorities or interests. Each of the perceived barriers could be overcome with training or changes in classroom culture. Hubbard (2013) argued that learner training is important for any technology and should not be taken for granted. Lai et al.'s (2016) finding that online training is effective for promoting autonomous use of language learning is also encouraging. Further, Lai's (2015) study speaks to the idea that when instructors and peers encourage student technology use, engagement improves. Future research will need to examine whether explicit training will increase engagement with technology in the L2. Learner training that focuses on exposing students to different social media outlets, news or general interest sites, blogs and discussion forums may help them gain a better understanding of the possibilities to engage with the target language and culture. Training may also focus on showing students that utilizing different technology is not time consuming and could be incorporated into daily routines. Instructors may demonstrate outlets that provide comprehensible linguistic input but also encourages longer-term commitments. As this would require extra effort on the part of instructors, research is needed that examines different strategies of adoption and incorporation as well as impact.

7 Conclusion

This study expands on previous work to improve understanding of the L2 technology habits of foreign language students. Spanish students with higher proficiency ratings do report higher levels of engagement with technologies in the L2 in informal contexts. This pattern also holds true for SA experience and declaring a Spanish language major. Understanding what factors impact students' habits with ICTs is a meaningful starting point for future research and pedagogy to develop action plans to improve. Many students do not participate in SA programs, and not every student will have a desire to pursue a degree in an L2. Those students who do not declare a Spanish major and who do not study abroad are not reporting the same levels of engagement with technology. Granted, this may also reflect language ability, but the

barriers identified by students also included not being aware of tools. Students who do not declare a Spanish major may not make using technology or engaging with the L2 outside of class a priority. Nevertheless, this is an area that merits further exploration before any definitive answers can be provided. Clearly, some groups of students in the FL curriculum receive differing amounts of exposure to practices involving digital literacy, something that could possibly be offset by shifts in pedagogical practice and cultures within the classroom.

It is important to note also that more in-depth analyses will offer a better understanding of what impacts the technology habits of students outside of class. I have provided an initial look at some of the factors and their relationships with reported levels of technology use. It is worth repeating that a major point for further exploration is the nature of the relationships between the factors that contribute to engagement in digital practices.

Longitudinal research designs and case study research would go far in understanding the nature of the relationship between technology use and the multiple factors investigated here. It is possible that combinations of identity and motivation, institutional and community beliefs, and personal experiences play an important role in the engagement with different technologies and practices in the L2. Understanding the nature of these relationships, and whether they may be impacted by training and shifts in practice is an important next step.

References

ACTFL. (2012). *ACTFL proficiency guidelines*. Alexandria, VA: ACTFL. Available from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Álvarez Valencia, J. A. (2016). Language views on social networking sites for language learning: The case of Busuu. *Computer Assisted Language Learning, 29*(5), 853–867. https://doi.org/10.1080/09588221.2015.1069361

Black, R. W. (2009). Online fan fiction, global identities, and imagination. *Research in the Teaching of English, 43*, 397–425. http://www.ncte.org/library/NCTEFiles/Resources/Journals/RTE/0434-may09/RTE0434Online.pdf

Bull, S., & Wasson, B. (2016). Competence visualisation: Making sense of data from 21st-century technologies in language learning. *ReCALL, 28*(2), 147–165. https://doi.org/10.1017/s0958344015000282

Burston, J. (2014). MALL: The pedagogical challenges. *Computer Assisted Language Learning, 27*(4), 344–357. https://doi.org/10.1080/09588221.2014.914539

Celik, S. (2013). Internet-assisted technologies for English language teaching in Turkish universities. *Computer Assisted Language Learning, 26*(5), 468–483. https://doi.org/10.1080/09588221.2012.692385

Chapelle, C. A. (2007). Technology and second language acquisition. *Annual review of applied linguistics, 27*, 98–114. https://doi.org/10.1017/S0267190508070050P

Chapelle, C. A. (2009). The relationship between second language acquisition theory and computer-assisted language learning. *The Modern Language Journal, 93*(s1), 741–753. https://doi.org/10.1111/j.1540-4781.2009.00970.x

Chen, X. B. (2013). Tablets for informal language learning: Student usage and attitudes. *Language Learning & Technology, 17*(1), 20–36. http://llt.msu.edu/issues/february2013/chenxb.pdf

Chik, A. (2014). Digital gaming and language learning: Autonomy and community. *Language, Learning & Technology, 18*(2), 85–100. http://llt.msu.edu/issues/june2014/chik.pdf

Chun, D., Kern, R., & Smith, B. (2016). Technology in language use, language teaching, and language learning. *Modern Language Journal, 100*(S1), 64–80. https://doi.org/10.1111/modl.12302

Cornillie, F., Thorne, S. L., & Desmet, P. (2012). ReCALL special issue: Digital games for language learning: challenges and opportunities. *ReCALL, 24*(3), 243–256. https://doi.org/10.1017/S0958344012000134

Garrett, N. (2009). Computer-assisted language learning trends and issues revisited: Integrating innovation. *The Modern Language Journal, 93*(s1), 719–740. https://doi.org/10.1111/j.1540-4781.2009.00969.x

Godwin-Jones, R. (2011). Emerging technologies: Autonomous language learning. *Language Learning & Technology, 15*(3), 4–11. http://llt.msu.edu/issues/october2011/emerging.pdf

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL, 25*(2), 165–198. https://doi.org/10.1017/s0958344013000013

Guikema, J. P., & Williams, L. F. (Eds.). (2014). *Digital literacies in foreign and second language education*. San Marcos, TX: Computer Assisted Language Instruction Consortium (CALICO).

Hafner, C. a., Chik, A., & Jones, R. H. (2013). Engaging with digital literacies in TESOL. *TESOL Quarterly, 47*(4), 812–815. https://doi.org/10.1002/tesq.136

Hubbard, P. (2013). Making a case for learner training in technology enhanced language learning environments. *CALICO Journal, 30*(2), 163–178. https://doi.org/10.11139/cj.30.2.163-178

Kern, R. (2014). Technology as Pharmakon: The promise and perils of the internet for foreign language education. *Modern Language Journal, 98*(1), 340–357. https://doi.org/10.1111/j.1540-4781.2014.12065.x

Kinginger, C. (2013). Identity and language learning in study abroad. *Foreign Language Annals, 46*(3), 339–358. https://doi.org/10.1111/flan.12037

Kline, P. (2013). *Handbook of psychological testing*. London, UK: Routledge.

Kramsch, C. (2006). From communicative competence to symbolic competence. *Modern Language Journal, 90*(2), 249–252. https://doi.org/10.1111/j.1540-4781.2006.00395_3.x

Lai, C. (2015). Modeling teachers' influence on learners' self-directed use of technology for language learning outside the classroom. *Computers & Education, 82*, 74–83. https://doi.org/10.1016/j.compedu.2014.11.005

Lai, C., & Gu, M. (2011). Self-regulated out-of-class language learning with technology. *Computer Assisted Language Learning, 24*(4), 317–335. https://doi.org/10.1080/09588221.2011.568417

Lai, C., Shum, M., & Tian, Y. (2016). Enhancing learners' self-directed use of technology for language learning: the effectiveness of an online training platform. *Computer Assisted Language Learning, 29*(1), 40–60. https://doi.org/10.1080/09588221.2014.889714

Lankshear, C., & Knobel, M. (2008). Digital literacies: Concepts, policies and practices. In C. Lankshear & M. Knobel (Eds.), *Digital literacies: Concepts, policies and practices* (p. 321). New York, NY: Peter Lang.

Lea, M. R. (2016). Reclaiming literacies: Competing textual practices in a digital higher education. *Teaching in Higher Education, 18*(1), 106–118. https://doi.org/10.1080/13562517.2012.756465

Lee, L. (2011). Blogging: Promoting learner autonomy and intercultural competence through study abroad. *Language Learning & Technology, 15*(3), 87–109. http://llt.msu.edu/issues/october2011/lee.pdf

Lee, C., Yeung, A. S., & Ip, T. (2016). Use of computer technology for English language learning: Do learning styles, gender, and age matter? *Computer Assisted Language Learning, 29*(5), 1035–1051. https://doi.org/10.1080/09588221.2016.1140655

Lenhart, A. (2015). *Teens, social media & technology overview 2015*. Washington, DC: Pew ResearchCenter.http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/

Levy, M. (2009). Technologies in use for second language learning. *Modern Language Journal, 93*(s1), 769–782. https://doi.org/10.1111/j.1540-4781.2009.00972.x

Levy, M., & Steel, C. (2015). Language learner perspectives on the functionality and use of electronic language dictionaries. *ReCALL, 27*(2), 177–196. https://doi.org/10.1017/s095834401500004x

Lin, H. (2015). A meta-synthesis of empirical research on the effectiveness of computer-mediated communication (CMC) in SLA. *Language Learning & Technology, 19*(2), 85–117.

