Language Learning & Technology May 2001, Vol. 5, Num. 2
http://llt.msu.edu/vol5num2/norris/default.html pp. 99-105

CONCERNS WITH COMPUTERIZED ADAPTIVE ORAL PROFICIENCY ASSESSMENT

A Commentary on "Comparing Examinee Attitudes Toward Computer-Assisted and Other Oral Proficient Assessments" by Dorry Kenyon and Valerie Malabonga

John M. Norris
University of Hawai`i at Manoa

There is no doubt that computers and related technology have already acquired considerable importance in the development, administration, scoring, and evaluation of language tests, as this special issue of *Language Learning & Technology* demonstrates (see also Alderson, 2000; Brown, 1997; Chalhoub-Deville & Deville, 1999; Chapelle, 2001; Dunkel, 1999). Given the integral role computers play in our lives, and advances in technology which will make possible the measurement of an expanding array of constructs, it is clear that the use of computer-based tests (CBTs) for language assessment and other educational/occupational assessment purposes will become increasingly predominant in the immediate future (Bennett, 1999). However, what is unclear is the extent to which CBTs will offer the most appropriate means for (a) informing the interpretations that language educators want to make about the language skills, knowledge, or proficiencies of L2 learners and users; and (b) fulfilling the intended purposes and achieving the desired consequences of language test use (Norris, 2000).

Of particular concern for language testing is the extent to which CBTs may contribute to assessment of productive language performances, especially those involving speaking abilities (Alderson, 2000; Bernstein, 1997; Chalhoub-Deville & Deville, 1999). Of course, computerized tests of speaking have been developed which elicit production on constrained tasks and automatically score isolated features such as fluency and pronunciation (e.g., Ordinate Corporation, 1998), and seminal work is underway in developing an integrated speaking component for the CBT Test of English as a Foreign Language (Butler, Eignor, Jones, McNamara, & Suomi, 2000). However, despite such efforts, it is questionable whether the full range of individual and interactive speaking performances that language educators are interested in will be adequately elicited in computerized formats; likewise, it is doubtful that the complexities of such performances and the inferences that we make about them will be captured by automated scoring and speech recognition technology (Burstein, Kaplan, Rohen-Wolf, Zuckerman, & Lu, 1999). Furthermore, because it is unlikely that complex speaking performances will be automatically scoreable, the applicability of computerized *adaptive* testing (CAT) for assessing speaking, among other complex abilities, remains unclear (see related discussions in Wainer, 2000).

Recent research and development efforts at the Center for Applied Linguistics (CAL) demonstrate one approach to combining available technology with advances in measurement theory (i.e., adaptive testing) in order to move beyond the testing of receptive language skills (e.g., Chalhoub-Deville, 1999) and towards a creative solution to the computerization of direct tests of complex speaking abilities. As such, the Computerized Oral Proficiency Instrument (COPI) presents the language testing community with a good opportunity to further consider just how CBT capabilities may best be matched with intended uses for language tests. In this brief commentary, I will address (a) what the COPI has to offer to language testing, (b) some of the key issues that should be addressed in future research on the COPI, and (c) several fundamental concerns associated with attempts to computerize L2 speaking assessment.

WHAT DOES THE COPI HAVE TO OFFER?

The COPI features several technical and procedural innovations which may offer improvements over other types of technology-mediated oral proficiency assessment (e.g., the tape-based Simulated Oral

Proficiency Interview, or SOPI), especially in terms of examinees' affective responses and efficiency in administration and scoring. As reported in Kenyon and Malabonga's article, the COPI utilizes technology to address affective concerns by introducing examinees to the computerized test format with a hands-on tutorial, by increasing examinee control over topic selection and planning/response time, and by introducing an adaptive algorithm in order to present examinees with speaking tasks that are not overly easy or difficult. Findings from CAL's initial investigations seem to indicate that these innovations achieve the desired effects. Thus, even though examinees generally evaluated the COPI and SOPI formats in equivalently positive ways on independent surveys, when they were asked to select between one of the two tests, they reported on average that they preferred the COPI because it (a) seemed less difficult overall, (b) featured a fairer set of questions and situations, (c) made them feel less nervous, (d) had clearer directions, and (e) enabled a more accurate depiction of their strengths, weaknesses, and current abilities to speak in the target language.

