An Interview with Dorry M. Kenyon[1]

Nathan T. Carr
University of California, Los Angeles

Viphavee Vongpumivitch
University of California, Los Angeles

PROFILE

Dorry M. Kenyon is the Director of the Language Testing Division at the Center for Applied Linguistics in Washington, D.C. and serves as the Book Review Editor for the international journal Language Testing. He has done extensive work in test validation studies, particularly in the validation of oral proficiency rating scales (e.g., Kenyon, 1998; Kenyon & Tschirner, 2000; Stansfield & Kenyon, 1993, 1996). He is also well known for his work in oral proficiency testing, such as the Basic English Skills Test (BEST) (Center for Applied Linguistics, n.d.b), Simulated Oral Proficiency Interview (SOPI) (Center for Applied Linguistics, n.d.d), and Computer-Based Oral Proficiency Instrument (COPI) (Center for Applied Linguistics, n.d.c).

INTRODUCTION

This interview with Dr. Kenyon addresses a range of issues based on his experience developing and validating a number of widely used language tests. In the first section, we ask Dr. Kenyon about his education and professional experience, what drew him to the field of applied linguistics, and how he became involved in language testing. In the next section, Dr. Kenyon discusses the Center for Applied Linguistics (CAL) and its past and current research and test development projects. CAL is a private non-profit organization headquartered in Washington, DC, which "aims to promote and improve the teaching and learning of languages, identify and solve problems related to language and culture, and serve as a resource for information about language and culture" (Center for Applied Linguistics, n.d.a). Dr. Kenyon explains the process through which the Center, and specifically the Language Testing Division, takes up research and development projects, and he details several of the projects on which he and his division are currently working.

The discussion in the third section turns to oral proficiency interviews and specifically to the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) (Language Testing International, n.d.).

Issues in Applied Linguistics
ISSN 1050-4273

Vol. 12 No. 1, 111-126

The OPI, which uses a single scale—the ACTFL Language Proficiency Guidelines (Breiner-Sanders, Lowe, Miles, & Swender, 2000)—to rate proficiency in a number of languages, is widely used in the foreign language education community in the United States. In some cases, however, it is either impossible or impractical to have a live rater available to conduct the interview, so institutions sometimes turn to tape- or computer-mediated versions of the OPI. Dr. Kenyon discusses these tests and how they differ from the traditional OPI. We also ask Dr. Kenyon about the ACTFL Guidelines themselves; he addresses such issues as the place of the scale in foreign language education, its origins, the reasons for its popularity, and concerns that have been raised regarding the scale.

In the fourth section, the interview takes up test validation. Validation studies are often underappreciated outside professional testing circles, but they are essential to determining whether a test measures what it purports to measure, and therefore whether its use for a given purpose is justifiable. Dr. Kenyon discusses issues involved in these studies, including real-world constraints on resources and the importance of convincing test users of the need for test validation. In the final section, we ask Dr. Kenyon to address issues he has encountered in the emerging use of computer- and web-based language testing. This is an area of widespread and growing interest in the language testing field, and one in which Dr. Kenyon has extensive experience.

THE INTERVIEW

Professional Background

Carr: *Please describe your academic training and what drew you into the field of applied linguistics and language testing.*

Kenyon: Well, as long as I can remember I've loved language and different languages. Listening to different languages on short-wave radio was a hobby when I was a kid, and trying to identify the different languages. I'd always loved math too, and my goal when I went to college was to be a math teacher, but because I also enjoyed languages so much, I wound up majoring in economics and German, and spent my junior year abroad in Germany.

I learned about a program at the American University in Cairo that offered a master's in teaching English as a foreign language (TEFL). Because I just enjoyed languages and wanted to experience a non-Western one, I decided to do my degree there. I chose TEFL because I was thinking about teaching math versus teaching languages, and math seemed to be more "by yourself," you know. With math, you're thinking things through and working by yourself, while teaching languages is more about working with other people. While working on my master's degree, I taught at the American University at Cairo, and I worked as a teaching assistant at a German high school where I'd help out the English teachers. I also taught four

summers at The American School in Switzerland (TASIS). In all those teaching situations, I was drawn to the testing, especially the placement testing. I always enjoyed that. I'd always be the first to volunteer to score essays and to give oral interviews and things like that.