Lin, C.-H., Warschauer, M., & Blake, R. (2016). Language learning through social networks: Perceptions and reality. *Language Learning & Technology, 20*(1), 124–147. http://llt.msu.edu/issues/february2016/linwarschauerblake.pdf

Liu, G. Z., Lu, H. C., & Lai, C. T. (2014). Towards the construction of a field: The developments and implications of mobile assisted language learning (MALL). *Literary and Linguistic Computing, 31*(1), 164–180. https://doi.org/10.1093/llc/fqu070

Meyers, E. M., Erickson, I., & Small, R. V. (2013). Digital literacy and informal learning environments: An introduction. *Learning, Media and Technology, 38*(4), 355–367. https://doi.org/10.1080/17439884.2013.783597

Pearson Student Mobile Device Survey 2015. (2015). Retrieved from http://www.pearsoned.com/wp-content/uploads/2015-Pearson-Student-Mobile-Device-Survey-College.pdf

Peterson, M. (2012). Learner interaction in a massively multiplayer online role playing game (MMORPG): A sociocultural discourse analysis. *ReCALL, 24*(2012), 361–380. https://doi.org/10.1017/S095834401200019

Peterson, M. (2016). The use of massively multiplayer online role-playing games in CALL: An analysis of research. *Computer Assisted Language Learning, 29*(7), 1181–1194. https://doi.org/10.1017/s0958344012000195

Reinhardt, J., & Zander, V. (2011). Social networking in an intensive English program classroom: A language socialization perspective. *CALICO Journal, 28*(2), 326–344. https://doi.org/10.11139/cj.28.2.326-344

Ryu, D. (2013). Play to learn, learn to play: Language learning through gaming culture. *ReCALL, 25*(2), 286–301. https://doi.org/10.1017/s0958344013000050

Steel, C. H., & Levy, M. (2013). Language students and their technologies: Charting the evolution 2006–2011. *ReCALL, 25*(3), 306–320. https://doi.org/10.1017/s0958344013000128

Sung, K. M., & Tsai, H. M. (2014). Motivation and learner variables: Group differences in college foreign language learners' motivations. *International Journal of Research Studies in Language Learning, 3*(2), 43–54. https://doi.org/10.5861/ijrsll.2013.561

Sylvén, L. K., & Sundqvist, P. (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL, 24*(3), 302–321. https://doi.org/10.1017/s095834401200016x

Tanaka, K., & Ellis, R. (2003). Study abroad, language proficiency, and learner beliefs about language learning. *JALT journal, 25*(1), 63–85. http://jalt-publications.org/recentpdf/jj/2003a_JJ.pdf#page=65

Thorne, S. L., & Reinhardt, J. (2008). "Bridging Activities," New media literacies, and advanced foreign language proficiency. *CALICO Journal, 25*(3), 558–572. http://www.u.arizona.edu/~jonrein/pubs/thorne_reinhardt2008.pdf

Thorne, S. L., Black, R. W., & Sykes, J. M. (2009). Second language use, socialization, and learning in internet interest communities and online gaming. *Modern Language Journal, 93*(1), 802–821. https://doi.org/10.1111/j.1540-4781.2009.00974.x

Trinder, R. (2016). Blending technology and face-to-face: Advanced students' choices. *ReCALL, 28*(1), 83–102. https://doi.org/10.1017/s0958344015000166

White, K. D. (2016). Cultures and communities in the virtual world: Beginning the exploration. *IALLT Journal of Language Learning Technologies, 43*(2). Retrieved from: http://www.iallt-journal.org/index.php/ialltjournal/article/download/84/75

Williams, L., Abraham, L., & Bostelmann, E. D. (2014). A survey-driven study of the use of digital tools for language learning and teaching. In J. P. Guikema & L. Williams (Eds.), *Digital literacies in foreign and second language education* (pp. 29–67). San Marcos, TX: Computer Assisted Language Instruction Consortium (CALICO).

Winke, P., & Goertler, S. (2008). Students' computer access and literacy for CALL. *CALICO Journal, 25*(3), 482–509. https://doi.org/10.1558/cj.v25i3.482-509

Jeffrey Maloney is an Assistant Professor of English at Northeastern State University. He received his PhD from the Second Language Studies program at Michigan State University. One of his main research interests is student technology practices in the L2 outside of formal language learning contexts. His other research is focused on language teacher education with technology, and bilingual and heritage learner identity.

Linking Proficiency Test Scores to Classroom Instruction

Charlene Polio

Abstract This study relates scores on listening, reading, and speaking proficiency tests in a Chinese program to what happens in the classroom. Data were collected as part of the Flagship Proficiency Grant to document progress on the ACTFL scale. Qualitative data in the form of classroom observations and focus group interviews with a subset of students and teachers were also collected. Information from the observations was put into activity charts that documented the focus of lesson segments as well as the type of interaction and the amount of Chinese spoken. Comments from Chinese students and teachers were coded to determine what themes emerged and whether or not there was consistency in views of the program among the students and teachers. These themes were then related to test scores and progress. In some cases, such as speaking scores, we can see that the type of speaking activities may not have pushed to students to the higher levels. Throughout the chapter, I discuss what data were collected and contrast them with data that would have been helpful to collect so that an ideal mixed methods study could have been conducted.

Keywords Program evaluation · Chinese language teaching · Mixed methods · Classroom research · Proficiency testing

1 Introduction

An important issue when thinking about proficiency testing results in language programs is to understand what might be contributing to those results. There are many potential contributing factors, including but not limited to any of the following: time on task (how many hours are spent in the classroom or in official

C. Polio (✉)
Second Language Studies, Michigan State University, East Lansing, MI, USA
e-mail: polio@msu.edu

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_15

classroom activities, such as online requirements), student demographics, or in the case of the present chapter, specific classroom activities and/or behaviors. I relate this study to the language program evaluation literature because ultimately, one of the major goals of proficiency testing is to improve instruction. I also discuss the challenges of interpreting the proficiency scores within a singular, large-scale study at one university.

1.1 Design Issues in Evaluation

This chapter is a formative evaluation of sorts in that the ultimate goal of the proficiency tests being considered in this chapter is to improve instruction. More specifically, I discuss an attempt to explain student outcomes in relation to classroom practices and curriculum. Language program evaluation saw a surge of research in the 1980s and 1990s paralleling the overall increase in research in the field of second language acquisition and teaching. Norris and Watanabe (2013) discussed this era through the lens of a series of assumptions that evolved from early discussions of language program evaluation including, among others, the fact that large-scale experimental research as a way to evaluate a program was virtually impossible. Indeed, this formative evaluation is not an experimental study in that no independent variables were manipulated.

In this chapter, in which I aim to provide insights on methods for program evaluation, I include information on language proficiency and language proficiency development from classroom observations and from focus groups with students and teachers. I attempt to link that information to the students' test scores and follow Norris and Watanabe's recommendation that in program evaluation, information from stakeholders, such as students and teachers, is essential. In this chapter I also present methodological considerations because I had considerable challenges in drawing conclusions. As such, I contrast the data that were collected with data that would have been ideal to collect. In this way, this study served as a pilot study on methods in using classroom observations, focus group data, and test scores as part of a package aimed at program evaluation. I hope that the chapter provides guidance for others conducting evaluations using a more ideal mixed methods design.

In an explanation of student-learning outcomes assessment, Norris (2016) noted:

> A key factor that distinguishes outcomes assessment of sort from accountability testing is that academic programs and institutions are largely left to their own devices to determine what outcomes need to be assessed as well as the preferred methods for doing so. Another key distinguishing factor is that assessment findings are supposed to be used as evidence in ongoing cycles of program monitoring for improvement. (p. 174)

In sum, the goal of this chapter is to understand how what happens in the classroom may be related to proficiency scores, with the ultimate aim of program improvement. The specific focus is the Chinese language program at one university involved

in the proficiency grant, described elsewhere in this volume. Specifically, national proficiency tests were used so that the test scores could be interpreted in the context of foreign language teaching in the United States. Of course, test scores alone cannot reveal the inner workings of a program without in some way linking them to instruction, so as part of understanding the test scores and suggesting program changes, qualitative data were collected as part of the larger initiative. It is those qualitative data that form the basis for this chapter.

Researchers in the areas of language program evaluation have discussed design challenges at length. Long (1984) noted that there are limits in evaluating only the outcomes of a program because *product evaluation* does not shed light on the reasons for the outcomes. He called for adding what he called *process evaluations* to any program evaluation. For him, this meant "*systematic observation of classroom behavior with reference to the theory of (second) language development*" (p. 415, italics in original). Lynch's (1996) book on language program evaluation includes a chapter on both quantitative and qualitative approaches, as well as a chapter on mixing approaches in which he included several examples of designs where outcomes were combined with observations and interviews with stakeholders to evaluate a program. He did not use the term *mixed methods,* but rather referred to mixed designs and mixed strategies and concluded that multiple strategies should be used whenever possible to better understand program outcomes.

1.2 The Potential of Mixed Methods Models

Mixed methods research has gained much interest recently in applied linguistics (e.g. Hashemi & Babaii, 2013; Polio & Friedman, 2017; Riazi & Candlin, 2014). One useful definition within applied linguistics came from Dőrnyei (2007):

> A mixed methods study involves the collection and analysis of both quantitative and qualitative data in a single study with some attempts to integrate the two approaches at one or more stages of the research process. In other words, mixed methods research involves the mixing of quantitative and qualitative research methods or paradigm characteristics. (p. 161)

The primary benefit of mixed methods research is that it can draw on the strengths of both quantitative and qualitative methods. If done correctly, the data should be integrated and at least one research question should draw on both quantitative and qualitative data (Brown, 2014; Cresswell & Plano Clark, 2011). Within applied linguistics, specifically L2 writing, Polio and Friedman (2017) found that several studies claiming to employ mixed methods did not truly integrate quantitative and qualitative data. I hope that this study can serve as an example of a more integrated research approach by using a concurrent-explanatory design. This type of study is detailed in the methods section, but in sum, classroom observations and interview data are combined to help explain outcomes, or test scores.

Because of various limitations discussed later, I was not able to collect all of the data that I would have liked; however, this is not unusual when using data from authentic classes. As Norris (2016) noted:

> Language teaching and learning as well as other educational and social endeavors related to language learners and users, play out not within sanitized laboratories where theories are carefully tested, but rather under the realities of geopolitical and economic forces; governmental budgets and policies; institutional affordances and constraints; and the everyday actions of administrators, teachers, learners, and others. (p. 169)

Therefore, in the design section of this chapter, I will propose what I consider to be a more ideal design for understanding language programs in order to do a more robust formative evaluation. I then explain which data were collected and how they were analyzed.

