The Oral Proficiency Interview: A Research Agenda

Micheline Chalhoub-Deville
University of Iowa

Glenn Fulcher
University of Dundee

Abstract: *Many researchers and practitioners maintain that ACTFL's efforts to improve instructional practices and promote proficiency assessments tied to descriptors of what learners can do in real life have contributed significantly to second language teaching and testing. Similar endeavors in the area of research, however, are critically needed. Focusing on the oral proficiency interview (OPI), this article argues that ACTFL has a responsibility to its stakeholders to initiate a research program that generates a coherent combination of logical and empirical evidence to support its OPI interpretations and practices. The article highlights a number of high-priority areas—including delimiting purposes, examining interview discourse, documenting rater/interlocutor behavior, explicating the native speaker criterion, and investigating the OPI's impact on language pedagogy—that should be incorporated into the research agenda.*

Introduction

With more than 20 years of history, one could say that the proficiency movement has come of age. Efforts to develop and disseminate proficiency principles and products have been well rewarded. The proficiency movement's primary products, the ACTFL Guidelines and the oral proficiency interview (OPI) have not only prevailed but have asserted themselves in the second language field.

At this juncture, it is incumbent on the movement's most conspicuous and important organization, ACTFL, to capitalize on its successes and promote research that documents the soundness of its products and practices. In this article focusing primarily on the OPI, we emphasize the need for an ongoing and systematic research agenda that provides a coherent account of, and support for, the interpretations and uses of OPI ratings. We identify issues in several fundamental areas that need to be incorporated into this research agenda.

First, however, we present an overview of the historical circumstances that have governed OPI research and development. In this overview, we describe some of the circumstances that have shaped ACTFL products and services and have determined research goals and priorities. Moreover, a historical perspective may help the reader better appreciate the underpinnings of the issues highlighted in the research agenda discussed later.

A Critical History

The Early Years

In the United States, the OPI originated out of practical necessity. During World War II, the majority of U.S. military personnel did not have the skills needed to perform key foreign-language communication tasks.

Micheline Chalhoub-Deville (PhD, Ohio State University) is Associate Professor in the Foreign Language and ESL Education Program at the University of Iowa.

Glenn Fulcher (PhD, University of Lancaster, UK) is Director of the Centre for Applied Language Studies, University of Dundee, Scotland, UK.

In 1942, the Army Specialized Training Program (ASTP) was established to teach speaking in the fields of engineering, medicine, and area studies. Processing 140,000 learners from 1943 to 1944, it was the first U.S. training program designed "to impart to the trainee a command of the colloquial spoken form of a language and to give the trainee a sound knowledge of the area in which the language is used" (Angiolillo, 1947, p. 32).

Apparently, there had been a perception that the war was not proceeding well, in part because of the lack of practical language skills among key personnel. As Kaulfers (1944) put it, "[t]he urgency of the world situation does not permit of erudite theorizing in English about the grammatical structure of the language for two years before attempting to converse or to understand telephone conversations" (p. 137).

Consequently, promoting the teaching and assessment of practical language use became a driving force in government language schools. Teachers in the ASTP programs designed their own tests to assess the communicative ability of their students (Agard & Dunkel, 1948). These tests consisted of picture descriptions, sustained speech, and directed conversation (Barnwell, 1996).

It was Kaulfers (1944), however, who provided an early example of a rating scale for such tasks. For a three-part test (securing services, asking for information, and giving information), Kaulfers suggested two scoring categories: scope of oral performance and quality of oral performance. Quality was understood as intelligibility. The four-point scale for scope was:

1. Can make known only a few essential wants in set phrases or sentences.
2. Can give and secure the routine information required in independent travel abroad.
3. Can discuss common topics and interests of daily life extemporaneously.
4. Can converse extemporaneously on any topic within the range of his knowledge or experience (Kaulfers, 1944, p. 144).

