Assessment Practices in STARTALK Language Programs: A View of Current Language Assessment Literacy

Margaret E. Malone, Megan J. Montee, Francesca DiSilvio

1. *Introduction*

Assessment is essential to education, because it provides information on students' progress toward learning goals. Reliable and valid assessment can provide not only important summative information, but also formative information to instructors and learners on both what has been learned and what remains to be learned. However, in order for assessment to be used effectively, instructors must understand the components of a reliable and valid assessment system and how to incorporate such a system into classroom testing. Many language instructors in the United States may lack basic knowledge of assessment and measurement (Popham, 2009).

Because assessment of world language learning is not currently required by *No Child Left Behind* regulations, there is no national system for K-12 world language assessment. Moreover, there is no clear documentation of the actual language assessment processes and practices in world language classrooms in the U.S. These issues affect the teaching of all languages; among the less commonly taught languages (LCTLs), the challenges are magnified. There is a shortage of nationally available, research-based LCTL assessments across language modalities, which are identified by the 2006 *Standards for Foreign Language Learning in the 21st Century* as Interpersonal Speaking, Interpretive Reading and Listening, and Presentational Writing and Speaking (National Standards in Foreign Language). Furthermore, many LCTL instructors lack the background needed to interpret those tests that are available, or to develop their own reliable and valid classroom tests (Nier, Donovan, and Malone, 2009). This dual lack of materials and of assessment knowledge presents a serious obstacle to documenting language attainment in LCTLs.

Since 2007, the U.S. Department of Defense has funded STARTALK, a program that provides summer learning opportunities in languages deemed critical to U.S. security. In 2010, over 5,000 students in grades K-16 participated in summer language programs in Arabic, Chinese (Mandarin), Dari, Hindi, Persian, Swahili, Turkish and Urdu. STARTALK programs are short-term, varying in

length from two weeks to two months, although many have emerged from existing academic-year LCTL programs. Each potential program applies for STARTALK funding individually. Funded programs develop their own curricula and assessment plans based on STARTALK requirements, and all programs report their assessment plans through required curriculum documents. Given STARTALK's centralized reporting requirements, an examination of these curricula and their included assessment plans provides an opportunity to investigate the assessment practices of a large group of LCTL programs.

This paper provides a systematic analysis of the formative and summative assessment practices of STARTALK programs and the extent to which these practices support stated program goals. Furthermore, it also discusses how increased language assessment literacy could help STARTALK and academic-year programs improve their assessment practices. The paper begins with an explanation of the two main types of assessment used in STARTALK, formative and summative assessment. It then provides background on language assessment literacy, followed by a description of STARTALK's relevance to assessment practices in world language programs nationwide. Next, the paper describes the methodology and results of a study conducted to analyze assessment practices across 2010 STARTALK student programs. Finally, it suggests ways that assessment literacy can be improved in STARTALK programs in particular, and U.S. world language programs in general.

2. *Formative and Summative Assessment*

Most assessment is categorized according to how the results will be used. For example, language assessments can be used to make placement decisions, to certify instructors, or to evaluate the outcomes of a program, among other purposes. Assessments can also be classified on a continuum from high stakes (summative assessments) to low stakes (formative assessments) (Stoynoff & Chapelle, 2005). High stakes or summative language assessments are generally used to make decisions, including whether a prospective teacher has sufficient language proficiency to receive certification to teach that language in a specific state. Other uses for summative assessments include assigning grades, determining whether graduation requirements have been met, or allowing admission to a specific language program for a course of study. On the other end of the spectrum, low stakes or formative assessments are generally used to provide information to students and instructors on progress toward specific learning goals in a course (Stoynoff & Chapelle, 2005). However, the dichotomy presented by these definitions is somewhat misleading; clearly, formative

assessment should be aligned to the goals of summative assessment. Similarly, summative assessments should reflect the learning goals of formative assessments. Such an alignment can improve the likelihood that formative assessments reflect progress toward the outcomes ultimately attained on summative assessments.

Although summative assessments can refer to standardized, nationally-available tests, many language instructors develop their own summative assessments to provide the basis for grades and to demonstrate the attainment of course outcomes. Many STARTALK programs provide students with grades or course credits, but these results are not the only reason that summative assessment is important for STARTALK programs. It can also provide important feedback for program staff and local stakeholders. Furthermore, STARTALK summative assessments in the aggregate help to reveal what can be taught and learned in this type of short-term language program.