Coming back to the States after finishing my master's, I taught for several years in the D.C. area, at George Mason University in adjunct positions. I realized that the only people who had stable positions were the directors of English language institutes; the others were all part-timers or paid by the hour and had no benefits or anything. So I said, "Well, I'd better get a Ph.D." The University of Maryland had good PR for its Department of Measurement and Applied Statistics and Evaluation. I had my math background, and they demonstrated that all the people got jobs when they graduated, so I decided to enter that program.

Just at the time that I'd decided to enter the University of Maryland, I met Charles Stansfield, who had come to the Center for Applied Linguistics (CAL), and he hired me to work on a language testing project there. I started my Ph.D. program at Maryland and my work at the Center for Applied Linguistics in the same week, in September of '87. I earned my Ph.D. in educational measurement, but always applied everything to language testing on projects at the Center for Applied Linguistics. It was kind of a natural. To me it was a very satisfying way of using a combination of my math and language backgrounds.

Research and Test Development at the Center for Applied Linguistics (CAL)

Vongpumivitch: *So you're now the Director of the Language Testing Division at CAL, the Center for Applied Linguistics. Would you mind explaining to us what CAL does, what its role is, and also in particular what your division does?*

Kenyon: CAL is a little bit hard to describe in the sense that we do so many things. We have six separate divisions now, and we're involved in all sorts of projects. Probably the most visible one to many would be the ERIC Clearinghouse on Languages and Linguistics, which we have the contract to run, so that's one of the 16 or 17 ERIC Clearinghouses, and we've had that contract for a long time. Maybe others would know NCLE, the National Clearinghouse on Adult ESL Literacy Education. It's now a center, actually, which produces a lot of publications. CAL itself has been around for over 40 years. It's a private nonprofit organization and has been involved in all sorts of issues at the intersection of language and society, language and education, and of course one of those issues in language learning is testing. What the Language Testing Division does is work on a whole constellation of projects. We bring in contracts and projects that relate to the assessment of English as a second language, or for example, Americans learning foreign languages.

Vongpumivitch: *So basically the Language Testing Division works on a project when someone contacts you and you sign a contract with them.*

Kenyon: Either as a contract, where they have something specified, or we go after research grants, and that gives us a little bit more leeway to propose what we'd like to do. A lot of our work in the simulated oral proficiency interviews—SOPIs—has come through those grants.

Carr: *When you pursue research grants, are you basically constrained by what someone is already willing to pay for to fund one of your projects, or are you able to decide "I want to look at this. I think CAL needs to explore this area," and then persuade funding bodies to cut a check for it?*

Kenyon: Both. Definitely both. For example, one government agency recently issued a request for the development of a test for a very specific purpose that we applied for. The purpose was to determine, in a machine-scorable way, whether candidates who were taking a translation test out of one language into English could write an English paragraph. But they wanted to have a pre-screen for people, so they wouldn't have to bother administering the hour-long writing test and scoring it. So that was very constrained and they knew exactly what they wanted and it had to be machine scorable. Within that boundary, then, CAL can say, "OK, but let's try it this way," or "Let's look at it this way." I don't know whether we are going to get that project, but I had to propose a different way of conducting it than the outline they sketched out, because the outline they sketched out wouldn't have worked and given them a defensible product at the end. So that's an example of where a request is very prescribed.

An example of the opposite, where you go to a funding source and say "Here's a need that needs to be addressed," was when the Center for Applied Linguistics developed what was called the Basic English Skills Test in the early 1980s for adult ESL. This test is kind of at the survival level, and it's being used more and more nationally for accountability purposes, but that's not what it was intended for at all. So we brought to the attention of the Office of Adult Vocational Education that "Hey, this needs to be updated and needs to be a little tighter for the purposes it's currently being used for." And over several years we were able to secure some funding for that.