1.3 Characterizing Classrooms Based on Observation Data

A variety of recent studies, often with no control groups, have examined student progress in specific instructional contexts. For example, Yasuda (2011) implemented a new genre-based writing instruction study in a Japanese EFL program and examined how students' writing changed over the course of one semester. She described the curriculum in detail and used student interview data to help triangulate her quantitative results. Mazgutova and Kormos (2015) examined student progress in an EAP writing program in terms of a number of measures of syntactic and lexical complexity. Hung (2015) conducted a quasi-experimental study of flipped versus traditional classes and provided a detailed description of the curriculum as well as student interview data. None of these studies, however, collected any classroom data. Of course, a researcher can collect only a limited amount of data, but data about what actually happens in classrooms is somewhat scarce. In fact, only one of the other chapters in this volume included observational classroom data (Dillard, this volume) in an attempt to evaluate a professional development intervention, but no test data were included.

Fortunately, other researchers have paved the way for a true mixed-methods study with the purposes of program evaluation and improvement. Duran, Roseth, and Hoffman (2015) supplemented a study of bilingual preschool instruction with observation data. In a 2-year experimental study, they randomly assigned children in a Head Start program to a predominant English (PE) class or a transitional bilingual education (TBE) class. The difference between the two programs was the language of instruction. In the experimental bilingual condition, all instruction and materials were in Spanish. In the English condition, no Spanish was used.

In Year 2, all instruction remained in Spanish during the first half of the school year in the TBE classroom and all instruction remained in English in the PE classroom. During this second year, more L2 English teaching strategies were

incorporated, such as using visuals and manipulatives to support instruction, using total physical response, and being intentional in teaching vocabulary. In January, however, English was gradually introduced in the TBE classroom until a ratio of 30% English to 70% Spanish was achieved during mid-February. Spanish was introduced in the PE classroom until the ratio of 30% Spanish to 70% English was reached (p. 926).

As a check to see if the independent variable, language of instruction, was being implemented as planned, the authors included three methods of characterizing classroom instruction. Specifically, Duran et al. (2015) first randomly selected 10-minute segments from six classes in each program and coded them for instance of English in the TBE class and Spanish in the PE class. Second, they used the Early Language and Literacy Classroom Observation (ELLCO; Smith, Brady, & Anastasopoulos, 2008) that focused on "organization of the classroom, curriculum integration, oral language facilitation, the presence and use of books, and embedded print and early writing" (p. 928). Third, the observational Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) was used to assess classroom quality. This study was impressive because of the range of observation tools, especially in contrast to studies that use few to none. The authors' use of the multiple tools provided guidance to researchers on how they can use a similar variety of tools to get at a more thorough representation of what happens in the classroom in evaluation studies.

2 Current Study

2.1 Goals

As mentioned earlier, the goals of this chapter are couched within the context of the Chinese program of one of the universities that participated in the proficiency testing program. The data were collected from four languages by several different researchers (hence the use of *we* when talking about data collection). The overarching research question is:

1. Is it possible to link test scores to qualitative data in a way that will inform future instruction within a program?

   In relation to research design, I also explore:

2. Given the successes and shortcomings of this study, how can future mixed methods evaluation studies be better conducted?
3. How successful is the use of a general observation form at capturing the nature of classroom instruction?

2.2 Context

This study focuses on 1 year in the Chinese program. The program offers a fall and spring sequence of language-focused courses (i.e., 101–102; 201–202; 301–302; 401–402) for 4 years as well as additional 300-level (350; 366) and one other 400-level class (466) that were included in this study. Chinese 350 is a Chinese linguistics class that is taught mostly in Chinese with readings in English, and Chinese 366 and 466 are culture, film, and literature classes that are taught in both English and Chinese. Students in these courses may have demonstrated proficiency somewhat higher or lower than students in the 301–302 and 401–402 courses because these courses do not have to be taken concurrently with 300–400 courses. Because the 350, 366, and 446 courses are conducted partially in English and are not general language proficiency courses that every student in the sequence takes, I do not focus on them here. The number of students in all of the Chinese courses who participated in the oral proficiency testing was 107 in the fall and 124 in the spring.

2.3 Study Design: Actual and Ideal

In describing the methods and procedures of this study, I will include not only a discussion of what was done, but also a discussion of what ideally could have been done. These design features are illustrated in Fig. 1. Figure 1 includes the type of data collected including test scores, interviews, and observations as well as data that could have been collected (but was not) to better inform the test results. These actual and ideal data sources are shown on a timeline representing the academic year during which the data were collected. The unshaded boxes indicate data that were actually collected and the shaded boxes indicate ideal data that would have been helpful to have. The bolded boxes indicate quantitative data. I elaborate on each of these data types here. In the results section, I report what we were able to find and why additional data would have been helpful. The type of mixed methods study used here is concurrent-explanatory (QUAN + QUAL) design (see Creswell & Plano Clark, 2011). *Concurrent* means that both quantitative data and qualitative data were collected during the same phase of the study, all during the first year. In other words, we did not look at the test data and then decide which qualitative data to collect. *Explanatory* means that the qualitative data are used to try to explain the trends in the quantitative data. And finally, *QUAN + QUAL* refers to the fact that neither data were privileged in the analysis.

Timing and Procedure This study was conducted during one academic year. Test data were collected near the end of each of the two semesters. The qualitative data (interviews and classroom observations) were collected at various times based on the availability of the researcher who was conducting the interviews and the schedules of the observed teachers and those doing the video recording.

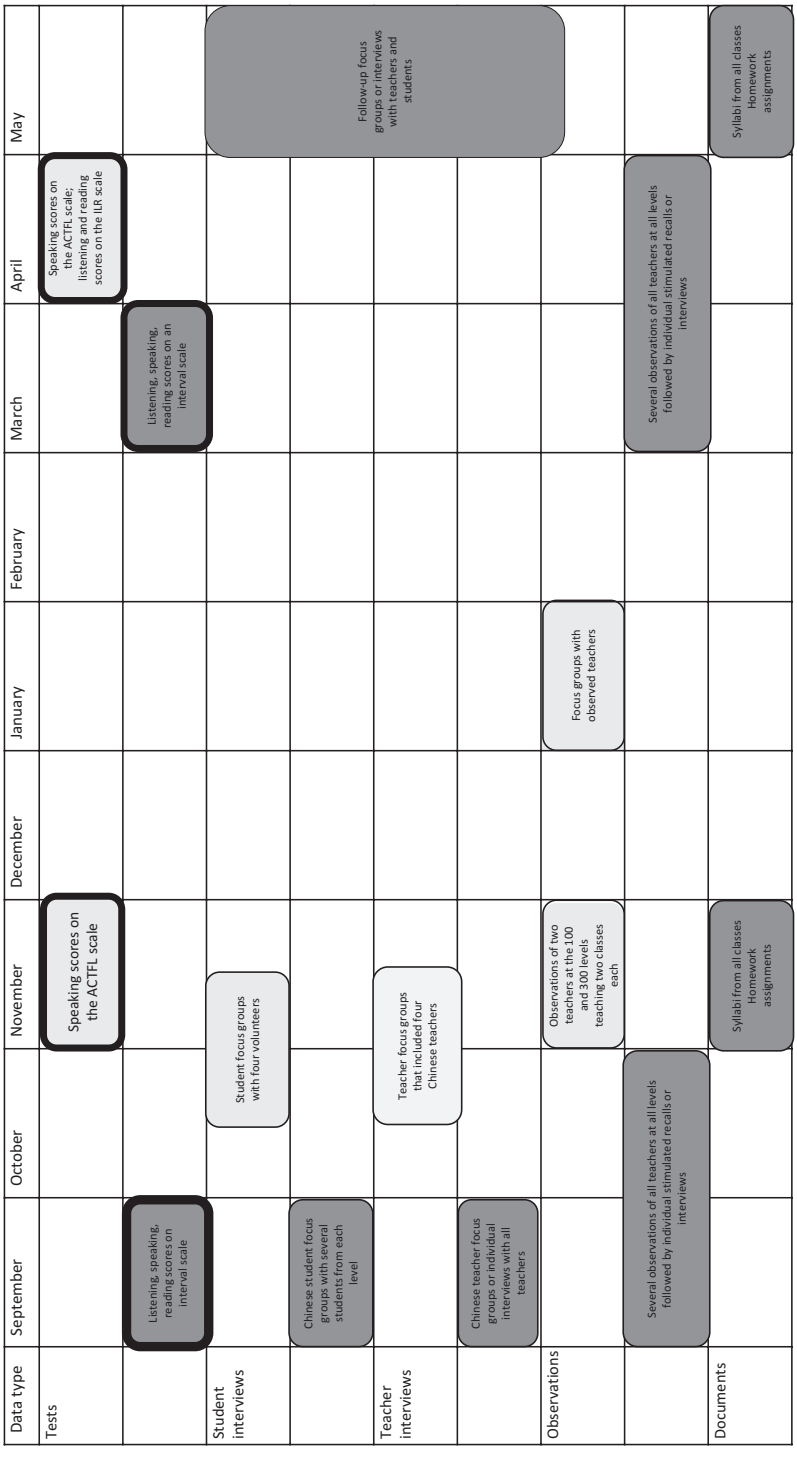


Fig. 1 Actual and ideal design

In terms of timing, it would have been best to collect the quantitative data very early in the academic year as well as at the end of the academic year. Had we had access to the initial test results, we could have purposively sampled students to participate in the focus group interviews. In other words, we could have chosen a stratified sample of students at different levels based on their proficiency scores to understand different types of student views on the curriculum and instruction. We also could have found students who made different amounts of progress and interviewed these students at the end of the year. However, the results were not used in this way. Instead, we asked for volunteers to participate in the qualitative aspects of this study, as the original data collection methods were employed in a more exploratory fashion in the first year of this proficiency grant project.

