

## RESPONSE TO THE **NORRIS COMMENTARY**

**Dorry M. Kenyon, Valerie Malabonga, and Helen Carpenter**  
Center for Applied Linguistics

Norris has provided a thorough and insightful commentary to our article. Nevertheless, we would like to respond to four points that Norris has brought up. The first two points are clarifications about the COPI's design and our most recent research findings about this new test. The last two points respond to Norris' discussions of broader issues in the assessment of speaking, specifically the validity of the ACTFL *Guidelines* (1999) and the complexity of assessing second language speaking.

First, a clarification about the adaptive algorithm of the COPI seems appropriate. In a lengthy performance-based assessment, efficiency is necessary. In the case of the COPI, raters need to listen to all responses made by examinees. With an essay prompt, in large-scale written tests, raters often evaluate in 1 to 2 minutes what examinees took 30 to 45 minutes to produce. With speaking assessments, however, if the examinee speaks for 20 minutes, the rater needs to listen for at least 20 minutes, assuming the rater listens only once.

While a typical full-length Simulated Oral Proficiency Instrument (SOPI) presents 15 tasks to the examinee, the first 7 tasks may be administered as a "short form" for examinees at ACTFL Intermediate and Advanced levels. These 7 tasks comprise 4 tasks at the *intermediate* level and 3 at the *advanced* level. Our experience indicates that this approach produces an adequate speech sample to make for raters to arrive at their ratings (e.g., Kenyon & Tschirner, 2000). Based on this experience and to improve efficiency, we decided a priori that the COPI algorithm should also present examinees with 7 tasks, 4 at their starting level (i.e., the level of the self assessment) and 3 at the next level above. If the starting level was *superior*, then four tasks at the *superior* level and three at the *advanced* level were presented. The only instances in which more than seven tasks were administered to examinees occurred when, during the course of the COPI, they chose to be administered a task below their starting level or two levels above their starting level. The algorithm used in the COPI continued testing until at least four tasks at the starting level and three at the next higher level were administered. In a small number of cases (11 out of 54, or 20%), more than seven tasks were administered before this criterion was reached.

One of the main goals in this approach was to ensure that examinees were not disadvantaged with a lower rating if they started at too low a level. Subsequent analysis of the examinees' COPI and SOPI scores, as reported in Kenyon, Malabonga, and Carpenter (2001), revealed that examinees' starting COPI level (based on their self-assessment) was problematic only for a small percentage (8 %) of the examinees.

Second, we agree with Norris that there is still work to be done in researching the COPI. The current article reports on our first analyses of the COPI, focusing on the attitudinal questionnaires. Kenyon, Malabonga, and Carpenter (2001), for example, present further results on the relationship between examinee self-assessment of speaking proficiency, teacher assessment of examinee speaking proficiency, and COPI results. These analyses show that the student self-assessment instrument had a high rank-order correlation with examinees' actual performance on the COPI (.88), and that teacher assessment of student proficiency also correlated highly with the COPI (.84). Kenyon et al. also presented a preliminary analysis on examinee use of planning and response time, showing that more proficient examinees used less planning time but had longer response times. We intend to conduct further analyses of the COPI, including examining raters' behavior, investigating equivalencies of ratings across the three test formats and among tasks from the same ACTFL level, and conducting discourse analyses of examinee speech under the different testing conditions.

Third, we feel it necessary to provide a broader perspective on Norris' comments regarding the ACTFL *Guidelines* (1999). He writes, "Nearly two decades of criticism and research have cast serious doubts on the usefulness of this metric for informing interpretations about learners' language abilities or for making decisions and taking actions within language classrooms and programs..." While we agree with the issues Norris raises from a research perspective, from a policy and practice perspective the *Guidelines* have been useful to many in the foreign language field. In fact, after 20 years, their reach seems to be extending, rather than receding.