Carr: *I gather CAL does a lot of work with government contracts. Is it a common problem with government projects that you create a test for one agency or one purpose and then somebody decides that they could save a little money, and says "Well, it's an English test or a Spanish test, so let's use it over here, too."*

Kenyon: Well, I don't think that anything CAL has developed has been used in that way, but that's often a big issue. In fact, we recently heard that the Office of

Bilingual Education is thinking about national standards for nursing tests for non-English speakers, and they wanted to find out more about the way performance-oriented assessment was developed in Australia in the Occupational English Test. Perhaps you're familiar with it, Tim McNamara (1996) used it to illustrate his book Measuring Second Language Performance. They wanted something like that, combined with the TOEIC (Chauncey Group International, n.d.), which is a multiple choice test. And you know, it's really like apples and oranges, and you have *to say so. Somebody will call you up and ask for information and your opinion about using a particular test, and you have to say "Well, what is it, what do you really want?". Ultimately, maybe that's a good thing. I guess you can try to convince them that they should develop something from scratch, or at least adapt something. But there is a lot of that going on. I don't think that has happened with anything that CAL has developed except for the BEST, but that's more because there just aren't many standardized adult ESL assessments out there. The BEST is just about the only oral test that's out there.*

Vongpumivitch: *What are the projects that you're working on at this moment?*

Kenyon: One of the main ones is what we're calling the CBEST, but we may have to change that name because there is a California test called the CBEST (California Commission on Teacher Credentialing & National Evaluation Systems, n.d.). We'll probably call it the Computerized BEST or BEST Plus, because with this new generation of the BEST that we're developing the administration will be assisted by a computer. For example, we're thinking about what questions we'll ask so that we can better get at the target language situation and the ability level of the examinee. That's a big project. Another project is item development for the Foreign Language National Assessment of Educational Progress, which will likely be administered in 2003. We have a new administration now, and their plan for the uses of data from NAEP—that's the National Assessment of Educational Progress (National Center for Education Statistics, n.d.)—is different; they're conceptualizing it a little bit differently.

Carr: *How is it different?*

Kenyon: It's given in a variety of subject areas, and finally after 30 years they've reached foreign language education. But the Bush Administration would like, from what I understand, to have the NAEP serve as a national benchmark against which the states can compare their state-level performances in the basic skills, so they would particularly like it to be given annually in reading and math. NAEP is currently mandated to be administered in grades 4, 8, and 12. If the funds go into doing reading and math on a yearly basis, they won't have funds left over to do this foreign language assessment. But we're moving ahead on that, and we are currently funded.

It's just a case of a limited pot of money, so if we're going to use the money that's there to do annual testing in fourth and eighth grade reading and math, you don't have the other funds left over. Last Saturday, President Bush did his weekly radio address in Spanish, so I don't think it's because the new administration doesn't like foreign language. It's just because conceptually, they see the purpose of NAEP differently. But it will take a while for any changes to take place, because although NAEP is funded by the government, there's a governing board that is independent, and so it has to pass through that. Anyway, we're working on the development of those items. In particular the Center for Applied Linguistics is responsible for developing the interpersonal task for the speaking assessment. And ETS (the Educational Testing Service) has the main grant to do that project.

Another project, which I think you both heard about at LTRC (the Language Testing Research Colloquium)[2], is the Web test project we're doing (Malone, Carpenter, Winke, Kenyon, 2001), which involves creating a framework for the development of a listening and reading comprehension test that can be scored on, and validly aligned with, the ACTFL Guidelines. Those guidelines have just become really entrenched in the foreign language field, and people like the idea of being able to say "Well, this person can read at the advanced level in Russian, and this person can read at the advanced level in Arabic," and think that we're comparing things that are similar. The importance of this project is that although the government has some ways of testing reading, outside of the government, there haven't been ways of testing listening and reading using the ACTFL scale that have proliferated. So we're developing a framework that could be applied to less commonly taught languages, and developing that in Arabic and Russian for delivery over the Web.

Then another major project that we're working on is something completely different. It's a large-scale project that is funded by the National Institutes of Health (NIH) ultimately, and OERI, the Office of Educational Research and Improvement, and one of the institutes in NIH is the NICHD, the National Institute of Child Health and Human Development. What they've done is they've pooled a lot of money to have a coordinated effort on research to find the best way that children entering our school system speaking Spanish can become literate in English. They funded two five-year programs, and CAL has one jointly with Harvard and Johns Hopkins University. The University of Houston has the other, and then there are other small, independent one-, two-, and three-year projects. They're all coordinated together, so at the end of five years we'll have some definitive research on these issues.