Test Data Test data were collected from Chinese students in the fall and spring. The Oral Proficiency Interview, computerized (OPIc) from Language Testing International (https://www.languagetesting.com/) was used to assess students' speaking in the all but first year courses in the fall and all levels in the spring. Tests developed by the American Councils for International Education (https://www.americancouncils.org) were used in the spring to assess listening and reading for the 200–400 levels. As such, the speaking results are reported using the ACTFL scale and the listening and reading results using the ILR scale. Some students took the test both semesters, but some did not. The test data here are considered cross-sectional given that students at different levels took the tests, but there are also some students who took the speaking test in both the fall and spring. We did not examine these students individually.

Ideally, we would have been able to collect test data from all students, with the exception of complete beginners early in the fall, on all three skills and track them over the year. In addition, as discussed later, raw test scores may have provided a better picture of progress. The ACTFL scale is not an interval scale, so students are not expected to progress across levels at the same rate. An additional problem is that because the ACTFL listening and reading tests were not yet available in Chinese, the American Councils tests of listening and reading were used. This resulted in a conversion from the ILR scale to the ACTFL scale, which may have resulted in a loss of information. Furthermore, it made it more difficult to relate the listening and reading scores to the speaking scores, which are presented according to the ACTFL scale.

Classroom Observation Data and Follow-Up Interviews The goal of the observations was to collect information about how the Chinese curriculum was being implemented in the classrooms. We chose to develop an observation scheme that could be filled out without having to transcribe each of the class sessions. To construct the observation form, we focused on issues that we thought would affect listening, speaking, and reading skills. We also wanted to try to get a picture of how the classes were regularly taught throughout the semester. The goal was not to quan-

tify interactional features such as question types or recasts (e.g., Zyzik & Polio, 2008) but to focus on classroom features that might affect the proficiency test outcomes. The characteristics that we were most concerned with were the amount of exposure to the target language during class, the amount of student production, and the focus of instruction. Thus, we created the form shown in Appendix A. Two native speakers of Chinese coded the classroom videos, and the form shown includes an example of how one class was coded by the two different coders. We asked the coders to add new rows when they believed that there was a change in activity, but we did not define what should be considered an *activity* or *task*.

We video recorded two instructors, one at the 100-level and one at the 300-level. For each instructor, two classes were recorded about two thirds of the way through the first semester. We hoped that by recording two sessions of each instructor, we would obtain a clear picture of a typical class. We then conducted a focus group interview with the two observed teachers together along with those from other languages who had also been observed. The questions for the focus group are provided in Appendix B; however, for this focus group and the others described in the next section, not all questions were asked because of time limitations. Instead, the focus group questions represent what we ideally would have liked to find out. At times during the focus groups, there were no responses at all from the Chinese teachers. The purpose of the observed-teacher focus groups was to allow the teachers to contextualize the observed classes in the curriculum, comment on their goals for the class, and note what they thought went well and what did not. They were conducted in part to help us understand what was happening, but also to triangulate the data from the observations. For example, a teacher may have ended up talking more than usual, but believed that he or she spoke too much that day and normally tried to give the students more talking time. Both the observations and follow-up focus groups interviews are represented in the study design shown in Fig. 1.

Generally, more qualitative data is desirable and so it would have been useful to have observed more classes throughout the year from more teachers in order to better understand the implementation of the curriculum. But more importantly, stimulated recall sessions with the teachers observing and commenting on videos of their teaching would have provided more insight into their reasoning of how they taught. Our focus group interviews were conducted up to 1 month after the observed classes, so it was unlikely that they could recall specifics of the class. The focus groups did not give them the time to comment on individual interactions.

Teacher and Student Focus Group Interviews As Fig. 1 shows, focus groups were conducted with a large number of teachers in the four language programs including four Chinese instructors during the first academic year. Two of these were the same teachers that were observed, but we did not attempt to link the comments from the general focus group interview to the specific class observation. These questions are provided in Appendix C. Focus group interviews were also conducted with four student volunteers who received payment for their time. These focus

groups included students from other languages as well, and the questions are provided in Appendix D. For both the teacher and student focus groups, responses from the Chinese teachers and students were extracted for analysis.

Focus groups were chosen because they were easier to schedule than individual interviews, in part because they were conducted by someone outside of our university. We felt that the students and teachers might be more willing to talk honestly with an outside investigator. They did know, however, that we would see the transcripts from the focus groups. Although they may have some advantages, we may have obtained more information from interviews because each teacher may have spoken at more length. As such, there are limited responses from each of the Chinese class participants.

We also relied on volunteers and were not able to sample students from a range of levels. As mentioned earlier, had we selected students based on test scores, we could have purposively sampled students at the beginning of the semester and then interviewed students who had made differential progress. In doing so, we could have better integrated the qualitative and quantitative data.

Document Data We did not collect any syllabi, assignment sheets, or materials. In retrospect, an analysis of the materials would have been especially helpful for better understanding progress or, in the case of listening and reading, the lack thereof. The classroom observations provided information about what was happing in class, but an analysis of the online materials and textbooks would have provided more information about how the oral and written input may have influenced test scores. In addition, we could have seen the types of grammar activities and homework that the students had to do. We also could have better triangulated our observations from the classroom.

Table 1 Results of proficiency testing

	Fall 14				Spring 15			
	101	201	300	400	102	202	300	400
Speaking								
N		58	37	12	34	53	17	20
Median		NH	IL	IL	NM	IL	IL	IM
Mode		NH	IM	IM	NM	NH	IL	IM
Listening								
N						46	19	18
Median						1	1	1
Mode						1	1	1
Reading								
N						42	18	19
Median						1	1	1
Mode						1	1	1

NM novice mid, *NH* novice high, *IL* intermediate low, *IM* intermediate mid

3 Results

3.1 Test Data

Because of the small numbers of students tested, I report the median and modes of the test scores in Table 1. As noted earlier, speaking scores are reported on the ACTFL scale and the listening and reading scores on the ILR scale. I chose to present the results according to the actual score assignment from the ACTL and ILR scales instead of converting them to numerical scores as done in many studies in this volume because in this sample of Chinese learners there are very few proficiency scale levels represented.

The first step was to look at the test results to see what was salient about the students' progress or lack thereof. What is most obvious is that the students made progress on the speaking test but not on the listening or reading tests. For example, in the fall, the students' performance in the 200-level course was at the Novice-high level, while the students in both the 300 and 400 level courses performed at the Intermediate-low or Intermediate-mid based on the medians and modes, respectively. In the spring, we can see a more steady progression across the levels. These results are not unexpected. The data for listening and reading, however, are quite different. Table 1 shows no progress at all; students have an ILR score of 1 for 3 years.

It is important to consider that a 1 on the ILR scale aligns with an IL/IM on the ACTFL scale, so it is quite possible that the students are improving but the score levels are not sufficiently nuanced to register any change. Thus, raw test scores might have been helpful. Another problem is that we did not have access to the listening and reading passages or even the topics of the passages. Although scores on a general proficiency test should not be affected by topic, it would have been helpful to see what topics and related vocabulary were being tested and to then consider whether these aligned with the curriculum, particularly considering the reading test where a lack of character knowledge can greatly impede comprehension. However, it was surprising, at first, to see the lack of progress on the listening scores as well. But in hindsight, we cannot infer that there was no progress; progress was likely unmeasureable by the American Council tests. It could be that the tests were less discriminating with learners at the lower (Novice or ILR level 1) levels of proficiency, but that is an empirical question not investigated directly in this study. A related question is whether proficiency can reliably be measured when students are at an ability level of Novice or ILR level 1, because proficiency assessment is supposed to be a measurement independent from one's coursework. But at such low levels of proficiency, it would be hard to argue that the learners' language knowledge is anything but dependent on the coursework: In the beginning, learners only know what they have been taught, so testing them on topics with which they are not familiar may be impractical and non-informative. In other words, testing the proficiency of low-level learners may be impossible if the learners themselves are not proficient in any way in the

language. Thus, the test data showing no growth or progress for low-level learners may need to be taken with a grain of salt: they must have been learning (arguably), but the learning being done was probably just not registered by the test.

3.2 Observation Data and Observed Teacher Interviews

Summary of Findings After each of the two coders coded one session, I compared the two charts from one session and found that one coder had divided the class into a greater number of activities. I met with the coders, and we discussed what should be considered an *activity*. This was the only type of coder norming that was done. Of the four classes for which charts were filled out, three had between 18–23 activities with a discrepancy of exactly two activities per session. For example, in one case one coder said that there were 21 activities and the other said that there were 23. One of the sessions, however, had one of the coders breaking up an activity into many sub-activities resulting in a difference of 27 versus 19 activities. Again, examples of the charts are shown in Appendix A.

To analyze the results, I compared the charts from the two coders for each class session. I first looked at the amount of target language used by the teacher because a large of amount of English could have impeded students' progress in listening proficiency. For the first semester Chinese class, both coders characterized the teacher's speech in the classes as being almost completely in Chinese. They noted only single word translations a few times during the two sessions. The PowerPoint (PPT) slides, however, regularly included instructions or translations into English. The 301 sessions were also almost completely in Chinese as well but included only a few slides with English instructions.