Developing formative and summative assessments can be a challenge for short-term language programs. While most of us are familiar with the multiple-choice approach used on most standardized tests in the U.S., STARTALK programs, like many year-long programs, often rely on performance assessments to gauge student outcomes. In contrast to multiple-choice tests, performance assessments require and expect students to use language in real-life situations; demand that students develop an understanding of what real-life situations are like in the language; and are based on performance or the active construction of language, rather than a demonstration of passive understanding only (Norris et al., 1998; Sandrock, 2010). Examples of performance assessments are numerous; for example, students might conduct **role plays**, in which they work with another student or students to simulate real-life situations. Such role plays are largely spontaneous, unlike a formal student **presentation**, which, as a performance assessment, generally requires planning. Other examples of performance assessments include **interviews,** in which a student participates in an interview with the teacher or another student. All performance assessments should be developed and rated according to a specific set of criteria that reflect the goals and purposes of the course.

3. *Language Assessment Literacy*

Assessment is essential to determining and documenting the outcomes of any language learning experience (Brown, 2004; Hughes, 2003). Furthermore, assessment results can provide valuable information for all participants in the language learning process (Shepard, 2000). Despite its acknowledged importance,

however, many administrators, instructors, and students do not understand the basics of assessment (Popham, 2009). As the central role of testing in the U.S. educational system has only expanded under *No Child Left Behind*, the gap between what stakeholders need to know about assessment and what they do know must be closed.

The term "assessment literacy" refers to familiarity with assessment processes (Stiggins, 2001). Robust language assessment literacy combines an understanding of testing concepts with knowledge of both language acquisition and language teaching methodologies. Recently, a number of researchers (McNamara and Roever, 2006; Inbar-Lourie, 2008; Malone, 2008; Taylor, 2009) have discussed the importance of language assessment literacy as well as strategies for improving it among stakeholders. Two recent studies (Nier, Donovan and Malone, 2009; Riestenberg et al., in press) have focused specifically on language assessment literacy for instructors of less commonly taught languages. Such publications suggest that researchers largely appreciate the importance of promoting assessment literacy among world language educators. However, there is still little documentation or analysis of current assessment knowledge and practice among world language instructors. Additionally, little research has yet been conducted on how exactly to promote assessment literacy in order to increase positive washback (Hughes, 2003) for world language teaching and learning.

One challenge to defining and addressing language assessment literacy is that, while there is nearly universal acceptance of the need for assessment literacy among language testing experts, and efforts to examine and recommend improvements to existing formal language testing courses, there is not yet any agreed-upon process for attaining different levels of language assessment literacy. The remainder of this paper describes a study that addresses one way to examine current levels of assessment literacy in a national language program.

4. *The STARTALK Initiative*

STARTALK is a U.S. presidential initiative directed toward teaching languages deemed critical to national security. STARTALK programs provide summer language instruction for students in grades K-16. STARTALK's stated mission is

> …to increase the number of Americans learning, speaking, and teaching critical need foreign languages by offering students (K-16) and teachers of these languages creative and engaging summer experiences that strive to exemplify best practices in language education and in language teacher development, forming an

> extensive community of practice that seeks continuous improvement in such criteria as outcomes-driven program design, standards-based curriculum planning, learner-centered approaches, excellence in selection and development of materials, and meaningful assessment of outcomes (National Foreign Language Center, 2011).

The STARTALK programs, which began in summer 2007, have added new languages and student grade levels each year. In 2011, STARTALK languages will include Arabic, Chinese (Mandarin), Dari, Hindi, Persian, Portuguese, Russian, Swahili, Turkish, and Urdu. There are two types of STARTALK programs: language learning programs for students, and professional development programs for teachers of these critical languages. Since 2007, STARTALK has served over 7,000 students learning critical languages, and 3,000 teachers and prospective teachers of these languages (National Foreign Language Center, 2011).

In addition to providing language learning and professional development experiences in critical languages, the STARTALK initiative has developed consistent standards and procedures by which participating programs must operate, beginning with their initial applications and ranging all the way through the final reporting process (see also Ingold, this volume). As stated in its mission, STARTALK supports the development and execution of excellent summer language experiences that embody the best of current research and practice. To this end, the standard documents and procedures have been developed based on both government reporting requirements and current best practices for language education. All funded programs are required to develop, submit, and, if necessary, revise a curriculum template. This curriculum template mandates that each program describe in detail its learning goals and objectives and how they will be assessed using both formative and summative tools.