What the Language Testing Division is doing is assisting the project researchers from Harvard and also those from CAL in developing two things. One is developing the instrumentation, because in a lot of this research, at the end of the day sometimes you feel the outcomes were an artifact of poor instrumentation, and so they'd like to have a more principled way of developing instruments. They're not necessarily for language proficiency. Often they're very focused on cross-linguis-

tic issues like morphology, or sound symbol knowledge, or spelling ability as well as more global measures. Another big issue that we're helping this program on is the use of standardized assessments for this population. For example, there's a Woodcock-Johnson test for English speakers (Woodcock & Johnson, n.d.), and there's Woodcock-Muñoz for Spanish speakers (Woodcock & Muñoz, n.d.). But these are bilingual students, so you have to make accommodations using either test. For example, in the test, in order to assess the children's ability to know sound-symbol correspondence, they use pseudo-words. Well, in Spanish, some of those pseudo-Spanish words are actual English words. So if the student pronounces it as a sight word the way it should be said in English because they know the word in English, it would be counted wrong – because they're not showing their knowledge of the sound-symbol correspondence in Spanish. So we have to standardize how those accommodations are going to be made across all the different projects. Those are the main projects I'm currently working on.

Oral Proficiency Interviews and the ACTFL Guidelines

Vongpumivitch: *You also have the Simulated Oral Proficiency Interview (SOPI), which is a tape-mediated Oral Proficiency Interview. What's the difference between the COPI (Computer-Based Oral Proficiency Instrument) and SOPI?*

Kenyon: The SOPI is on tape, so when the tape starts playing, the students get the same exact tasks, the pauses are timed for them, and the response time is timed for them. In the COPI, which is administered over the computer, the tasks are very similar, but the students has more control. They can pick from several tasks. They can say "Oh, this task was too easy, give me one that's harder," or "This task was too hard, give me one that's easier," and they can control the thinking time and the response time. Also, because you can store so much more in the computer than on a tape, they have other choices about the language of instructions. Generally, these are foreign language tests, so the instructions are given in English and the responses are in the foreign language. But higher-level examinees would like to have the instructions in the language that they're studying, so they're not switching back and forth. We couldn't do that on the tape version because you can only put so much on a side of a tape.

Vongpumivitch: *What are the uses of those tests?*

Kenyon: They are mostly developed as research projects, and the COPI specifically was a research format. But they are made available, and we have what we call the self-instructional rater training kits, so people can buy it off the shelf, learn how to rate it, and administer and rate it themselves. They are often used in programs that want to assess the speaking ability of students. For example, a small college might use it because they want to evaluate their majors in French or Ger-

man with a more standardized oral assessment, and in the SOPI we try to relate the outcomes to the ACTFL performance levels so that they get some sense of where their students are in the ACTFL guidelines. So that's the main use. Also, high schools use it for students who have had several years of study, for evaluation purposes, but those are for internal purposes, because they're rating their own.

Carr: *You said the COPI is self-adaptive, more or less?*

Kenyon: Yeah, more or less. There's an underlying algorithm, so that students have to be assessed at tasks at their starting level and also more challenging tasks, whether they like it or not, because you don't want to disadvantage students who really can do more, but self-assess themselves at the beginning at a lower level. So that's really the danger there.

Carr: *Or a student who's too intimidated to try a more challenging level.*

Kenyon: Right, but who may really be able to handle it.

Vongpumivitch: *Does CAL also administer the OPI?*

Kenyon: No, we don't. We do work more and more with ACTFL who provide training for the OPI. Also particularly through their LTI, Language Testing International, they arrange for official OPIs to be given. Often they're given telephonically.

Vongpumivitch: *And the use of OPI is—as opposed to SOPI, which is a research project—I mean the OPI is an actual "test." Right?*

Kenyon: Through LTI, you can take the test, pay money and get a certificate, and then there's very vigorous quality control. There are only certain people who can do that. ACTFL will also train interested people in how to administer and score the OPI. Those people may go back to their universities and administer the OPI, but again, like the SOPI score, it wouldn't be certified outside of that particular university.