The two 101 sessions were somewhat different in terms of focus and modality. One of the classes was focused on grammar points while the other was more focused on tasks or functions such as making an appointment or writing a note. Grammar was taught through a variety of modalities including listening and response, constructed response activities, and guided dialogues where students had to create a new dialogue based on a sample. Some of these activities were led by the teacher, and some included pair work. The second class was almost completely teacher-fronted and involved a lot of choral repetition as the teacher and students read through examples of written notes and phone call dialogues.

The 300-level language class sessions were more vocabulary focused and teacher centered with students having no more than five minutes of pair work. When the students did work in groups, they worked from their textbooks to answer comprehension questions, which the teacher then went over. In addition, one of the sessions included a three-minute excerpt from a television show. The teacher played it once, went over some pre-listening questions, and then played it a second time.

The follow-up focus group interviews with the two observed teachers did not reveal any inconsistencies in terms of what was observed in the classes. To the con-

trary, the teachers confirmed that what happened in the classes was quite typical in terms of amount of Chinese spoken, the types of materials used, and the types of activities. They did both express concern that the students did not use Chinese consistently during the class activities, but this was not captured in the observations.

Relationship to Quantitative Results The lack of progress in listening cannot be explained by a lack of exposure to Chinese in the classroom. It may be, however, that at the lower levels, there is a lack of sustained listening activities that are similar to the type of listening that students do on the proficiency test. When the teacher is speaking Chinese, students can rely on interactional or visual cues to follow along in the classroom, cues that are absent in the audio-only testing context. On the other hand, it also seems that at the 300-level, (and this was confirmed by the observed teacher), authentic videos from YouTube are used on a regular basis in class, listening practice that includes visual information for scaffolding and interpretive help. Regarding the lack of progress in reading, at the 100 level, there was much use of English on the PPT slides, and the reading was limited to the textbooks. At the 300 level, there was some emphasis on reading comprehension, but the teacher noted that the readings came exclusively from the students' textbooks.

Regarding speaking scores, the students did interact in Chinese in the classroom, both in teacher-fronted activities and in pair work, but in all observed classes, teacher-talk dominated. Nevertheless, the students showed consistent progress in their speaking scores. This progress was made despite the teachers' concerns that their students were speaking too much English during their group work.

3.3 Teacher and Student Focus Groups

Summary of Findings and Relationship to Observations To analyze the focus group responses from the four Chinese-language students, a research assistant coded the responses and extracted quotations related to four topics that we thought could help explain the test scores. These included the amount of target language use, the amount of classroom interaction and speaking time, and the skill focus or content. The last topic came up when students were asked what they thought the program goals were. The teachers' responses focused more on both their own and their students' goals, which are somewhat related, and assessment, which is beyond the scope of this chapter.

When asked about the amount of target language use, one of the students said that although the upper-level classes were taught in Chinese, the lower-level classes were taught about 60% in Chinese. Meanwhile, one of the teachers said that she spoke only in Chinese, but had online exercises that included a lot of English explanations. Interestingly, two of the students spoke about their own use of English versus Chinese. One said that in small groups, students usually slipped back into speaking English, while another student said that one of the teachers was very strict

about forcing them to speak Chinese, and though they "hated it," it helped them improve a lot. Because only one student commented on the lower percentage of Chinese in the lower-level classes, we cannot be sure that this was a regular occurrence.

In terms of interaction, the students who responded were from the 201 and 350 courses, which were not observed. The responses about the skills and content of the Chinese program were not specific enough to be insightful and instead referred to all skills and not anything specific about the 100 or 300-level classes. One exception was that one student stated that reading was emphasized at the higher levels.

Relationship to Quantitative Results The main finding that needs to be explained is why the students demonstrated (via testing) progress in speaking but not discernably in listening or reading. The teachers' responses were heavily oriented toward oral language production even though two mentioned integrating all skills. They mentioned their students' goals as "to communicate with the native speakers" and for them "to survive to talk to a native speaker." It is possible that because of the teachers' orientation toward teaching speaking, they did not place much emphasis on reading or listening. One student seemed to confirm that reading was the focus of only the higher levels. The reasons why are difficult to ascertain in part because we, as researchers, are not sure if the American Councils tests of listening and reading gave us useful information about the Chinese language learners' growth in listening and reading. More nuanced tests may be needed to get at such change, or such early-level development may need to be documented in ways other than generalized, language proficiency tests. This research reinforces the need to more fully understand when language is proficient enough to undergo proficiency assessment, and whether before that threshold is obtained, if classroom-based (and instructionally-content-dependent) assessments are better measures of student success in learning.

4 Conclusions

The questions put forth at the beginning are addressed below.

1. Is it possible to link test scores to qualitative data in a way that will inform future instruction within a program?

The results show that the classroom traits that we chose to focus on in the observations, as well as the questions that we asked in the interviews, had the potential to explain the tests scores. In other words, the qualitative data addressed central issues such as the amount and type of exposure to the target language, classroom interaction, and types of activities. However, we did not obtain information from the students about their goals and efforts in the classroom, and this is an essential piece of information.

A more serious problem is the lack of raw test scores, or perhaps reliable or meaningful test scores, on the listening and reading tests as well as examples of the

kinds of listening and reading that was being tested. In this study, the black box is the content of the listening and reading tests and whether that content aligned with the current curricula and in-class instructional content. On one hand, general proficiency tests allow a program's scores to be interpreted in a wider context, but they may not be the best tests for programs wishing to effect curricula changes. It would be ludicrous to think that students made no progress in listening or reading, but as presented in this chapter, the results could be seen as disheartening and difficult to link to the qualitative data. Thus, I strongly caution researchers using general (and broad ranging) proficiency tests to measure very low-level learners to think carefully about the application of the test data to theories about development and growth, and to also think carefully about whether the data from low-level learners are accurate for interpreting evidence of learning.

2. Given the successes and shortcomings of this study, how can future mixed methods evaluation studies be better conducted?

Figure 1 provided a graphic representation of the ideal data collection. One of the major shortcomings was the lack of data in terms of quantity and quality that were obtained from the student and teacher focus groups. Because of time limitations, not all of the questions could be addressed nor could participants be probed for further details. Although interviews might have provided more information, organizing focus groups by language might have been more informative. Students would be able to focus on the different courses and provided more details about what they were learning at each level. Similarly, had all the Chinese teachers talked together, they might have come to clear conclusions about the strengths and weaknesses of the program. The focus group for the observed teachers was more successful in that we confirmed what we observed in the classes. Stimulated recalls might have shed light on specific interactional features in the classroom, but these were traits not addressed in this study.

A better way to integrate the quantitative and qualitative data would have been to choose student participants based on their class level and test scores. This would have allowed us to better understand their likely differential goals and reasons for studying Chinese. Another important missing piece here is the teachers' views on the quantitative data. End-of-year interviews could have included showing the teachers' the quantitative test results and asking for their interpretations of them.

3. How successful is the use of a general observation form at capturing the nature of classroom instruction?

It was possible to come up with a chart that gives a general idea of what is happening in the classroom in relation to certain classroom variables. Because we were not quantifying the data, we did not do extensive norming. Nevertheless, it was very helpful to have two coders both to confirm what was happening in the classroom from two perspectives but also in case one coder chose not to record something. For example, one coder did not note the extensive use of English on the PPT slides in the 100-level class.

Another approach would have been to use the charts for individual follow-up interviews with the teachers. Given that a stimulated recall with the videos would have been time consuming and perhaps unnecessary, the charts could have served as a reminder about what was happening in the classroom.

In addition to answering the above questions, one of the goals of this chapter was to provide suggestions for identifying the need for, and then implementing, program changes. Despite problems with the interpretation of the listening and reading test scores, it seems appropriate to suggest that there should be more focus on listening and reading with activities that go beyond listening to the teacher and reading the textbook. A positive outcome from the listening and reading test results was an exploration of the types of listening and reading activities that exist in the current classrooms, and whether those activities are sufficient, and how more (and more types of them) could be integrated into the general language program.

Sustained reading of authentic materials in low-level Chinese classes can be challenging because of the students' lack of character recognition that cannot be overcome through cognate clues. Nevertheless, some supplemental readings could be introduced and students could then return to those readings throughout the semester to try to better understand them. There was a focus on vocabulary in the 300-level class, but if students are to read better, there may need to be a greater emphasis on character recognition. Learning Chinese characters is, of course, time-consuming, so programs need to consider how important reading is within the curriculum.

With regard to listening, practice interacting with authentic listening texts of progressively increasing lengths might address the kind of listening that is assessed on general proficiency tests. The data showed that there was some use of authentic listening materials in the 300-level classes, but authentic listening materials could be introduced, in small chunks, earlier in the curriculum. Although they might be difficult for beginners, teachers could return to them throughout the semester and review with students what they heard and understood. Alternatively, non-authentic materials that go beyond dialogues could be used as well. Although listening to the teacher speak Chinese is important, an overreliance on teacher-talk and visual cues might be ineffective for progress beyond the beginning levels. In addition, the data presented the program with more, but beneficial, questions. A healthy discussion ensued on whether audio-only listening (as measured on the listening test) is the true goal for Chinese language listening, which helped all teachers understand the types of proficiency-test data that teachers, researchers, and programs can currently obtain from companies like Language Testing International and American Councils. An agreed upon conclusion is that the Chinese language program needs different types of listening assignments, practice, and assessment to better represent the larger construct of listening that the program would like to focus on.