The STARTALK initiative therefore presents a rare grouping of a large, diverse collection of less commonly taught language programs that utilize common standards and forms to describe their curricula and assessment practices. It therefore provides an optimal opportunity to investigate how assessment practices are described and used in LCTL programs across the U.S. The next sections of this paper describe the results of such an investigation, including research questions and a methodology for identifying assessment practices in STARTALK programs.

5. *Methods*

In 2010, researchers conducted a descriptive and qualitative study of the self-reported assessment practices of the 2010 STARTALK student programs. By both describing the types of assessment practices reported in STARTALK curricula and examining the relationship between assessment practices and instructional goals, the study provided a picture of assessment practices across STARTALK programs, while also identifying areas where additional training and resources may be needed. The study had two phases: the first examined assessment activities and patterns across STARTALK programs, and the second provided a qualitative analysis of the ways in which formative and summative assessment was used to measure curriculum goals. Research questions for the study were:

1. What types of assessment activities do STARTALK programs report using?
2. What patterns emerge across languages, proficiency levels, and grade-level clusters?
3. To what extent are STARTALK programs assessing stated curriculum goals?
4. How are STARTALK programs using formative and summative assessment and how do these reflect current levels of assessment literacy in STARTALK programs?

The researchers answered questions 1 and 2 through a content analysis of the *Student Program Curriculum Template and Guide* (referred to in this paper as simply "curriculum template") submitted by the 156 STARTALK student programs that operated in 2010. Each curriculum document contained information on the assessment activities that students would be expected to perform as evidence of learning. Section F of the curriculum template, "End of Program Performance Tasks," asked programs to list the culminating performance tasks that students would complete to demonstrate their achievement of the program learning objectives. Section G asked programs to list "Other Types of Assessment and Evidence of Learning." Items in sections F and G were first coded into emerging categories, each representing a different type of assessment activity; all coding was checked by two additional researchers. In case of disagreement, the coders discussed and reached consensus. In all, 34 assessment activity codes were created; each curriculum document contained 3-19 codes, depending on the amount of detail included. Then, to address research question 2, coded data were broken down by program language, proficiency level, and grade level for further analysis.

In the second phase of the study, the curriculum documents were reviewed as part of a qualitative analysis. A researcher reading the documents identified a set of preliminary themes for research questions 3 and 4. These preliminary themes were added to and revised during the analysis process, and curriculum documents were iteratively compared to the emerging themes to ensure that the descriptions accurately captured the data. The curriculum documents were then reviewed again during analysis to ensure the trustworthiness of the data. Finally, two additional members of the research team familiar with the curriculum documents checked the validity of the findings and recommendations. This cyclical method of analysis resulted in a final set of major themes related to each research question, as well as recommendations based on the findings.

6. *Results*

Research Question 1

The first research question asked: What types of assessment activities do STARTALK programs report using? To address this question, the researchers coded and quantified the different types of assessment practices reported in the curriculum templates, including classroom-based assessments and standardized tests. Table 1 shows the frequency with which various types of assessment practices were reported across the 156 STARTALK programs by raw number and percentage of total programs. As Table 1 below shows, the assessment activity reported by the greatest number of programs was role play, followed by presentation, and then reading comprehension. Overall, oral assessments, including presentations, role plays, interviews, and assessments of oral comprehension, were among the most frequently reported types of assessment. Programs also reported the use of standardized assessments. A small number of programs (17 total, or 11%) reported using nationally available standardized tests; in some cases, programs reported using more than one standardized test. Table 2 shows these tests and the number of programs that reported using each. Furthermore, a total of 26 programs (17%) reported general use of tests or exams, most likely indicating that some program administrators or instructors developed tests internally.

Research Question 2

The second research question asked: What patterns emerge across languages, proficiency levels, and grade-level clusters? To address this question, assessment

activities were analyzed for patterns within program target language, proficiency level, and grade-level clusters.