Vongpumivitch: *We'd like to turn to another tool associated with oral proficiency interviews. I think that for a period of time you did a lot of studies validating the ACTFL scale for oral proficiency interviews within many different situations (e.g., Kenyon, 1998; Kenyon & Tschirner, 2000; Stansfield & Kenyon, 1993, 1996). So what would you say are the main findings?*

Kenyon: This is a big issue, and unfortunately, there's a big divide between foreign language education on one side and ESL/applied linguistics on the other. So I

have to say that up front. I have to say up front too that in the foreign language field, the ACTFL Guidelines are quite entrenched. They're not really used much in ESL, EFL, and applied linguistics. One of the reasons is because clients want things interpreted on the ACTFL scale. It's just a given in foreign language education, a little bit like what's developing in Europe with the Council of Europe's (1996) six levels, with three main levels, each with two sublevels. CAL has sought to develop products that meet the needs that people have and so we've been working with the ACTFL scale with the SOPI for a long, long time. And, as we've had opportunity, we've tried to look at some of the issues related to the use of the scale. We can't get funding for a project that includes validating the whole ACTFL scale, but as we're doing a project we can look at validating different pieces of the scale.

The main published research that you might be thinking about is on students' perceptions of difficulty in the SOPI. In the OPI, the questions are targeted at these different main levels. And one issue is, are those tasks really more difficult? Finding the difficulty of tasks is a big problem, a big issue, and how do you really define the issue? There are so many things that make a task more or less difficult. So, some of that research (e.g., Kenyon 1998) was taking tasks like the ones used in the SOPI which are developed to assess at these different ACTFL levels, and seeing if students put them in the same kind of difficulty order as would be predicted by the scale. There are several studies of that through a self-assessment instrument that we've done. Of course, the scale and this all come from the government beginning in the 1950s.

Carr: *The ILR (Interagency Language Roundtable)?*

Kenyon: The ILR skill level descriptors. This is also catching on a whole lot in business too, because they're the ones that ACTFL mostly does language testing for. So it's just become a common standard out there. I think a big concern, both in the earlier days and still today, is inter-rater reliability. People who are using this test seem to be focused a lot on that kind of research. So as I say, we've been trying to look at it — as we've had opportunity — in different ways. My sense of it, when it all boils down, is that in broad brush strokes it has some use for painting things broadly, but it's not the end-all and be-all and I think that some people, proponents especially in the early days, were saying that it was. I think better describing what it is and what it isn't is probably the most honest thing you can do with it. Because it's there to stay, so people should understand the nature of the beast, so to speak.

Vongpumivitch: *I can understand why it's so popular, because when you need to make a quick decision, to make a quick judgement on someone's proficiency level, those scales come in so handy. Apart from those scales, is there any other kind of scale that is that clear and "comfortable?"*

Kenyon: I think you've said something that's really very important. It's that it's the practitioners in foreign language education who like it for that reason. Because I think it responds to something in their experience and in broad-brush strokes. Let's take intermediate-mid, because that's a very broad level. Students who are intermediate-mid level have many different profiles, and if you did diagnostic testing, individual strengths and weaknesses would be very different. But in a broad brushstroke, people find it useful just to call them intermediate-mid. However, I digressed a little bit. The acceptability of a scale is a bit of a social phenomenon, too. Researchers could say many things about that scale, or about something else. Like, you know, the TOEFL (Educational Testing Service, n.d.b) has the TOEFL 2000 Project, which is supposed to respond to some of its weaknesses in the past, and I applaud ETS for doing that. But the TOEFL is pretty entrenched. Researchers can critique it, but if the schools and colleges and universities use it and find that it's doing its job, or it's doing a job, or it's helpful, that, in a sense, can have more strength than what the researchers say. I really applaud ETS for trying to bring the two together. I know that they've been putting a lot of resources into that and I expect to see really good things in the future. But users are what make it entrenched, and in the United States there aren't any other scales that have been adopted. In adult ESL education there's something called student performance levels, or SPLs, in the MELT (Mainstream English Language Training) Project, which gave rise to the BEST, and they have a somewhat similar currency in adult ESL. Most of the adult ESL tests try to map onto some of the SPLs. I can't think of another scale, with the exception of the one in Australia[3] (Wylie & Ingram, 2001) which I understand as being derivative from ILR and ACTFL. Then there also are the Council of Europe proficiency levels.