These categories were to be made meaningful by trialing test tasks with bilingual speakers "whose oral efficiency in real life situations is already known from outside evidence, such as types of professional employment abroad, etc. The scores used by these bilingual subjects can then be used to provide norms that can be interpreted in terms of quality and range of ability to speak the language in actual life" (Kaulfers, 1944, p. 141).

In other words, Kaulfers recognized that to certify practical language skills, scoring should be grounded in observations of actual performance, and he embarked on setting a research agenda for language performance tests. These early attempts to develop a research program, unfortunately, were not followed through and Kaulfer's fledgling research agenda never came to fruition.

Fluctuations in foreign language support and funding, often related to national security, could partly explain why the research program was not implemented. In an article in *The Washington Post* on October 23, 2001, U.S. Senator Paul Simon wrote:

> In every national crisis from the Cold War through Vietnam, Desert Storm, Bosnia and Kosovo, our nation has lamented its foreign language shortfalls. But then the crisis "goes away," and we return to business as usual. One of the messages of Sept. 11 is that business as usual is no longer an acceptable option. (p. A23)

Simon's statement clearly explicated the ebb and flow in support for foreign languages over the years. When he cautioned against "business as usual," he implied the need to marshal resources to provide sustained support for foreign language training. Unfortunately, such commitments are typically forgotten in times of relative peace. Furthermore, in an environment where language instruction is not a priority, support for research may be considered superfluous.

Given the absence of a research program to support language testing procedures, one would think that efforts to disseminate assessment methods would be held in check. This was not the case, as dissemination of the interview procedure and the corresponding scale in both government and academic communities sped ahead of research with unconstrained vigor.

The Interagency Language Roundtable

Immediately after World War II, the impetus to teach and test the speaking of foreign languages was lost. Although the Foreign Service Institute (FSI) had been set up soon after the war (Kramsch, 1986), interest in training and testing in speaking was only revived with the Korean War (1950–1953). Military need led to the Civil Service Commission requiring a register that documented personnel's familiarity with foreign languages and cultures (Liskin-Gasparro, 1984b). The FSI then drew up a six-band scale that rated the oral language proficiency of personnel. Sollenberger (1978) reported on the development of this intuitive, holistic rating scale, which defined the lowest band (or level) as "no ability" and the highest band as "native speaking ability."

The work of the FSI was the beginning of something that was no longer just a matter of pragmatic answers to practical military communication problems; it represented the bureaucratization of language testing. Testing became a tool within the government and military bureaucracy, and there was no place for the research initially envisaged by Kaulfers. Gone was the notion that levels should be tied to descriptions of the performance of competent bilinguals, a concept that was replaced by a six-level linear scale ranging from no ability to native speaker ability.

The new rating scale was first used in 1956 and published in 1958. Considered to be the grandfather of all language-rating procedures (Fulcher, 1997), the scale underscored five traits: accent, comprehension, fluency, gram-

mar, and vocabulary (Adams, 1980; Wilds, 1979). Each trait was measured on the 6-point rating system. In practice, however, it appears that a holistic scale was used in live rating and the role of the primary traits was to allow the raters to reflect on the possible meaning of the holistic score (North, 1993a).

Confidence in the new testing procedures developed by the FSI was so high that during the 1960s they were adopted (and adapted) by the Defense Language Institute, the Central Intelligence Agency (CIA), and the Peace Corps (Barnwell, 1996). For example, Quinones (no date) detailed how the CIA adapted the FSI system for its own use.

The most important feature of the adapted FSI scale was the use of multiple raters and an averaging system in an attempt to increase reliability. (Rater reliability has consistently been a concern for users of the interview procedure, but, as argued below, rater reliability is a necessary but not sufficient condition for rating quality.)

In 1968, the diverse agencies mentioned above came together to produce a standardized version of the levels, known as the Interagency Language Roundtable (ILR) (Jones, 1975; Lowe, 1987), which is still in use today and available on the Internet at http://www.dlielc.org/testing/round_table.pdf.