Appendices

Appendix A: Classroom Observation Charts

Coder 1

Time	Summary of activity	Participant organization	Content or focus of activity	Student modality focus	Materials and visuals used	T amount of English use/ Reasons for using English
00:00–00:07	Greetings	T to the whole class	How are you today?	Listening & response		0%
00:07–00:46	Introduction of lesson	T calls on Ss to respond	Review Lesson 6: Make an appointment	Listening & response	PPT	0%
00:46–1:45	Grammar point review by asking questions (fill in the blanks)	T to one S at a time/ Choral repeat	A(在……)给B打电话 A calls B (at…)	Listening & response	PPT	"location" 1%
1:45–5:41	Grammar point (要) review by asking questions (Fill in the blanks & form questions) based on the cues)	T to one S at a time/ Choral repeat	要 (indicates a future action or commitment)	Listening, response & repeat, a little reading	PPT	"indicates a future action or commitment" "Wh-question" English instructions on the PPT 5%
5:41–11:42	Practice talking about one's schedule using 要 (information gap)	T guides/to one S at a time	Schedule	Reading, listening & speaking	PPT	0%

Coder 2

Time	Summary of activity	Participant organization	Content or focus of activity	Student modality focus	Materials and visuals used	Amount of English used by T/Reasons for using English
0:00–0:44	Greetings with Ss. Introduced today's lesson.	T to all Ss. Some Ss responded.	"Class starts!"	Listening	PPT slides	0%
		Other Ss sat and listened.	"我们上课啦!"			
			Let's review Lesson 6."			
			"我们今天复习第六课"			
0:45–1:45	Reviewed a grammar point by making new sentences	T asked questions and called on Ss. Ss listened and responded.	A calls B	Listening and some speaking.	PPT slides	Almost 0%
		Choral repetition.	A 给B打电话			T translated one word ("location") for Ss.
			A calls B at(location)			
			A 在……给B打电话			
1:45–5:43	Reviewed a grammar point by making new sentences/ filling in the blanks	T asked questions and called on Ss. Ss listened and responded.	Going to do something (Chinese future tense marker) 要	Listening and responding to questions	PPT slides	10%
		Choral repetition.				T explained the meaning of the word in English
						PPT contains English translations.
5:44–11:49	Reviewed how to talk about schedule	T asked questions and called on Ss. Several Ss responded.	Available	Listening and Ss who were called on did some speaking	PPT slides and blackboard	5%
			有空			T used "schedule" instead of the Chinese both when teaching and in PPT.
			When someone is available?			
			……什么时候有空?			
			Schedule			

Appendix B: Focus Group Questions for Observed Teachers

1. How do you feel about the classes that were observed?
   (a) Were they typical?
   (b) If not, why not?
2. Objectives
   (a) What are your objectives for your courses?
   (b) What do you want your students to learn by the end of the course?
   (c) What do you want the students to be able to do with the language?
3. Classroom discourse
   (a) How much of the target language and how English do you speak in your different classes?
   (b) What do you do to make the target language comprehensible?
   (c) When and why do you use English?
4. Materials
   (a) What is the focus/approach of the textbook that you are using?
   (b) What is your opinion of it?
   (c) How do you supplement the textbook? What other materials do you use?
   (d) Which, if any, authentic materials do you use? What do you see as obstacles to using more?
   (e) What kind of homework do you give the students? What do you see as the purpose of the homework?
5. Assessment
   (a) How do you assess the students? What challenges do you see in the area of assessment?
   (b) Does the assessment match your curricular objectives? Why or why not?
6. Planned focus on language
   (a) How do incorporate grammar and vocabulary into your lessons? Is this usually the main focus?
   (b) For classes where your focus is on content, how do you focus on language?
7. Incidental focus on language and feedback
   (a) How do you deal with student errors in class?
   (b) How do you deal with student errors in writing?

8. Amount of student talk
   (a) What kinds of speaking activities do your students do in class? Outside of class?
   (b) Do you do group work? If so, what do the students do in the groups?
   (c) What do you think the ratio of teacher to student talk is?
9. General
   (a) What do you see as the strengths of the program and your classes?
   (b) In what areas do you think the curriculum needs to be revised?
   (c) What do you see as the challenges to teaching the courses that you do?

Appendix C: Focus Group Questions for All Teachers

1. How long have you taught Chinese, French, Russian or Spanish at MSU? How long have you taught elsewhere?
2. How do you define the term "proficiency-based instruction"? [Probe: Please tell me more about ___.]
3. What are the goals of your current language program? What are your students expected to be able to do by the time they finish the program?
4. How do you determine whether students have reached these goals?
5. What are your students' goals for studying the language?
6. What are your experiences with testing and assessment in your current program? What kinds of assessment practices does your program currently use?
7. What are your experiences assessing the oral proficiency of your students?
8. Have you ever been trained to assess your students' oral proficiency using a large-scale test like the ACTFL OPI, or something like it? If so, what was your experience with that training?
9. What kind of professional development do you think would benefit you? [Probe: Describe. Why?]
10. Is there anything else you would like to tell us about your experiences teaching language, assessing your students' oral proficiency, or a workshop on how to assess oral proficiency, like the ACTFL OPI workshop?

Appendix D: Focus Group Questions for Students

1. Which language have you studied most recently (Chinese, French, Russian or Spanish) [Probe: Which one(s) have you studied at MSU?]
2. What are your experiences learning this language/these languages? How long have you studied it, and in what situations?

3. What are the goals of your current language program? What are you expected to be able to do by the time you finish the program?
4. What do you think about your progress? Are you progressing at the rate you expected? Please explain.
5. What do you like about the classes you're taking? How could they be improved?
6. What are your experiences with testing and assessment in your current language program?
7. What kinds of assessment practices does your program currently use? Is your speaking assessed? If so, when and how?
8. Do you feel that you understand the way that your language skills are assessed?
9. Have you heard of the ACTFL Proficiency Guidelines? If so, what is your experience with them?
10. Have you ever completed a self-assessment of your language skills? What was it like?
11. What questions do you have about assessments of your language skills?
12. Is there anything else you would like to tell us about your experiences learning and being assessed?

References

Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh, UK: Edinburgh University Press.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*.Oxford: Oxford University Press.

Durán, L., Roseth, C., & Hoffman, P. (2015). Effects of transitional bilingual education on Spanish-speaking preschoolers' literacy and language development: Year 2 results. *Applied Psycholinguisitcs, 36*, 921–951.

Hashemi, M. R., & Babaii, E. (2013). Mixed methods research: Toward new research designs in applied linguistics. *The Modern Language Journal, 97*, 828–852.

Hung, H.-T. (2015). Flipping the classroom for English language learners to foster active learning. *Computer Assisted Language Learning, 28*, 81–96.

Long, M. H. (1984). Process and product in ESL program evaluation. *TESOL Quarterly, 18*, 409–425.

Lynch, B. K. (1996). *Language program evaluation: Theory and practice*. Cambridge, MA: Cambridge University Press.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for academic purposes programme. *Journal of Second Language Writing, 29*, 3–15.

Norris, J. M. (2016). Language program evaluation. *The Modern Language Journal, 100*, 169–189.

Norris, J. M., & Watanabe, Y. (2013). Program evaluation. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Chichester, UK: Wiley-Blackwell.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system™: Manual K-3*. Baltimore, MD: Paul H Brookes Publishing.

Polio, C., & Friedman, D. (2017). *Understand, evaluating, and conducting second language writing research*. New York: Routledge.

Riazi, A. M., & Candlin, C. N. (2014). Mixed-methods research in language teaching and learning: Opportunities, issues and challenges. *Language Teaching, 47*, 135–173.

Smith, M. W., Brady, J. P., & Anastasopoulos, L. (2008). *Early language and literacy classroom observation tool, pre-K (ELLCO pre-K)*. Baltimore: Paul H. Brookes Publishing.

Yasuda, S. (2011). Genre-based tasks in foreign language writing: Developing writers' genre awareness, linguistic knowledge, and writing competence. *Journal of Second Language Writing, 20*, 111–133.

Zyzik, E., & Polio, C. (2008). Incidental focus on form in Spanish literature classes. *Modern Language Journal, 92*, 50–73.

Charlene Polio is a Professor in the Second Language Studies program at Michigan State University. She is the co-editor of *TESOL Quarterly* and recent past associate editor of *Modern Language Journal*. Her main area of interest is second language acquisition as seen through writing. Her two recent books include *Understanding, Evaluating, and Conducting Second Language Writing Research* (Routledge, with Debra Friedman) and *Authentic Materials Myths* (Michigan, with Eve Zyzik).

Afterword and Next Steps

Margaret E. Malone

Abstract The afterword provides a broad overview of the volume and a forward-looking approach to the implications of the research presented for the field of language teaching. First, this chapter reflects on the papers in the volume and the individual and aggregate contributions. Next, the chapter contextualizes the book and the issues raised in current-day language teaching and learning and the challenges and opportunities facing the community. Finally, the chapter provides recommendations for ways that ongoing research and assessment can continue to support the field.

Keywords Assessment · Teaching · Curriculum · Policy · Language · Pedagogy

1 Introduction

What levels of proficiency do university students attain after specific courses of study and across different skills? How can such results provide positive washback on language teaching in higher education? How can we use these results to help students, instructors, and administrators alike to understand what are reasonable expectations for language learning and teaching? These are critical questions for the language learning and teaching field, and questions that have been left unanswered via any kind of organized approach for decades. The chapters in this volume, both collectively and individually, offer some insight into these questions and explore related questions as well. Moreover, the chapters in this volume provide inspiration and paths forward for researchers who want to explore these questions in their own contexts. Even more so, the chapters investigate a variety of efforts to investigate not only proficiency outcomes, but also deeper issues in language teaching and learning in higher education.