Table 1. Assessment practices across programs

	Total	Percentage
Role play	135	86.5
Presentation	128	82.1
Reading comprehension	115	73.7
Poster/graphic project	101	64.7
Interview	97	62.2
Oral comprehension	94	60.3
LinguaFolio	82	52.6
Skit/performance	75	48.1
Other project	70	44.9
Writing/composition	68	43.6
Podcast/video	61	39.1
Information gap	60	38.5
Song /rhyme	57	36.5
Journal	54	34.6
Venn diagram	54	34.6
Informal observation	39	25.0
Research project	36	23.1
Games	34	21.8
Field trip	32	20.5
Quiz	31	19.9
Worksheet/assignment	29	18.6
Test/exam	26	16.7
Portfolio	22	14.1
Survey	20	12.8
Dialogue	19	12.2
Nationally available standardized tests	17	11.0
Food/cooking project	15	9.6
E-portfolio	14	9.0
Oral proficiency interview	14	9.0
Self-assessment	11	7.1
Computer/technology	8	5.1
Student participation	4	2.6

Table 2. Nationally available standardized tests used by STARTALK programs

	Total
Computerized Assessment of Proficiency (CAP)	5
Standards-based Measurement of Proficiency (STAMP)	5
Assessments provided by the Center for Advanced Study of Language (CASL)	4
ACTFL Assessment of Performance toward Proficiency in Languages (AAPPL)	3
Computerized Oral Proficiency Instrument (COPI)	1
ALOFA (Arabic Language Oral Fluency Assessment)	1

Assessment Practices by Target Language

Assessment practices within language groups were analyzed for Arabic, Chinese, and Hindi programs. Persian, Swahili, Turkish, and Urdu programs were not included in this part of the analysis, as there were fewer than ten programs in these languages, making consistent patterns difficult to identify. Table 3 lists the top five assessment practices for each language, along with the percentage of programs within that language that reported using this practice.

Consistent with the most frequently reported assessment practices for all languages noted above, the top assessment activities by target language included role play, followed by presentation (equal in frequency to role play within Chinese programs and to oral comprehension within Hindi programs), and reading comprehension for Arabic and Chinese (equal in frequency to oral comprehension within Arabic programs). For Arabic and Hindi, all or almost all programs reported using role plays and presentations. There was less uniformity in reported assessment practices across Chinese programs, perhaps due to the larger number of Chinese programs represented in the data. Notably, Arabic diverged from the other programs in its relatively higher frequency of quiz, test/exam, and worksheet assessment activities. More Arabic programs also reported the use of standardized tests (23%). By contrast, no Hindi program reported using a standardized test (possibly because few standardized tests are available in Hindi). Furthermore, although there were fewer Hindi programs represented in the data, it may also be worth noting that 75% of Hindi programs reported the use of journals compared with 40% of Arabic programs and 24% of Chinese programs.

Assessment Practices by Proficiency Level

Assessment practices were also analyzed by proficiency level. In its initial funding application, each program was required to indicate the proficiency level of the students at which it would aim its courses. Most, but not all programs, reported this information in terms of the ACTFL proficiency guidelines

(American Council for the Teaching of Foreign Languages, 1999). For analysis, programs were grouped in the four categories listed in the table below.

Table 3. Top five assessment practices by target language

Arabic (N=30)	Chinese (N=91)	Hindi (N=16)
Role play (93.3%) Presentation (83.3%) Oral comprehension (70%) Reading comprehension (70%) LinguaFolio (66.7%)	Role play (79.1%) Presentation (79.1%) Reading comprehension (75.8%) Poster/graphic project (65.9%) Interview (57.1%)	Role play (100%) Presentation (87.5%) Oral comprehension (87.5%) Interview (81.3%) Poster/graphic project (81.3%)

Table 4 below shows the top five most frequently reported assessment practices for each proficiency level cluster. Table 4 demonstrates that assessment practices do not seem to vary greatly according to proficiency level. Role plays, presentations, and reading comprehension assessments were reported at all levels. However, posters and graphic projects were more common at the Intermediate and Advanced levels, while oral comprehension and interview assessments were more common for the Novice learners.