The standardization of the proficiency scale across agencies was prompted by lessons learned during the Vietnam War, when difficulties in assigning personnel to language-related tasks were encountered. These problems were noted to result from inconsistencies among the descriptions of language proficiency levels used by the various agencies.

In conclusion, the ILR scale's history of use and institutionalization, combined with the continued importance of language and culture to U.S. foreign policy, suggest that the scale will remain entrenched within U.S. government testing for the foreseeable future. However, the spread of the ILR scale has not been accompanied by a research program examining its fundamental assumptions about language use and development (Fulcher, 2003). Interestingly, this absence of a coherent research agenda in government language circles is also observed in the academic language community, which has uncritically adopted the FSI interview procedure and scale.

The American Council on the Teaching of Foreign Languages

The FSI interview procedure and corresponding scale had an impact on language education in general; use of the interview and scale spread from the defense agencies and Peace Corps to schools (Liskin-Gasparro, 1984b). In the 1970s, the interview and scale were adopted by many universities and states for the purpose of bilingual teacher certification (Liskin-Gasparro, 1984b). This expanding use was accelerated through Testing Kit Workshops (Adams & Frith, 1979), which were initially conducted for teachers of French and Spanish.

In terms of research or documentation of the quality of the interview and scale, investigations have focused primarily on reporting interrater reliability. The correlations for speaking test ratings between teachers and FSI raters and reliability indices were reported to be consistently higher than .84 (Adams, 1980). As will be argued later in this article, interrater reliability is a necessary first step to validation but does not by itself validate the interpretations and uses of the ratings. Nevertheless, the popularity of the OPI and reports of high interrater reliability quickly led to the overt, if unsubstantiated, belief that the interviews and rating scales in themselves possessed "psychological reality" in terms of second language use and development.

The work of Carroll (1967) became the focus of considerable attention during efforts to revise the FSI system to accommodate practical language instruction. Carroll had administered the OPI to college majors of French, German, Russian, and Spanish in the United States, and concluded that very few college majors in these foreign languages were capable of achieving a level above 2/2+. (The FSI rating system set Level 3 as the minimum for professionals working in the various government agencies.) Carroll's study was replicated in 1979 (Liskin-Gasparro, 1984b).

It was argued that if the ILR approach to testing speaking was to be used by universities, colleges and schools, the rating scales would need to allow more discrimination below the ILR 2+ level. It was argued that it would not be appropriate for students to spend many hours studying foreign languages and register little or no progress. As a result, the Educational Testing Service (ETS) and ACTFL became involved in revising the ILR rating scale to suit the academic language community (Clark, 1988; Liskin-Gasparro, 1984a, 1984b; Lowe, 1983, 1985, 1987).

Revisions included creating subdivisions at the lower levels of the ILR scale to accommodate and describe smaller increments in proficiency. Categories at the upper levels of the scale (i.e., above 2+) were collapsed, because few students in academic language programs were expected to achieve those levels. In addition, the ILR levels "were renamed to correspond to the needs and purposes of the academic community" (Omaggio, 1986, p. 12).

Academia's adoption of the FSI interview procedure and scale, albeit in a revised format, was also prompted by the President's Commission on Foreign Language and International Studies' report "Strength through Wisdom: A Critique of U.S. Capability" to President Carter (Strength through Wisdom, 1979). Among the recommendations of the report was the setting up of a National Criteria and Assessment Program to develop language tests and assess language learning in the United States, and the FSI interview and scale were seen as a valuable step in this direction. As a result, the *ACTFL Provisional Proficiency Guidelines* appeared in 1982 (ACTFL, 1982). The complete

Guidelines were published in 1986 (ACTFL, 1986) and the revised Guidelines in 1999 (Breiner-Sanders et al., 2000).

With the inception of the Guidelines, ACTFL proponents were engaged in the demanding task of developing training and testing materials, primarily to help change classroom practices (Omaggio, 1986). Similar efforts, however, were not manifest in a research program.