Unfortunately, because the questionnaires utilized in the study did not allow for examinees to respond that "neither" test format was appropriate (see Tables 3, 6, 9), it cannot be ruled out that a proportion of the response patterns in favor of the COPI may be attributable to a shortcoming in the research design (i.e., it is unclear how many of the examinees would have chosen "neither," had it been an option). It also seems evident (Table 9) that much of the observed pattern of preference for the COPI over the SOPI was a result of substantial differences expressed by lower proficiency examinees, while middle and upper proficiency examinees were more evenly divided in their preferences. Nevertheless, even if improvements in examinee affect are only found among lower proficiency examinees, the COPI's innovations are probably warranted. Furthermore, it should be pointed out that, regardless of infelicities in the research design, evidence does show that examinees felt at least equally comfortable, and in all likelihood more so, with the computerized format of the test. This is a particularly important finding in light of the oft-raised concern that features of the computerized context may negatively influence examinees' perceptions of a language test and, as a result, alter their performances (Chapelle, 2001; Douglas, 1998; Dunkel, 1999).

In addition to potential improvements over the SOPI in terms of examinee affective variables, the COPI offers distinct advantages in eliciting examinee performances and in facilitating the rating process. Administering the COPI will certainly be easier than the SOPI, or the live oral proficiency interview (OPI) for that matter, since the computer program does away with the need for a test proctor or interviewer to distribute and collect test materials, to monitor and advance test activities, to capture examinee performances (in the form of audio tape recordings), and so forth. One of the most useful technological advances featured in the COPI is the replacement of tape recordings with computer-based digital audio recordings. Computerized recordings will prove beneficial for raters, who will be able to listen to examinee performances on test tasks in any order, in part or whole, with instantaneous repetition of particular segments of speech, etc., all without the cumbersome demands of rewinding or fast-forwarding accompanying tape recordings used in SOPIs and OPIs. In addition, either CD ROM-based or Web-based formats of the COPI can be easily integrated with internet technology, such that recorded examinee performances are automatically distributed to certified raters as data files, making the scoring process that much more efficient. Finally, the COPI program also eliminates one common source of scorer error found in SOPIs by automatically assigning final global ratings based on a scoring algorithm rather than leaving this sometimes complex task up to the raters themselves. While the details for such automated delivery, administration, recording, and scoring processes will need to be carefully thought through and pilot-tested, especially in light of system design and hardware/software requirements (see Green, 2000), it is apparent that the COPI offers several important benefits over existing forms of oral proficiency assessment.

WHAT ARE THE KEY ISSUES TO BE ADDRESSED IN FUTURE RESEARCH ON THE COPI?

Although its innovations may offer certain advantages, the extent to which the COPI will serve as an adequate platform for assessing oral proficiency according to the ACTFL *Guidelines* (1999) as well as a model for the development of other CBTs of complex performance abilities, depends in large part on whether research reveals that changes in the testing format influence both the quality of speaking performances and the accuracy of proficiency ratings. The most important innovation in the COPI involves the use of an adaptive algorithm for the selection of test items. The COPI achieves adaptivity by first asking examinees to assess their own speaking abilities. Then, on the basis of this self-assessment, the computer suggests a starting difficulty level for tasks. The algorithm continues by alternating between (a) allowing examinees to select subsequent tasks that are easier or more difficult than the task just completed, and (b) automatically presenting somewhat more difficult tasks in order to probe the upper limits of an examinee's abilities.

At issue for research on the COPI is whether this adaptive approach will result in the selection of tasks which adequately reflect an examinee's actual abilities, and whether the number and range of tasks will provide sufficient evidence for accurate rater judgments. First, the influence of examinees' initial self-assessments on the difficulties of tasks selected by the algorithm will need to be investigated. For example, what is the impact on initial and subsequent tasks when advanced learners under-estimate or novice learners over-estimate their proficiencies on the self-assessment, a phenomenon observed in previous research (e.g., Heilenman, 1990)? Second, although allowing examinees to select the difficulty level of every second item may improve their affective responses, it will be essential to investigate whether examinees' selection strategies may systematically bias performance outcomes. For example, what is the impact of an examinee selecting items that are consistently easier than (or equivalent to) the previous item or consistently more difficult? Also, to what extent would the algorithm present the same examinee with sets of items that reflect the same constellation of difficulty levels on independent trials of the COPI? Third, if the use of this algorithm reduces the number of tasks performed (typically to seven, as reported in Kenyon and Malabonga's article, although it is unclear how the algorithm makes the decision to *stop* presenting items), it will be necessary to investigate the extent to which raters find this truncated data set sufficient for accurately assigning a global proficiency rating. For example, will a rater be able to make an accurate judgment on the basis of two or three Advanced-level tasks and three or four Intermediate-level tasks? The real power of a computerized adaptive test comes from the fact that it can accurately adjust the difficulty of items presented to examinees based on the automatic scoring of responses to previous test items, and that it can continue presenting items until a desired level of precision is achieved (Thissen & Mislevy, 2000). The COPI obviously does not have this capability; as such, the use of self-assessment as an adaptive surrogate will need to be carefully researched.