Assessment Practices by Grade Level

Finally, assessment practices were analyzed by program grade level; all programs listed at least one grade level and some programs fit into more than one grade level cluster. Table 5 shows the top five assessment practices used with each grade level cluster. As Table 5 below shows, all grade level clusters reported frequent usage of role plays. Within the early elementary cluster, oral comprehension was the most frequently reported assessment activity; it ranked third most frequent among upper elementary programs. Interviews were among the most frequent assessment activities reported by middle school, high school, and college programs, but not elementary programs. As might be expected, reading comprehension was not frequently reported as an assessment activity by programs serving early elementary students. Assessment of reading comprehension played a relatively steady role in the four higher grade clusters, though, ranging from use by 64% to 78% of programs at these levels. Writing assessments were reported by 20% of early and upper elementary school programs, and this number increased with each successive grade cluster; at the college level, writing assessments were reported by 73% of programs. The elementary clusters reported little usage of standardized tests or of internally

developed, classroom-based exams. The use of testing, both standardized and classroom-based, was most frequently reported at the middle and high school levels.

Table 4. Top five assessment practices by proficiency level

Novice (N=94)	Both Novice and Intermediate (N=32)	Intermediate (N=20)	Both Intermediate and Advanced (N=10)
Role play (84.0%)	Role play (93.8%)	Presentation (95.0%)	Role play (90.0%)
Presentation (78.7%)	Presentation (84.4%)	Reading comprehension (85.0%)	Poster/graphic project (90.0%)
Reading comprehension (68.1%)	Reading comprehension (81.3%)	Role play (85.0%)	Presentation (80.0%)
Oral comprehension (63.8%)	Interviews (68.8%)	Poster/graphic project (75.0%)	Reading comprehension (80.0%)
Interviews (60.6%)	Oral comprehension (62.5%)	LinguaFolio (70.0%)	Interviews (60.0%)
			Skit/ performance (60.0%)
			Other project (60.0%)

Research Question 3

The third research question asked: To what extent are STARTALK programs assessing their stated curriculum goals? As part of their initial funding application, each program was required to identify curriculum goals related to each of the five *Standards for Foreign Language Learning* (National Standards in Foreign Language, 2006): Communication, Cultures, Connections, Comparisons, and Communities. To address the third research question, each program's stated goals were compared with its assessment practices. Additionally, the researchers also considered each program's declared *Standards*-based goals in relation to its entire curriculum document in order to glean other relevant contextual

information about instructional materials and practices, program themes, and the specific knowledge and skills covered by the program. As described in the methods section above, data for research questions three and four were analyzed qualitatively and are presented descriptively by theme.

Table 5. Top five assessment practices by grade level

Early Elementary (N=18)	Upper Elementary (N=36)	Middle School (N=51)	High School (N=122)	College (N=11)
Oral comprehension (94.4%)	Role play (80.6%)	Role play (82.4%)	Role play (88.5%)	Role Play (100.0%)
Role play (72.2%)	Presentation (72.2%)	Poster/graphic project (78.4%)	Presentation (86.1%)	Presentation (90.9%)
Other project (66.7%)	Oral comprehension (69.4%)	Presentation (78.4%)	Reading comprehension (76.2%)	Interview (81.8%)
Poster/graphic project (66.7%)	Reading comprehension (69.4%)	Reading comprehension (78.4%)	Interview (63.9%)	Skit/ performance (81.8%)
Song/rhyme (66.7%)	Poster/graphic project (63.9%)	Interview (56.9%)	Poster/graphic project (63.1%)	Information gap (72.7%)
		LinguaFolio (56.9%)		Writing/ composition (72.7%)

Assessment and the Standards

Programs varied in how they reported goals and expected outcomes. While some programs included multiple detailed outcomes for each *Standard*, other programs reported general outcomes across *Standards*. Furthermore, when reporting their assessment practices, many programs only discussed the *Communication Standard*, making no mention of how they might assess the other four. This indicates that programs may not be planning assessments for areas that are more difficult to measure, such as culture or interest in future language study.

While programs varied in the specificity of their responses, most programs included some detail related to specific program content or instructional activities with mention of general proficiency-based outcomes rather than program-specific outcomes aligned with proficiency levels.