The ACTFL venture is more than 20 years old and its influence on instructional practices has been declared a success (Liskin-Gasparro, 2001). To continue to grow in leadership and to help guard against damaging litigation, it is critical for the organization to design and sponsor a research agenda that provides coherent documentation of the OPI's qualities with respect to the various interpretations and uses of the procedure by its proponents. As we point out in the next section, testing in the foreign language profession should be held to the same standards that apply to the rest of the U.S. educational testing industry.

The remainder of the article identifies several areas of high priority for an ACTFL research agenda. A comprehensive research agenda for the ACTFL Guidelines, as well as ACTFL's products and services, is beyond the scope of this article. Therefore, we address a few of the most important issues concerning the OPI procedure and ratings, which have been widely discussed in the published literature and still require a concerted research effort. These issues are validity and reliability, purpose, interview talk, rater behavior, native speaker criterion, and classroom impact.

An OPI Research Agenda

ACTFL founded a professional testing agency, the Language Testing International (LTI), to handle the growing demand for OPI ratings. These ratings are increasingly being used to make high-stakes decisions—decisions that have a considerable effect on individual lives in terms of licensure and certification, employment, promotion, admission, and graduation (see Swender, 1999). It is critical, therefore, for the OPI to yield ratings that provide high-quality information and for ACTFL to provide evidence that validates the intended interpretations and uses of these ratings.

Moreover, in this age of accountability, it is incumbent on a leading professional organization to ensure that its products meet high standards. For example, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 1999)—a widely recognized publication that has been endorsed by the International Language Testing Association—aims to promote rigorous and ethical test development. It provides principles for the evaluation of tests, test practices, and test use impact, and outlines the responsibilities of those involved in the testing operation.

As stated in the preface, although these standards for education and psychological testing are "prescriptive, the *Standards* itself does not contain enforcement mechanisms" (AERA, 1999, p. viii). Nevertheless, it is expected that the principles set forth be carefully considered before a test is made operational. In cases in which "test developers, users, and when applicable, sponsors have deemed a standard to be inapplicable or unfeasible, they should be able, if called upon, to explain the basis for their decision" (p. 3). In summary, the *Standards for Educational and Psychological Testing* ("Standards") represents an accepted code of practice for all testers, including those in the foreign language field.

It is important to note that adherence to the principles established by the Standards is not just a theoretical exercise. In the case of a test taker or subgroup of the test-taking population challenging the test or its consequences, litigation is the likely outcome. The Standards have been referred to in diverse court cases. For example, in the United States versus North Carolina (1975), the state was forced to withdraw a test because the educational authority could not show that cut scores had been established according to principles laid down in the Standards (McDonough & Wolf, 1988).

Kleinman and Faley (1985) discussed cases involving performance tests in which the defendants were required to show that rating scales were not "ambiguous" and that raters had been appropriately trained in their use. Indeed, in litigation where there was no research to support key validity issues as described in the Standards, the outcome was often in favor of the plaintiff (Fulcher & Bamford, 1996).

In following the Standards' guidelines, a variety of issues need to be investigated and documented. First and foremost, validity and reliability must be determined.

Validity and Reliability

The OPI tester training manual (Swender, 1999) states:

> [T]he OPI is a valid and reliable assessment of spoken language ability. It is valid because it measures the language functions, contexts and context areas, text type and accuracy features as described in the *ACTFL Proficiency Guidelines–Speaking (Revised, 1999)*. It is reliable because large groups of trained testers and raters consistently assign the same ratings to the same samples. (p. 4) (italics in original).

These validity and reliability statements fall short of documenting sound evidence.

> The Standards (AERA et al., 1999) defines validity as: . . . the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. (p. 9)

This conceptualization clearly highlights that ACTFL's notion of validity is not in line with the standards by which

testing professionals are guided. Validation, according to the Standards, is based on an evaluation of the interpretations of scores in the context of their use and not of the test content itself, as asserted by ACTFL. Although content-based evidence (that is, gathering information about the relevance and the representativeness of test content with regard to the specified second language domain) is important, it is not sufficient for constructing the validity argument.