In addition to adaptivity, other innovations in the COPI may affect the language performances elicited. First, examinees are offered a range of content areas and topics to choose from on each COPI item. While this innovation may raise examinees' interest in the exam, it also implies that a very large pool of tasks will need to be developed in order to adequately represent "interesting" topics at all of the ACTFL levels, such that the adaptive algorithm may select a number of unique tasks close to the examinee's proficiency (see related discussion in Flaugher, 2000; Wainer & Mislevy, 2000). In order to meet this requirement, tasks will have to be pilot-tested, and performance outcomes equated and calibrated according to the ACTFL *Proficiency Guidelines* (1999). Although a handful of tasks in any one language have been developed by CAL for the purposes of the SOPIs, it is questionable whether sufficient numbers of tasks of differing difficulties and featuring different topics are currently available. In addition, all potential COPI tasks will need to be individually investigated in order to determine whether ostensibly "parallel" tasks actually serve to elicit equivalent examinee performances in terms of language qualities as well as the perceptions of test takers and raters.

Another innovation in the COPI gives examinees substantial control over how much time they use in planning and responding to tasks, in contrast to the SOPI, where planning and response times are fixed. Whether examinee control over these features will translate into differences in performance outcomes, in terms of the amount and quality of speech elicited as well as the ratings assigned, is an empirical question which will need to be addressed by CAL, especially in light of research findings which suggest that such features may substantially affect performance outcomes (e.g., Ortega, 1999).

A final issue has to do with the extent to which examinee ratings on the COPI will be reliable as well as comparable to ratings on either the SOPI or the OPI. If each of these three test formats is intended to provide a trustworthy indication of an examinee's oral proficiency according to the ACTFL *Guidelines* (1999), then final global ratings should not differ depending on the test. In the preceding article, high rank-order correlations are noted between COPI, SOPI, and OPI ratings for the same examinees. As CAL researchers well know, rank-order correlation coefficients may easily mask actual rating differences, since they only compare the relative ordering of examinees (see discussion in Kenyon & Tschirner, 2000). The only means for addressing the equivalency of ratings on the three different test formats will be for researchers to calculate exact agreement coefficients for individual examinee ratings on each of the tests. Likewise, CAL should investigate the extent to which individual examinees are assigned the same ratings on subsequent trials of the COPI, especially in light of the fact that an examinee may face two different set of tasks from one testing occasion to the next. Finally, given the low levels of exact agreement that have been noted in previous studies of SOPI and OPI raters (e.g., Kenyon & Tschirner, 2000; Stansfield & Kenyon, 1992; Thompson, 1995, 1996), it will be essential to investigate agreement levels between pairs of COPI raters, as well as individual raters' confidence in judging COPI performances, especially in light of the reduced number of fixed-level tasks used to elicit speech samples on the COPI.

WHAT OTHER CONCERNS ARE THERE WITH COMPUTERIZING ORAL PROFICIENCY ASSESSMENT?

Many language testers are currently interested in exploring how CBT and CAT technology can be applied to the problems of language assessment. It is interesting that, by contrast, the authors of the most seminal work to date on computerized assessment have redirected their own concerns away from the "how" of CBTs/CATs to a more fundamental questioning of whether computerization is worth the effort and expense (Wainer, 2000). Wainer emphasizes, "The questions we now must address deal less with 'how to use it?' but more often 'under what circumstances and for what purposes should we use it?'" (p. xxi).

The development of the COPI offers a good example of how available technology may be applied in transforming a pre-existing test of oral proficiency to a CBT format, and innovations in the COPI offer insights into how certain speaking abilities may be effectively and efficiently elicited and scored with the help of computerization. However, the question that remains to be asked of the COPI is, "is it worth it"? In one sense, of course, CAL's research and development efforts are certainly worth it, if only to help language testers extend their understanding of the complex issues involved in computerizing language performance assessment. At the same time, it is necessary to take a critical look at the circumstances and purposes associated with the intended uses for the test itself.

The obvious purpose for the COPI is to provide another means whereby the speaking proficiencies of foreign language learners may be assessed according to the ACTFL (1999) *Guidelines.* According to Breiner-Sanders, Lowe, Miles, and Swender (1999), the *Guidelines* provide "a metric against which to measure learners' functional competency; that is, their ability to accomplish linguistic tasks representing a variety of levels" (p. 13). However, 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 (see recent overviews in Norris, 1997; Salaberry, 2000). What is long overdue, and absolutely essential before we can determine

whether tests like the COPI are worth the effort and expense, is for ACTFL and associated test developers to take seriously a comprehensive program of validation such as that outlined by the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), investigating (a) whether interpretations that are made on the basis of the *Guidelines* are warranted, (b) the extent to which such interpretations lead to sound decision making and related educational actions, and (c) whether using the *Guidelines* results in desired consequences for the language learners they are ostensibly designed to serve.

