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

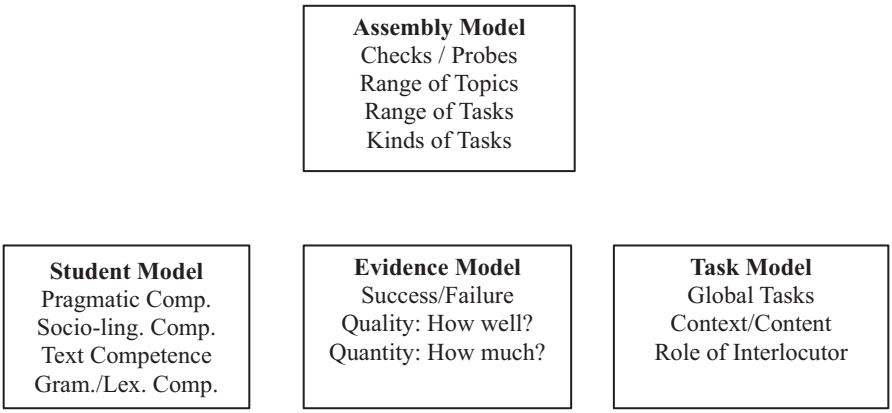


Fig. 1 The structure of the OPI in the ECD framework

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.

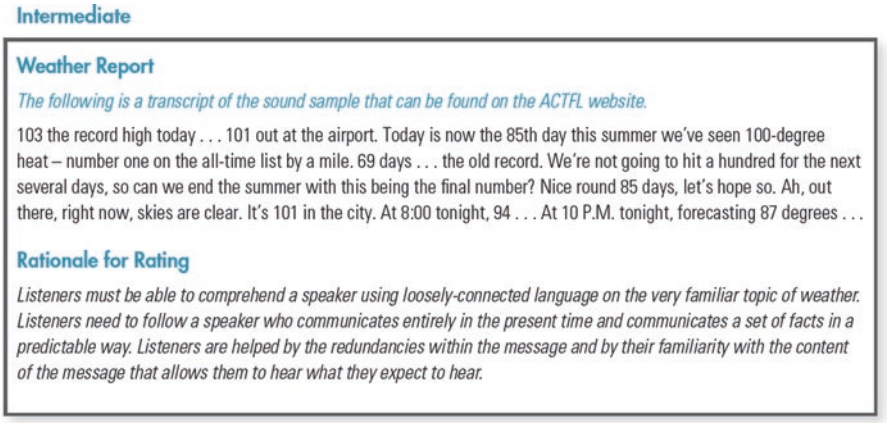
Mislevy et al. (2002) suggest that if we want an assessment to provide valid evidence of the student's abilities, we need to design it from both a task-centered and a construct-centered perspective (p. 493). In the next two sections, we discuss potential task- and construct-centered explanations for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