Indeed, a study by Chalhoub-Deville (2001) showed that several OPI-based assessments that were closely related to the ACTFL Guidelines in their design and development failed to provide a meaningful representation of test takers' performances. The study revealed strong method effects that "mask the knowledge and skills that underlie performance ratings and undermine appropriate interpretation and use of test scores" (p. 225). In short, the incorporation of the Guidelines' attributes into the design of a test provides only preliminary evidence of validity.

Future research should investigate not just the content of the OPI but also the inferences of second-language proficiency represented in an OPI rating. The appropriateness of these inferences cannot be assumed or declared by simply asserting a connection to the ACTFL Guidelines, but needs to be supported by research. The sections that follow give special attention to validation research for a variety of OPI issues that influence score interpretation and use.

In reporting reliability, the Standards (AERA et al., 1999) emphasize the importance of examining the extent to which scores obtained are "dependable, and repeatable for an individual test taker" (p. 180). In other words, reliability indices document the degree to which scores or ratings represent test takers' true scores, and not errors of measurement due to varied testing conditions.

Given the OPI reliance on the interviewer and the subjective scoring system, differences in interviewers and inconsistency in ratings among raters are both sources of measurement variability. Therefore, in addition to rater agreement, which is occasionally reported by ACTFL (e.g., Thompson, 1995), other analyses such as generalizability studies (Brennan, 2001) need to be undertaken to help document the consistency of the assessments and the ensuing scores. Generalizability studies provide more appropriate estimates of variability arising from different sources (e.g., rater and interviewer), because these sources are simultaneously examined.

The interviewer and the scoring system are but two sources of measurement error. Another significant source of variability is the method of testing. Several studies (e.g., Clifford, 1981; Chalhoub-Deville, 1995a, 1995b; Henning, 1983; Shohamy, 1983) found that test scores were greatly influenced by the method used to measure test takers' proficiency. Research that compares proficiency ratings from the OPI with those obtained from other oral measures is needed. This article focuses only on the OPI, and so the research agenda proposed here does not address the comparison of testing methods. However, a larger conceptualization of an OPI research agenda would include such test method comparisons.

In conclusion, the dependability of the inferences about test takers' second-language proficiency, as summarized in an OPI rating, cannot be established simply via an interrater agreement index. Although the interrater agreement index quantifies a critical aspect of score reliability, it does not document variability caused by other salient aspects of the OPI procedure. As indicated, other types of reliability evidence are needed to document the extent to which ratings are consistent representations of test takers' second-language proficiency.

Purpose

The OPI tester training manual (Swender, 1999) asserts that "the applications of the OPI are limitless" (p. 5). The OPI is claimed to be suitable for a variety of academic, research, and professional functions. A claim that a single test can serve limitless functions or purposes is a cause for serious concern. Typically, different test purposes entail different test design considerations and require differentially targeted validation research. If "limitless" applications are intended, it becomes challenging—if not impossible—to delineate and accommodate all the knowledge, skills, and processes salient in different contexts.

Moreover, a test that suits all purposes creates validation chaos. In such a situation, it is not clear what research evidence should be targeted or given priority. For example, designing a test to admit students into a university typically requires an assessment of "macro" second-language skills, whereas a test for diagnostic purposes demands a fine-grain description of abilities. Furthermore, whereas research for admission tests emphasizes predictive evidence, a research agenda for diagnostic tests prioritizes identification of learners' strengths and weaknesses that can be addressed via follow-up instruction. Thus, it is critical that ACTFL delineates the intended purposes or uses of the OPI so that research investigating the appropriateness of OPI scores for those uses can be designed and conducted.

The Standards (AERA et al., 1999) explicitly states that "[n]o test is valid for all purposes or in all situations" (p. 17). According to Standard 1.1, "[a] rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation" (p. 17). ACTFL test developers should invest in clarifying the primary purpose(s) of the OPI, recommend specific interpretations and uses of the ratings, and present the empirical evidence and arguments that support these recommendations.