M. E. Malone (✉)
Department of Linguistics, Georgetown University, Washington, DC, USA

American Council on the Teaching of Foreign Languages (ACTFL), Alexandria, VA, USA
e-mail: malonem@georgetown.edu

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_16

Writing this afterward is a professional joy, because the field of world language learning, teaching, and testing in the United States has desperately needed more research on student outcomes for decades; as Winke, Gass and Heindrich pointed out, many applied linguists consider Carroll's 1967 article the last definitive work on the topic. Thus, a volume based on several universities' explorations of language outcomes allows the topic to be addressed across multiple aspects of the higher education enterprise. In this volume alone, the authors investigate and describe the context of language learning in the United States and in higher education, the multifaceted and interdependent roles of curriculum development and instructor professional development, general proficiency assessment as well as assessment of reading, listening, speaking, outcomes in specific languages, the relationship between proficiency and program evaluation, as well as how learners view language teaching, learning, and assessment. The table of contents alone showcases the breadth and depth of research on world language learning as well as a diversity of authors and perspectives on such efforts. The research and reflection put forth in this volume emphasize the strong interest in and commitment to world language learning in higher education.

Despite the well-articulated importance of language learning for commerce, diplomacy, security, and education in general, federal funds for language teaching and learning, after a post September 11th upswing, have diminished in recent years. As Damari et al. (2017), Callahan and Gándara (2014), Jackson and Malone (2009), and others have emphasized, there is a need for individuals with a documented proficiency in another language to fill jobs in U.S. business as well as federal employment and security. In addition to the well-documented practical needs for bi- and multilinguals in the workforce, the recent American Association for the Advancement of Sciences report (AAAS, 2016) noted that the ability to use more than one language is a critical twenty-first century skill; globalization has made the world smaller and more amenable for international travel and communication. Furthermore, a great deal of empirical research supports the cognitive benefits that language learning conveys. Additional empirical research demonstrates a healthy relationship between language study and academic student performance in other areas, including English and math (Olsen & Brown, 1992). Adesope, Lavin, Thompson, and Ungerleider (2010) conducted a meta-analysis that examined the cognitive benefits of bilingualism, demonstrating that bilingualism goes hand-in-hand with increased attentional control, working memory, metalinguistic awareness, and abstract and symbolic representation skills. Thus, the skill of being able to not only speak more than one language but also to be able to read and write in another language represents an important domain for students beyond simply learning language.

Although research shows that learning more than one language, specifically during the early years, confers academic and cognitive benefits, and that language proficiency confers economic benefits to the individual and to society at large, language enrollments in K-12 schools (ACTFL, 2010; Pufahl & Rhodes, 2011) and in higher education (Goldberg, Looney & Lusin, 2015) continue to fall. These documented falling enrollments mean that the number and variety of language courses are fewer, and, thus, there are diminishing opportunities for students in the United States to

develop the high levels of proficiency needed to obtain the kind of employment detailed by Damari et al. (2017) and others. This disconnect is puzzling: If proficiency in world languages can secure well-paying jobs and support the general U.S. economy and security, why are language enrollments falling? Perhaps it is because of the public's limited understanding of language and the lengths of sequences needed to attain specific levels of proficiency; it may certainly be related to the lack of understanding of the proficiency needed to perform in the positions and roles needed. In a test-driven society, why is there not more attention paid to language and its advantages? Is it because U.S. society not only does not understand the advantages that language learning confers and the time needed to learn language, but also what it means to learn language and what language outcomes are and can be after specific learning sequences?

2 Reflections on This Volume

As Winke, Gass and Heidrich pointed out, there has been no national study of language outcomes in higher education since Carroll's (1967) study of foreign language majors across the United States. Although none of the chapters in this volume address specifically why these data have not been gathered, it is likely due to the decline of language study in higher education and a decline in the resources needed (including expertise, time, and cost) to conduct such studies. Thus, this volume, and specifically Winke, Gass and Heidrich's chapter, represent an important step toward beginning to replicate Carroll's work: What level(s) proficiency do majors attain? Other chapters in this volume address an even more important issue for language teaching and learning: What levels of proficiency do students enrolled in language courses attain at different points in their study? Are there differences across modalities (listening, speaking, reading)? When compared to the documentation of and research on the professional needs for proficiency, do our higher education students make the grade? And if they do not, how do we address this gap?

The other chapters in this volume explore additional issues in language learning within higher education, issues that extend past questions of proficiency outcomes. Cox, Brown, and Bell examine an issue in language testing that has been ignored for a number of decades: Do students with high levels of proficiency (ACTFL Advanced and Superior) do as well on tests with the questions in the target language instead of the L1? In which language should test questions be written and should it differ based on test takers' global proficiency? Although Cox, Brown, and Bell's outcome is different from Shohamy's (1984) results and largely inconclusive, the effort taken is important and the research methodology will be easily replicated in future research. In addition, the authors showed that many applied linguists' assumptions about language testing, including the language(s) in which test questions should be provided, are still topics to investigate and would benefit from additional focused research.

In this current era, applied linguists hear a great deal about language-student disengagement, in particular during class or during study abroad, because of the students' immediate access to mobile phones and social media and the prevalence of English in such digital tools. Maloney investigated the timely question of student engagement based on the students' digital literacy in the target language. Maloney suggested approaches for supporting students' digital literacy in target language activities so that they can engage in extended activities both during structured language classes and study abroad. Such outreach has the potential to turn the narrative about social media and digital engagement from a con to a pro for language learning.

One interesting and important finding in many studies shows that listening appears to lag behind the development of other skills. Davidson and Shaw (this volume) suggested that reading and speaking gains are associated with gains in listening as well. Perhaps, then, one important outcome from the research in this volume is the interrelatedness of the language skills. Although linguists frequently test and report language outcomes by skill, all skills are important, and their growth is interdependent.

Understanding the proficiency levels and what is attainable during different learning sequences is crucial for language instructors to move their students to high levels of proficiency without expecting either too little or too much from them. Soneson and Tarone described the influence of assessment and follow-up professional development sessions with language faculty. Not only do the assessments show what students have attained, but the follow-up professional development allows the instructors to understand not only actual student results but also the support students need to move to the next sublevel of proficiency. In addition, such professional development allows instructors to work together to share best practices as well as frustrations, limitations, and misconceptions. Vanpee and Soneson conducted what is essentially an extended case study of the influences of assessment, professional development, and professional connections to program changes. Their transparency in reporting results and working with faculty reveals the nature of the honest conversations needed about how and when to move the proficiency needle. The accompanying self-assessment efforts that they showcased (self-assessments that they gave alongside regular proficiency tests at their university) demonstrated how self-assessment, even when students are not entirely accurate in their self-assessment, allows for beneficial self-reflection and planning for improvement. In addition, their work shows, that on even a small scale, intervention and support can help improve teaching and learning, at least inasmuch as it is measured via assessment.

One outcome of the research efforts presented in this volume is simply the sheer number of tests conducted, including the accompanying student self-assessments. Data were also collected on the language teachers: for example, Polio described in her chapter how teachers participated in focus group sessions and reflected on their teaching the learning occurring in their programs. Such student and instructor reflection and professional development opportunities are perhaps the most important washback from this volume. By modeling self-assessment for students and

examining outcomes with instructors, linguists create a culture of language proficiency, transparency, trust, and continuous improvement. Moreover, the strength of using common, reliable instruments, developed, administered, and rated by external raters, shows the validity and generalizability of the results and supports the methodology and outcomes for future studies. In addition, based on the clarity of the methodologies, all of the studies are replicable or partially replicable in a variety of higher education contexts. As Polio pointed out, program evaluation and proficiency assessment are intertwined. Applied linguists and language educators cannot evaluate programs without conducting a variety of language proficiency assessment and examining the outcomes, both quantitatively and qualitatively. As program reaccreditation continues and accompanying ongoing improvement strategies are planned, the use of all types of educational measures (needs assessments, self-assessments, standardized proficiency tests, portfolios, focus-group sessions, exit interviews, and interviews with all stakeholders) will be influential in twenty-first century teaching and learning.

Although the chapters in this volume address some of and go beyond much of Carroll's 1967 work, it is important to contextualize both his article and the current state of language teaching and learning. As one might expect, much has changed in the past 50 years. First, the ILR scale used to rate students' proficiency in 1967 was transformed in the 1980s to better represent the needs of academic audiences. The focus of the scale became no longer solely for the purpose of measuring the language development of government employees, for whom the ILR scale was originally developed. Since 1986, the ACTFL Proficiency Guidelines (ACTFL, 2012) have been used in U.S. higher education and K-12 programs, as well as in the Peace Corps, and in business and industry, to assess language learner outcomes for a variety of purposes, including program evaluation, eligibility for study abroad, teacher certification, employment, course placement, and readiness of volunteers to perform their work in the target language. In addition, the scale has been used to develop not only the original ACTFL Oral Proficiency Interview (OPI) (see https://www.languagetesting.com/oral-proficiency-interview-opi), but, as we have seen in this volume, computer-based, speaking, listening, reading, and writing tests. All of these innovations represent the increasing availability of assessments to help programs, instructors, and learners understand the levels of proficiency attained after different formal, informal, intensive and sporadic, immersive and non-immersive, heritage and new, language learning experiences.