Level of Planning

The second theme that emerged from the analysis relates to the level of planning in which programs engaged relative to their assessment tasks. Programs typically described their assessment plans with less detail than their expected outcomes and planned instructional activities. For example, in Section E of the template, "Specific Knowledge and Skills," an expected outcome listed by one Arabic language program indicated that students would be able to "introduce themselves and others and give appropriate greetings and responses," and then provided a list of the specific vocabulary students would use for this function. This information was explicitly requested in the curriculum template and most programs provided such detailed responses. This same program, however, reported a planned assessment task (Section F) in which students would "produce recordings of short conversations and dialogues about Greetings/Introductions." Though this planned assessment relates directly to the course material, the program provided no information about the quantity and quality of language that the students would be expected to produce, or how student responses would be scored on this task. The template, in fact, does not require programs to list any explicit assessment criteria, and most programs did not volunteer such information. Overall, few programs explained how students would be expected to respond to assessment tasks, the criteria for successful task completion, or how the tasks would be scored or rated. For example, only a small number of programs mentioned the use of rubrics to score performance assessments. This indicates that programs may not yet be defining explicit assessment criteria or systematically scoring tasks.

Programs' general lack of firm assessment plans suggests that they may not be implementing backward design (Wiggins and McTighe, 2005), a curriculum planning method in which a final summative assessment is designed first, followed by plans for the lessons that will directly lead up to this assessment. CAL's experience in teaching assessment training courses to STARTALK instructors supports the observation that few instructors employ the backward design strategy. The level of assessment planning shown in the curriculum templates also demonstrates a need for increased assessment literacy.

Proficiency Levels

The third theme that emerges from the analysis relates to the degree to which programs accounted for their students' level of language proficiency when designing assessment practices and predicting outcomes. Analysis showed that outcomes and assessment plans were often less focused on proficiency levels than on the *Standard* being taught. (Of course, the alignment to the *Standards* is likely related to the structure of the curriculum template, which emphasizes themes and *Standards*.) In some cases, programs' reported assessment tasks were clearly inappropriate for their students' proficiency level, such as an essay writing task for a program with Novice-level students. In most cases, however, programs simply did not report assessment information.

The curriculum template also gives comparatively little attention to the role of proficiency in assessment design. This lack of focus on proficiency level points to larger issues within STARTALK programs related to proficiency, assessment, and the proficiency outcomes that can be expected for short-term language programs. Several ongoing projects are exploring standardized proficiency measures that could be used across programs. Still, data from the curriculum templates indicate that programs are not currently focused on aligning assessment tools and program outcomes with proficiency levels.

In summary, the researchers' analysis revealed that STARTALK programs' assessment plans were generally less developed and less detailed than their instructional plans. This may reflect both the limitations of the curriculum template as well as a generalized lack of assessment planning among programs. Finally, while assessment plans did reflect *Standards* and program themes, they were often disconnected from their students' proficiency levels and desired proficiency outcomes.

Research Question 4

The fourth research question asked: How are STARTALK programs using formative and summative assessments? While the results of Research Questions 1 and 2 provided a detailed list of the types of assessments programs reported using, this data did not show the extent to which the reported assessment activities were being used summatively or formatively. In the curriculum documents, programs were asked to list formative and summative practices separately in different sections (Sections F and G, respectively). However, analysis during the initial phase of the study indicated that reporting practices were inconsistent, rendering straightforward coding into formative and summative categories impossible. Given this limitation, the researchers instead

reviewed assessment practices for themes related to formative and summative purposes.

This qualitative analysis showed that programs reported formative assessments to a greater extent than summative assessments. In Section F of the curriculum template, programs were instructed to report summative performance tasks, defined in the template as "culminating performance tasks [that] will provide evidence that students have achieved the program learning objectives." However, programs often used this section to describe activities generally categorized as formative rather than summative. For example, programs often reported daily assessment tasks in Section F, though these are generally considered formative in nature, as they inform the day-to-day planning of classroom activities. Many programs also used Section F to report project-based assessments, often conducted in groups. While such activities between students provide excellent language practice and solid formative assessment information, the programs' reports generally do not indicate that they were developed to be particularly summative. Overall, very few programs used Section F to describe summative performance tasks intended as final assessments of students' proficiency or achievement.

Additional analysis suggests that many programs used neither formative nor summative performance task-based assessments. Performance tasks are authentic and contextualized assessments of language use (Norris et al, 1998); most programs' assessments did not require students to use the target language in any such authentic context. Thus, while programs are moving beyond traditional assessment practices such as multiple-choice testing, it is unlikely that such programs are implementing a task-based approach to assessment.