A more general concern with the computerization of all types of L2 speaking assessment is whether the critical features of speaking performance may be captured in computer simulations such that educators will be able to make warranted interpretations about learners' knowledge and abilities. Of course, the full range of capabilities offered by computer technology will need to be explored, including those featured in the COPI as well as others, such as the potential for multi-media to provide extended context, the recording of additional performance attributes like time on task and the steps taken in performing tasks, etc. However, language test developers need to begin their deliberations about speaking assessment not by asking what computers are capable of doing, but rather by asking (a) what kinds of interpretations actually need to be made about L2 speaking abilities; (b) what kinds of evidence a test will need to provide in order to adequately inform those interpretations; and (c) what kinds of simulation tasks will provide the required evidence (see Mislevy, Steinberg, & Almond, 1999). Adhering to such principles of evidence-centered design should enable language test developers to better judge the potential applicability and/or added value of a CBT/CAT, especially when interpretations need to be made about interactive (e.g., Young & He, 1998), integrated (e.g., Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000), or otherwise multifaceted L2 speaking abilities.

ABOUT THE AUTHOR

John Norris is a student in the Ph.D. program in SLA at the University of Hawai`i. He has worked as an ES/FL teacher in Brazil and Hawaii, and he has lectured on language assessment, curriculum development/evaluation, and research methods in Brazil, Japan, Spain, and the US. His research has been reported in journals such as *Language Learning and Language Testing*, as well as in several co-authored books with the University of Hawai`i Press.

E-mail: jnorris@hawaii.edu

REFERENCES

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Alderson, J. C. (2000). Technology in testing: The present and the future. *System, 28,* 593-603.

American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines -- speaking: Revised 1999.* Hastings-on-Hudson, NY: Author.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational measurement: Issues and practice, 18*(3), 5-12.

Bernstein, J. (1997). Speech recognition in language testing. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment -- Proceedings of LTRC 96* (pp. 534-537). Jyväskylä, Finland: University of Jyväskylä.

Breiner-Sanders, K. E., Lowe, P., Miles, J., & Swender, E. (1999). ACTFL proficiency guidelines -- speaking: Revised 1999. *Foreign Language Annals, 33*(1), 13-17.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44-59. Retrieved January 15, 2001 from the World Wide Web: http://llt.msu.edu/vol1num1/brown/default.html.

Burstein, J. C., Kaplan, R. M., Rohen-Wolf, S., Zuckerman, D. L., & Lu, C. (1999). *A review of computer-based speech technology for TOEFL 2000* (TOEFL Monograph Series Report No. 13). Princeton, NJ: Educational Testing Service.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper.* (TOEFL Monograph Series Report No. 20). Princeton, NJ: Educational Testing Service.

Chalhoub-Deville, M. (Ed.). (1999). *Issues in computer adaptive testing of reading proficiency.* New York: Cambridge University Press.

Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics, 19,* 273-299.

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research.* Cambridge, UK: Cambridge University Press.

Douglas, D. (1998). Testing methods in context-based second language research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 141-155). Cambridge, UK: Cambridge University Press.

Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology*, *2*(2), 77-93. Retrieved January 15, 2001 from the World Wide Web: http://llt.msu.edu/vol2num2/dunkel/default.html.

Green, B. F. (2000). System design and operation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 23-35). Mahwah, NJ: Lawrence Erlbaum.

Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 37-59). Mahwah, NJ: Lawrence Erlbaum.

Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7,* 172-198.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper.* (TOEFL Monograph Series Report No. 16). Princeton, NJ: Educational Testing Service.

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.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *On the roles of task model variables in assessment design* (CSE Technical Report 500). Los Angeles, CA: Center for the Study of Evaluation, Graduate School of Education & Information Studies at the University of California, Los Angeles.

Norris, J. M. (1997). The German Speaking Test: Utility and caveats. *Die Unterrichtspraxis/Teaching German, 30*(2), 148-158.

Norris, J. M. (2000). Purposeful language assessment. *English Teaching Forum, 38*(1), 18-23.

Ordinate Corporation (1998). *PhonePass test validation report.* Menlo Park, CA: Author.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21,* 109-148.

Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing, 17*(3), 289-310.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the Oral Proficiency Interview and the Simulated Oral Proficiency Interview. *System, 20,* 347-364.

Thissen, D., & Mislevy, R. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 101-133). Mahwah, NJ: Lawrence Erlbaum.

Thompson, I. (1995). A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: Data from ESL, French, German, Russian, and Spanish. *Foreign Language Annals, 28,* 407-422.

Thompson, I. (1996). Assessing foreign language skills: Data from Russian. *Modern Language Journal, 80,* 47-65.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd edition). Mahwah, NJ: Lawrence Erlbaum.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 61-100). Mahwah, NJ: Lawrence Erlbaum Associates.

Young, R., & He, A. W. (1998). *Talking and testing: Discourses approaches to the assessment of oral proficiency.* Philadelphia: John Benjamins.