Carr: *Some have criticized the ILR and ACTFL scales, though, because they contain both content and context as parts of the definition of proficiency. Specifically, they measure the ability to use vocabulary and grammatical structures in the context and under the conditions that are included in the testing procedure. You've worked extensively on projects using these scales, and they do have a lot of currency. What do you think would be the response from the ACTFL side?*

Kenyon: I don't know what to respond from the ACTFL side. I can say how I'm looking at it right now. I think proficiency is much, much narrower than communicative competence. It's kind of necessary—maybe not always necessary, I mean, people communicate in the real world without knowing each others' languages, but that can only go so far. I'd say it's necessary but not sufficient to accomplish real-world tasks. I think when all these things started in the government, it needed to develop a scale to know what the people who it was training for embassies could do in the real world. So they—and this was the 1950s, too— were trying to replicate the type of language in the interview context that might be found outside the interview context. Since that time, I think we've become sensitive to the many

other factors that are involved in being successful in a communicative situation, and I think what you were talking about was content?

Carr: *And the specifics, and problems with generalizability...*

Kenyon: I think in those early days they overgeneralized and some claims were made that were very, very broad. They were probably hard to support because having a face-to-face interview with one person, well, how will you do outside that context? On the other hand, if you don't have that proficiency, that's necessary but not sufficient, that's what you get out of an OPI, I think, in terms of what's been automatized, what can they speak freely about, what are the content areas, what's the breadth of their grammatical control and their vocabulary and the way that they can express themselves. That's going to be necessary out there, but it clearly isn't sufficient. It's not everything that's needed out there in the world. I would agree with you that the claims for generalizability outside of the context are probably exaggerated in that sense.

Vongpumivitch: *You mentioned raters, and that in using the scale-based test, raters are a very important group of people, and that a lot of research has been done on rater training. Who are the raters for these tests, such as the OPI?*

Kenyon: For the OPI, again, there's the official LTI OPI, and those are people who work very closely, and they're rating almost daily now. I mean, the volume of work has gone up exponentially. So there's tight quality control on those people, and they're meeting, they're benchmarked, there's all sorts of double-rating and triple-rating for an official ACTFL OPI through the LTI. The unofficial ones might be analogous to the SPEAK (ETS, n.d.a), you know, to what you do here at UCLA. Hopefully you have competent people there who have been trained who are making these ratings. Maybe there's a calibration test and everything like that.

Test Validation Studies

Vongpumivitch: *As someone who has a lot of experience doing validation studies, either for rating scales or for tests, what is the goal, in your opinion, of a validation study, either for a test or for a rating scale, and what is the best way to go about doing a good validation study?*

Kenyon: The rating scale in and of itself isn't validated apart from the whole assessment process. That's just one component there, but as you know, the standard issue of validity is what are the inferences being made, what are the actions being taken about the student on the basis of this test, and what theoretical considerations along with empirical evidence can we use to demonstrate that it's appropriate to make those inferences. Essentially, the question is: How do you justify

the use that's being made of these test results? So how do you do that? Well, first you have to have a very clear understanding of how the test results are being used. And I think that's one thing that may not be clear to everybody about the ACTFL guidelines, because they're used in so many different contexts. For example, with the LTI, now it's often used for correct placement in employment positions, so in that context you'd have to understand the decision that's going to be made about that candidate for that position.

Let's give the example of someone who's going to be employed with the AT&T Language Line, a service you can dial up for translation in one of 140 different languages. Anybody can do that on line. So if you're a border policeman and you stop somebody and you can't understand what language it is, you call up this Language Line. I don't know whether AT&T still runs it, but the Language Line service is still out there. You call them up, get the language identified, and they get somebody who can translate on the phone while you're trying to question this person that you stopped at the border. So who fills those positions? Well, people who fill those have to know English, and they have to know the language that's being spoken. How well do they have to know it? Well, let's say they say that ACTFL "superior" is what's needed. So the company that runs that business has to hire some people in Thai and say that these are "superior" level people. So they give the OPI in Thai and find out whether these people are "superior" or not.