Finally, programs generally did not report the criteria they used to conduct formative or summative assessments. Rubrics and scoring procedures were not frequently mentioned in the curriculum template, and no clear system for integrating formative and summative assessment emerged. This suggests that, although activities such as role plays were occurring between students and sometimes between students and instructors, clear criteria for evaluating student performance on such activities were not developed. Without such criteria, these activities cannot provide helpful and valid formative or summative assessment.

7. *Discussion*

The STARTALK curriculum templates provide a useful source of information about the range of assessment activities used by STARTALK programs. Because this information is reported on a standardized form that includes guidelines and

suggestions, data were uniform and may not represent the full range of assessment practices included in STARTALK programs. Additionally, information on how these practices are being implemented, rather than simply how they are being planned, is not consistently captured in the template, and is therefore not reported here. Despite these limitations, however, this analysis revealed several trends in the self-reported data. The data provide important information about current understanding of language assessment literacy in STARTALK programs as reflected in the STARTALK curricula documents.

Focus on Oral Assessment and Communication

Assessments of oral skills including presentations, role plays, interviews, and assessments of oral comprehension were among the most frequently reported assessment practices across STARTALK programs. This remained true when the data were analyzed by language, proficiency level, and grade-level cluster. These assessments cover both Interpersonal and Presentational Speaking, indicating that many programs are assessing both Communicative Modes. Interviews were used less frequently to assess oral skills, possibly because this method may be more time-consuming for teachers to conduct and more challenging to assess than presentation or role play tasks. The overall focus on oral proficiency may indicate that programs need additional resources for assessing literacy skills, particularly writing. While reading comprehension was assessed frequently across programs, writing assessment was limited in the lower grade level clusters. In addition, few programs reported how the assessments were being rated and how the results were being reported, which are two crucial aspects of assessment.

Patterns by Grade Rather than Proficiency Level

Consistent patterns in assessment practices emerged by grade level, while patterns across proficiency levels seemed less systematic. This may indicate that the target grade level(s) of a program play a greater role in determining their choice of assessment practices than do the target proficiency level(s). For example, while assessment activities such as games and songs were frequently reported by lower and upper elementary programs, these activities decreased in frequency with increases in grade level. The data also show that different assessment practices are being used across a variety of proficiency levels. For example, interviews were reported at relatively high rates across all proficiency levels. Additional resources may be needed to help programs to understand how to adapt these practices for target proficiency levels in order to make them level-

appropriate. These results also indicate that professional development for STARTALK program directors and instructors by target grade level and proficiency level may be useful.

Limited Use of Standardized Tests

Programs reported limited use of standardized testing. In many STARTALK languages or for certain grade levels, there may be limited standardized materials available. Programs may also need guidance in how to use standardized tests appropriately in short-term programs, as these types of tests may have limited uses in such contexts. This finding highlights the need for projects such as the present study that examine the usefulness and efficacy of testing in STARTALK programs.

The three main findings indicate that STARTALK programs report using a variety of assessments, yet explain little about how these assessments are rated, outcomes reported, and results used for program improvement and reflection. Therefore, professional development and increased awareness of language assessment fundamentals are likely needed for many STARTALK language programs.

8. *Recommendations to Improve Assessment Literacy*

Based on the analysis of the curriculum documents and assessments, recommendations can be made for improving assessment literacy among stakeholders in STARTALK programs in particular and U.S. world language programs in general.

Facilitate Alignment between Outcomes, Instruction, and Assessment

The current STARTALK curriculum template includes sections for programs to describe the outcomes they expect to achieve, their instructional practices, and the assessments they plan to use. For STARTALK, the authors recommend that the curriculum template be modified so that for each *Standard*, programs must list one or more expected outcomes as well as what assessment tools will be used to measure students' progress toward these outcomes.

It is also recommended that year-long and STARTALK programs alike investigate ways to include professional development training on the alignment between curriculum and assessment. The researchers recommend that all world language teachers participate in relevant training on developing, administering, reporting, and explaining the results of contextualized, authentic tasks that measure stated program outcomes.

Emphasize proficiency levels and understanding

Currently, the STARTALK programs' reported assessment practices reflect the *Standards* and incorporation of program themes. However, these programs have not yet aligned these practices with the *ACTFL Proficiency Guidelines*. A similarly limited relationship exists between the Guidelines and the programs' expected learning outcomes. While no data exist to confirm a similar gap between expected outcomes and assessment practices and the *ACTFL Proficiency Guidelines* in school-year programs, the researchers believe that additional resources and professional development on proficiency levels may also be useful for these world language instructors. Developing prototypical tasks and benchmark samples of can-do tasks could greatly help world language instructors better understand the proficiency levels.