Perhaps the most striking change in the last 50 years is not simply the availability of these tests, but the transformation of the tests themselves. They have moved from a grammar-based testing system, to audio-lingual, to proficiency-based, and are now emerging as a combination of proficiency, project, and task-based language teaching and learning. According to ACTFL, over 34,000 OPIs, 218,000 OPIcs, 9,000 WPTs, 6,300 RPTs, and 2,000 LPTs were administered in 2016 (ACTFL, 2016); in addition, the ACTFL Assessment of Performance toward Proficiency in Languages test (AAPPL), a test with versions for students at different levels of proficiency and varied grade levels (upper middle school through university), was administered to 225,000 students (ACTFL, 2016). That these tests exist and are

used in such volume demonstrates that our approaches to language teaching and testing has changed a great deal since 1967. Indeed, the popularity of these measures shows that the ACTFL Guidelines and proficiency-based assessments have become the ideal, if not the norm, in U.S. higher language education. As a result, language teaching and learning in higher education continues to move forward despite the current, limited state of funding and slowly dropping enrollments.

As described by Gass and Winke in this volume, much of the research reported on in this book was funded through the National Security Education Program (NSEP; https://www.nsep.gov/) and the Language Proficiency Flagship, which is part of the larger Language Flagship, a national K-20 initiative to change the way Americans learn languages (https://www.thelanguageflagship.org/). With NSEP's support, the vision and execution of this body of research will have immeasurable impact on the teaching and learning of world languages. Yet even with these data and NSEP's support, applied linguists still do not have a complete, national picture of proficiency outcomes within higher education. The data represented within the chapters in this volume tend to come from a small sample of universities or colleges and thus cannot be generalized to represent learning outcomes across the whole of the United States. Several of the studies in this book have large (for applied linguistics) sample sizes, but compared to what Americans know about other skill development across the American educational landscape (notably reading and math), applied linguists' knowledge of how Americans learn foreign languages, even with the chapters in this volume, is still rather sparse. The data in this volume show the richness of language-learning information available, information that needs to continually (and even more robustly) be tapped. Beyond NSEP's funding, the studies in this volume were also enriched by generous university-internal and external support, which were dedicated to the projects to pursue efforts in tracking and understanding foreign language development. The question is how to sustain and promote more studies like these, as a significant challenge is large-scale assessment when there is a lack of funding and resources or both.

On the other hand, the research and development that NSEP has spawned represents the deepest and broadest examination of language outcomes from public institutions of higher education since 1967, as well as new information on what contributes to and improves language programs. The research activity in this volume demonstrates the interest, both at the institutions funded by the project and beyond, in this kind of research and applications of the data to improve programs. The research, additionally, has implications for a large educational and lay audience, including community colleges, high schools, middle schools, the government, and private sectors. The data stress the importance and centrality of assessment to measure learning results and to serve as a measure of accountability. However, all proposed changes in language teaching and learning go beyond assessment. Perhaps the most important contribution of the ACTFL Guidelines and accompanying ACTFL OPI tester professional development program is the washback they have on teaching and learning. Thousands of language professional, from ACTFL raters, to K-20 language educators, have used these Guidelines to transform their classrooms

and to help students understand how language learning and the subsequent proficiency attained is useful for real-life tasks.

As individuals read these chapters, they should be cautious about the extent to which one can generalize the data. It has been decades since Carroll's 1967 study, and it is tempting to generalize the results in this volume, both individually and in aggregate, beyond the samples and contexts at hand. The research in this volume represents a great step forward in using proficiency outcomes for improvement of the field. Applied linguists can employ the same instruments and replicate these methodologies in future studies. There is still much work to be done. By reporting the results of future, parallel studies, applied linguists can move toward a full generalization of the results, and hopefully they can also publish an extensive meta-analysis of proficiency results in higher education. Thus, despite the inability to generalize beyond these data, this volume nonetheless makes an important contribution in exploring language proficiency outcomes, in using self-assessment with students, and in developing articulated sequences for professional language-teacher development. Early results show that such efforts may correlate with, and possibly influence, language teaching and learning. The results show what influence proper assessment and reporting of outcomes can have on a program, and how reflection and planning can influence improvement and understanding.

As Vanpee and Soneson stated, in a culture of transparent assessment, all stakeholders benefit. What should applied linguists do next? Below I outline how they can increase language proficiency outcomes by helping stakeholders understand what is possible developmentwise, and how they can take appropriate action to support such outcomes.

3 Recommendations and Next Steps

This volume and the articulated efforts it represents is just a beginning. I make five recommendations: Continued work at the participating institutions; replication of studies; renewed efforts for student self-assessment; increased professional development; and support for a national study or sets of studies.

3.1 Recommendation 1: Continued Work

The three original institutions (Michigan State University, the University of Minnesota, and the University of Utah) have conducted extended and intensive work during the past 4 years. Although the funding was provided to develop and execute the assessment projects, each institution's efforts far exceed the resources provided. In addition to continuing to mine the data each institution has collected, both individually and collectively, each institution has introduced a culture of assessment and improvement to its campus. By continuing these efforts, on perhaps

a smaller scale, students and instructors alike will continue to expect that outcomes will be measured, reported, and discussed, as well as used to maintain excellent work toward improvement. The existing data represent a robust data set from which additional research can and must emerge.

3.2 Recommendation 2: Replication Studies and Sharing of Materials

As mentioned frequently in this afterword, each chapter represents the potential for multiple replications. Although the range of assessments conducted across modalities, languages, and levels may seem daunting to many programs, each chapter shows one or more studies that a program can undertake. As Polio suggested, purposeful sampling of students can support the collection of qualitative as well as quantitative data. Then, the resulting professional development, as Soneson and Tarone described, can lend itself to a culture of transparent assessment. If the original authors can make the instruments they used available to other researchers and practitioners, then future research can show both the results of such efforts and the influence of professional development on language instructors.

3.3 Recommendation 3: Student Self-Assessment

Although student self-assessment may not always be accurate, it still allows students to better understand what is expected of them and helps increase the dialogue about language, language outcomes, and how to improve language learning. In addition to providing more opportunities for students to self-assess, I suggest providing concrete examples of the tasks, functions, and actual language expected at each level of proficiency. Tigchelaar, Bowles, Winke, and Gass (2017) have investigated ways to increase the accuracy of student self-assessment through more structured and longer self-assessment. If such assessments included samples of work and language at a variety of proficiency levels, students may not only more accurately self-assess, but better understand the underpinnings of proficiency.

3.4 Recommendation 4: Increased Professional Development

Unlike K-12 teachers, whose professional development is often monitored at a local and state level, university-level professional development is less structured. In addition, language programs may reside across different departments, thus minimizing the chances of collaboration and development of shared understanding and goals. By using actual assessment outcomes, coupled with student self-assessments, and

combined with reflections on these outcomes, professional development on the ACTFL Proficiency Guidelines (ACTFL, 2012) can help language instructors develop articulated sequences based on empirical data.

3.5 *Recommendation 5: Support a National Study and Sets of Studies*

A national study of representative students at different levels of study, including majors and minors, would go a long way to determining whether the outcomes in this volume are generalizable or not. In addition, if local programs replicate and publish the results of their studies, the field will develop a national picture of student proficiency in higher education.

This volume is only the beginning. Applied linguists have much work to do, and, thanks to the efforts of the authors of the chapters in this volume, future researchers have clear pathways to follow.

References

AAAS. (2016). *America's languages: Investing in language education for the 21st century*. Cambridge, MA: AAAS.

ACTFL. (2010). Foreign language enrollments in K-12 public schools: Are students prepared for a global society? Alexandria, VA: ACTFL. ISBN: 0615408273, 9780615408279

ACTFL. (2012). *ACTFL proficiency guidelines*. Alexandria, VA: ACTFL. Available from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

ACTFL. (2016). *2016 Annual report*. Alexandria, VA: ACTFL. Available at https://www.actfl.org/sites/default/files/reports/annualreport2016/index.html

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207–245.

Callahan, R., & Gándara, P. C. (Eds.). (2014). *The bilingual advantage: Language, literacy and the US labor market*. Tonawanda, NY: Multilingual Matters.

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

Damari, R. R., Rivers, W. P., Brecht, R. D., Gardner, P., Pulupa, C., & Robinson, J. (2017). The demand for multilingual human capital in the US labor market. *Foreign Language Annals, 50*(1). https://doi.org/10.1111/flan.12241

Goldberg, D., Looney, D., & Lusin, N. (2015, February). *Enrollments in languages other than English in United States institutions of higher education, fall 2013*. New York, NY: Modern Language Association. Available at https://www.mla.org/content/download/31180/1452509/EMB_enrllmnts_nonEngl_2013.pdf

Jackson, F. H., & Malone, M. E. (2009). *Building the foreign language capacity we need: Toward a comprehensive strategy for a national language framework*. Washington, DC: Center for Applied Linguistics. Available at http://www.cal.org/resource-center/publications/building-foreign-language-capacity

Olsen, S. A., & Brown, L. K. (1992). The relation between high school study of foreign languages and ACT English and mathematics performance. *ADFL Bulletin, 23*(3), 47–50.

Pufahl, I., & Rhodes, N. C. (2011). Foreign language instruction in US schools: Results of a national survey of elementary and secondary schools. *Foreign Language Annals, 44*(2), 258–288. https://doi.org/10.1111/j.1944-9720.2011.01130.x

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*(2), 147–170. https://doi.org/10.1177/026553228400100203

Tigchelaar, M., Bowles, R. P., Winke, P., & Gass, S. (2017). Assessing the validity of ACTFL can-do statements for spoken proficiency: A Rasch analysis. *Foreign Language Annals, 50*(3), 584–600. https://doi.org/10.1111/flan.12286

Margaret E. Malone (Ph.D., Georgetown University) is Director of the AELRC and Research Professor at Georgetown University and Director of the Center for Assessment, Research and Development at ACTFL. Her current research focuses on language assessment literacy, oral proficiency and intercultural assessment, and the relative difficulty of learning different languages for speakers of different L1s.