Well, the question is: In validating that, is being "superior" in Thai, and being "superior" in, let's assume English, sufficient to do that job or are there other skills necessary? I think there might be other skills involved in interpreting that might be necessary to train on, but if you don't have those language skills, it may not matter much. So if an organization is hiring people and putting them in this job just on the basis of an OPI, that might be insufficient because there might be other skills involved. So that might be a validity issue. If the OPI were to say "All you need to do is have an OPI and you can do this job." Well no, there's probably more. Again, going back to that sense of being necessary but not everything. Validation is so contextual; it's hard to say there's one way to validate. If the organization were to say, "Well, we can't provide training, we can only make decisions based on the OPI," and the people from the OPI are saying, "That's fine, sure, this is sufficient," well that might be problematic.

Vongpumivitch: *As graduate students, we can take classes about validation studies, and read about validation studies, but I think we really need to hear the views of people who have actually done validation studies.*

Kenyon: Well, the issue – and this is serious – is that in academia when you're not working with it, you think there are unlimited resources. Resources are very, very limited, and validation is the last step. It is easy to use all your money up before you get to that. I think one thing that language testers have to remind projects from the very beginning is that, at the end of the day, you're going to have to provide

demonstration for the uses that are being made of the assessment results. We have to set aside funding for that. I think that really is important, but often that can be overlooked in the real world. One of my pet peeves with working with people in the real world is they often start with what the task looks like. They don't think through the issues. And that's why my workshop at this conference is very different from what was in the abstract. I changed it. It's from thinking through those issues and then thinking about what the task is going to look like. But people say, "Oh, this TOEIC test looks like a good one. Now we need one for Spanish. Let's just translate it into Spanish, you know. But we need it for a whole different purpose. We want it for teachers." when it was written for business contexts. They'll look at something, at the superficial form, and say "That's what I want, I just want you to revise it. It's not going to take very much to revise it."

Computer-Based and Web-Based Language Testing

Carr: *You already talked to us about the ways that the COPI differs from the SOPI. In general, what would you see as some of the advantages and disadvantages of computer-based testing, web-based testing? What are the or doors that it opens and limitations that it imposes?*

Kenyon: That's a good issue for us in our situation. Again, given that we have limited resources, there seems to be tons of potential out there for what could be done. But when you get into real-world projects, often people will fund the tried and true. So basic research in what could be done and really exploited to make computer-based tests more than just paper-based tests on computer is really necessary and valuable. But my general sense of it is that large-scale programming in language testing, at least, computer-based testing , especially computer adaptive testing, has not been giving the return on investment, so to speak, that was originally foreseen and desired and hoped for. I'm aware of some computer adaptive tests that have run into issues and problems.

Carr: *What kind of issues have you seen occurring?*

Kenyon: The big issue is with the development of the number of items. There's usually an incredibly large item bank that's necessary to support a computer adaptive test. So that's one big issue, getting that number of high-quality items, having them all calibrated somehow, having them all field tested, revised, and calibrated before going into the pool. It's an expensive and big task for smaller assessment programs, and even for big ones.

CONCLUSION

Dr. Kenyon raises a number of important issues regarding language testing. Perhaps most significant are his comments relating to the use of the ACTFL Guidelines and OPI in the foreign language education community and some of the suggestions he provides on how to promote test validity and validation in real-world projects.

Regarding the ACTFL Guidelines and the OPI and its "cousins," he ar*gues that in spite of issues involving generalizability of performance beyond the context of the interview, these tests are widely popular in the foreign language community and unlikely to be displaced in the foreseeable future. This may be in part a symptom of a wider problem: In describing the entrenched position occupied by the ACTFL Guidelines, Dr. Kenyon mentions a disconnect in the United States "between foreign language education on one side and ESL/applied linguistics on the other." This is a regrettable and somewhat disturbing trend, as such a gap is potentially harmful to both communities, posing the risk of cutting off foreign language education professionals from research in applied linguistics, and of limiting the opportunities open to applied linguists in general and language testers in particular for doing research in languages other than English.*

Finally, Dr. Kenyon notes that in the real world resources are limited, making it important to point out from the beginning to test users that funding must be set aside to pay for validation in order to support the uses that will be made of test results. In addition to recommending the development of a new test or adaptation of an appropriate existing test, when asked for advice on the potential adoption of an inappropriate assessment procedure, Dr. Kenyon proposes another way in which language testing professionals can help encourage more valid test use. The other way, which might be termed a "half a loaf" approach, is somewhat more subtle, but well worth noting: When working on a project involving a portion of a larger testing program, opportunities should be found to do limited validation studies of some specific aspect of the test. While it is obviously preferable for test users to have invested in comprehensively investigating the validity of a test's uses, when they have failed to do so, language testers may be able to use this approach to at least partially correct matters.