Focus on assessment criteria, scoring, and feedback

In order for assessment tasks to be meaningful and useful, methods for scoring and providing feedback should be planned out as part of the initial assessment planning process. Very few STARTALK programs reported the criteria by which they score performance tasks. STARTALK programs and academic-year world language programs alike would benefit from performance task rubrics that could be adapted and shared across programs in the U.S. Stakeholders could also benefit from professional development on criteria, scoring, and feedback that includes benchmark samples; instruction on how to score performance tasks; and tips for providing feedback to students, instructors, and other stakeholders.

STARTALK provides language learning opportunities for increasing numbers of students every year, and its centralized structure has provided a window into its programs' current, self-reported assessment practices. The current study suggests that STARTALK instructors, many of whom are language instructors during the school year, would benefit from proficiency-oriented, developmentally appropriate professional development on language assessment. While such measures will certainly assist the STARTALK program, the same activities also have the potential to improve world language learning across the U.S. For world language programs in general, the recommendations suggested in this paper provide steps to improved assessment literacy that will close the gap between what world language instructors need to know about assessment and what they do know.

References

American Council on the Teaching of Foreign Languages. 1999. "ACTFL Proficiency Guidelines- Speaking: Revised 1999." http://www.actfl.org/files/public/Guidelinesspeak.pdf

Brown, H. Douglas. 2004. *Language Assessment: Principles and Classroom Practices.* White Plains, NY: Pearson Longman.

Hughes, Arthur. 2003. *Testing for Language Teachers*. 2nd ed. Cambridge: Cambridge University Press.

Inbar-Lourie, Ofra. 2008. "Constructing a Language Assessment Knowledge Base: A Focus on Language Assessment Courses." *Language Testing* 25: 328-402.

Malone, Margaret. 2008. "Training in Language Assessment." In *Encyclopedia of Language and Education* (2nd ed): *Vol. 7 Language Testing and Assessment*, edited by Elana Shohamy, and Nancy Hornberger, 225-239. New York: Springer Science and Business Media, Inc.

McNamara, Tim, and Carsten Roever. 2006. *Language Testing: The Social Dimension*. Oxford: Blackwell.

National Foreign Language Center. 2011. *STARTALK – Start Talking!* http://startalk.umd.edu/

National Standards in Foreign Language. 2006. *Standards for Foreign Language Learning in the 21st Century*. Lawrence, KS: Allen Press, Inc.

Nier, Victoria C., Anne E. Donovan, and Margaret E. Malone. 2009. "Increasing Assessment Literacy among LCTL Instructors through Blended Learning." *Journal of the National Council of Less Commonly Taught Languages* 7: 105-131.

Norris, John M., James Dean Brown, Thom Hudson, and Jim Yoshioka. 1998. *Designing Second Language Performance Assessments*. Honolulu: University of Hawaii Press.

Popham, W. James. 2009. "Assessment Literacy for Teachers: Faddish or Fundamental?" *Theory into Practice* 48, no. 1: 4-11.

Riestenberg, Kate, Francesca Di Silvio, Anne Donovan, and Margaret E. Malone. In press. "Development of a Computer-based Workshop to Foster Language Assessment Literacy." *Journal of the National Council of Less Commonly Taught Languages* 8.

Sandrock, P. (2010) The Keys to Assessing Language Performance: Teacher's Manual. Alexandria, VA; ACTFL.

The Shepard, Lorrie A. 2000. "The Role of Assessment in a Learning Culture." *Educational Researcher* 29, no. 7: 4-14.

Stiggins, Richard. 2001. *Student-Involved Classroom Assessment.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall, Inc.

Stoynoff, Stephen, and Carol A. Chapelle. 2005. *ESOL Tests and Testing: A Resource for Teachers and Program Administrators.* Alexandria, VA: TESOL.

Taylor, Lynda. 2009. "Developing Assessment Literacy." *Annual Review of AppliedLinguistics* 29: 21-36.

Wiggins, Grant, and Jay McTighe. 2005. *Understanding by Design.* 2nd ed. Alexandria, VA: Association for Supervision and Curriculum Development.