The ACTFL *Guidelines* (1999), emerging from the U.S. government's Interagency Language Roundtable Descriptors originally developed in the 1950s, responded to a need for a "common yardstick" to interpret growth in functional language competency across foreign languages, the primary goal for which they were developed (see, for example, Lowe, 1988). Since the ACTFL *Guidelines* appeared in 1986, they have been institutionalized in many foreign language programs throughout the U.S. and used in teacher training textbooks (e.g., Omaggio-Hadley & Terry, 2000). They have also served as a basis in the national *Standards for Foreign Language Learning* (National Standards in Foreign Language Education Project, 1996) and for many state-level curriculum frameworks, and have informed the first-ever foreign language National Assessment of Education Progress (NAEP), scheduled to be administered in 2003 (Kenyon, Farr, Mitchell, & Armengol, 2000). As a "common yardstick," the *Guidelines* of necessity may be broad, imprecise and approximate in their ability to describe "functional competency." As Norris states, the appropriateness of inferences and decisions made on the basis of results of assessment instruments based on them should be subject to continuing research. Nevertheless, users in academia, government, and, increasingly, in business, have found the levels and criteria of the ILR *Descriptors* and the ACTFL *Guidelines* useful for their purposes. No doubt, they will continue to find them useful in the absence of any other common metric.

The validation of the way in which users use assessment results based on the ACTFL scale will be a mammoth and lengthy undertaking. While some language testing researchers may desire a moratorium on the development of assessment instruments based on ACTFL criteria, this seems unlikely. In fact, research aimed at understanding the nature of such assessments and exploration of alternatives to them continues to be a worthwhile undertaking for those who serve a broad constituency in foreign language education.

Finally, we agree with Norris that the assessment of speaking in a second language is complex, whether with or without a computer or other technological assistance, and that evidence-centered design is critical. Indeed, in an early application of that approach, Mislevy used the ACTFL *Guidelines* (1999) for illustrative purposes (Mislevy, 1995), again attesting to the usefulness of such criteria, however coarse they may be. For those whose decisions are based on evaluating speech performances using the ACTFL criteria, despite their imperfections, our investigation begins to build a much-needed knowledge base about the use of computer capabilities in speaking assessments.

## **ABOUT THE AUTHORS**

Dorry M. Kenyon is director of the Language Testing Division at the Center for Applied Linguistics (CAL) in Washington, DC. Valerie Malabonga is a Research Associate at the Language Testing Division at CAL. Helen Carpenter is xxxxxx

E-mail: [dorry@cal.org](mailto:dorry@cal.org), [valerie@cal.org](mailto:valerie@cal.org), [helen@cal.org](mailto:helen@cal.org)

**REFERENCES**

- American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines--speaking: Revised 1999*. Hastings-on-Hudson, NY: Author.
- Kenyon, D. M., Farr, B., Mitchell, J., & Armengol, R. (2000). *Framework for the 2003 Foreign Language National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board. (Pre-publication edition) [A complete copy may be downloaded from the publications page of the National Assessment Governing Board's Web site at <http://www.nagb.org/pubs/pubs.html>.]
- Kenyon, D. M., Malabonga, V., & Carpenter, H. (2001, February 20-24). *Effects of examinee control on examinee attitudes and performance on a computerized oral proficiency test*. Paper presented at the 23<sup>rd</sup> Annual Language Testing Research Colloquium, St. Louis, MO.
- Kenyon, D. M., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *Modern Language Journal*, 84(1), 85-101.
- Lowe, P., Jr. (1988). The unassimilated history. In P. Lowe, Jr., & C. W. Stansfield (Eds.), *Second language proficiency assessment: Current issues* (pp. 11-51). Englewood Cliffs, NJ: Prentice Hall Regents.
- Mislevy, R. (1995). Test theory and language-learning assessment. *Language Testing*, 12(3), 341-369.
- National Standards in Foreign Language Education Project. (1996). *Standards for foreign language learning: Preparing for the 21<sup>st</sup> century*. Lawrence, KS: Allen Press.
- Omaggio-Hadley, A., & Terry, R. (2000). *Teaching language in context (3<sup>rd</sup> edition)*. Boston, MA: Heinle & Heinle Publishers.