NOTES

[1] The interview took place at the Fourth Annual Conference of the Southern California Association for Language Assessment Research held in Pasadena, CA (May, 2001). Nathan Carr and Viphavee Vongpumivitch are Ph.D. students in applied linguistics at the University of California, Los Angeles, and are specializing in language assessment.

[2] The 23rd Annual Language Testing Research Colloquium (LTRC) was held in St. Louis, MO in February, 2001

[3] The International Second Language Proficiency Rating Scale (ISLPR), formerly the Australian Second Language Proficiency Rating Scale (ASLPR), which is "widely used in Australia to assess the general language proficiency of adult ESLK learners" (Brindley, 1995, p. 3).

REFERENCES

Breiner-Sanders, K. E., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL Proficiency Guidelines—Speaking, Revised 1999. *Foreign Language Annals, 33,* 13-18.

Brindley, G. (1995). Introduction. In G. Brindley, (Ed.), *Language assessment in action* (pp. 1-9). Sydney, Australia: National Centre for English Language Teaching and Research.

California Commission on Teacher Credentialing & National Evaluation Systems (n.d.). *California basic educational skills test (CBEST).* Sacramento, CA: Authors.

Center for Applied Linguistics (n.d.a). *About CAL.* Available: http://www.cal.org/admin/about.html

Center for Applied Linguistics (n.d.b). *Basic English skills test (BEST).* Washington, DC: Author.

Center for Applied Linguistics (n.d.c). *Computerized oral proficiency interview (COPI).* Washington, DC: Author.

Center for Applied Linguistics (n.d.d). *Simulated oral proficiency interview (SOPI).* Washington, DC: Author.

Chauncey Group International (n.d.). *Test of English for international communication (TOEIC).* Princeton, NJ: Author.

Council of Europe (1996). *Modern languages: Learning, teaching, assessment. A common European frame of reference.* Available: http://culture.coe.fr/lang/eng/eedu2.4.html.

Educational Testing Service (n.d.a). *Speaking proficiency English assessment kit (SPEAK).* Princeton, NJ: Author.

Educational Testing Service (n.d.b). *Test of English as a foreign language (TOEFL).* Princeton, NJ: Author.

Kenyon, D. M. (1998). An investigation of the validity of task demands on performance-based tests of oral proficiency. In A. J. Kunnan (Ed.), *Validation in language testing: Selected papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 19-40). Mahwah, NJ: Lawrence Erlbaum Associates.

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,* 85-101.

Language Testing International (n.d.). *American Council on the Teaching of Foreign Languages oral proficiency interview (ACTFL OPI).* White Plains, NY: Author.

Malone, M., Carpenter, H., Winke, P., & Kenyon, D. (2001, February). Development of a web-based listening and reading test for less commonly taught languages. Work in progress session presented at the Language Testing Research Colloquium, St. Louis, MO.

McNamara, T. F. (1996). *Measuring second language performance.* London and New York: Longman.

National Center for Education Statistics (n.d.). *National assessment of educational progress (NAEP).* Washington, DC: Author.

Stansfield, C. W., & Kenyon, D. M. (1993). Development and validation of the Hausa Speaking Test with the "ACTFL Proficiency Guidelines." *Issues in Applied Linguistics, 4,* 5-31.

Stansfield, C. W., & Kenyon, D. M. (1996). Comparing the scaling of speaking tasks by language teachers and by the ACTFL Guidelines. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 124-153). Clevedon, UK: Multilingual Matters.

Woodcock, R., & Muñoz, A. F. (n.d.). *Batería Woodcock-Muñoz-Revisada (Batería-R).* Itasca, IL: Riverside Publishing.

Woodcock, R., & Johnson, M. B. (n.d.). *Woodcock-Johnson III Complete Battery.* Itasca, IL: Riverside Publishing.

Wylie, E., & Ingram, D. (2001). *International second language proficiency ratings (formerly the Australian second language proficiency ratings): About the ISLPR.* Available: http://www.gu.edu.au/centre/call/content4.html.