Interview Talk

The OPI and the FSI interview are not only deemed "direct," but also claim to represent "natural conversation" (Clark, 1975, 1980; Jones, 1985). Bachman and Savignon (1986) and Bachman (1990) have questioned the assertions of directness. These researchers and others have contended that all language measures are indirect indicators of the second-language constructs. Moreover, claims of directness are discussed in the literature in terms of situational and interactional correspondence to real-life contexts. "Directness" needs to be examined and documented in terms of the extent to which developers are successful in designing tasks that incorporate language use features that resemble those observed in intended real-life contexts (Kane, Crookes, & Cohen, 1999). Additionally, and perhaps more importantly, directness points to the degree to which interaction between test task features and second-language abilities resembles that observed in proposed nontesting situations (Bachman & Palmer, 1996).

With respect to natural conversation, several studies have shown that test interviews like the OPI generate a special genre of language, different from normal conversational speech (Johnson, 2001; Lazaraton, 1992; Perrett, 1990; Silverman, 1976; van Lier, 1989). Features examined have included turn-taking patterns, lexical and syntactic structures, sequences of speech acts or rhetorical scripts, topic management, and others. Differences between interview talk and nontesting conversational speech are not surprising, according to interactional competence arguments based on Vygotsky's ideas (see Chalhoub-Deville, 2003; Chalhoub-Deville & Deville, in press). Test takers' performances are engendered by the transactions in which they engage. The dynamic language exchange between the interviewer and the test taker plays a primary role in shaping OPI performances (see Brown, 2003) and creating an interaction style or genre unique unto itself.

In conclusion, further research is needed to operationalize the OPI "natural conversation" features by documenting the relationship between interview talk and conversations observed in pertinent nontesting situations.

Rater Behavior

ACTFL has invested in an extensive training program for OPI interviewers and raters. Published research has reported high levels of agreement among OPI raters. For example, Thompson (1995) reported interrater reliability (.85–.90) for 795 OPIs in five languages. Although this rater reliability information is reassuring, similar evidence must be provided for all languages. Moreover, as already with regard to reliability, other analyses such as generalizability research need to be undertaken. Finally, while training and consistency are reassuring, they are not sufficient. Evidence is needed to investigate raters' behaviors.

Several studies reported that background factors (e.g., professional training, place of residence, linguistic experiences) affect the severity of rater (both native and nonnative speakers) judgments (Barnwell, 1989; Galloway, 1980; Hadden, 1991).

In addition to differences in severity, research by Brown (1995) and Chalhoub-Deville (1995a, 1995b) showed that background factors influence the type of criteria that raters adopt when evaluating speaking performance. Finally, Brindley (1991), Brown (1995), McNamara (1990), and North (1993b) maintained that raters do not necessarily apply the criteria learned during their training in their assessments of learners' second language oral ability. Rather, raters seem to employ their built-in and idiosyncratic rating criteria schemes. Thus, although documenting rating consistency is important, investigations that examine the criteria that raters with different backgrounds use in judging OPI performances is still needed.

Native Speaker Criterion

Both ACTFL and ILR scale descriptions at different levels identify the native speaker as the criterion or norm against which test takers' performances are compared.[1] For example, speakers are eligible for an Advanced-Mid rating if they "are readily understood by native speakers unaccustomed to dealing with non-natives" (Swender, 1999, p. 114). Speakers qualify for an Intermediate-High rating if they "can generally be understood by native speakers" (Swender, 1999, p. 115). Similar criteria appear in the government scales. Lowe (1986) stated that "[t]he ILR approach has permitted successful use of the WENS (well educated native speaker) concept as the ultimate criterion in government for over thirty years" (p. 394).

Several researchers (Bachman & Savignon, 1986; Lantolf & Frawley, 1985) questioned the notion of a monolithic native speaker criterion. Studies cited earlier on the idiosyncrasies of native-speaker rater behavior are relevant here as well. Despite such arguments and evidence, the ACTFL Guidelines and ILR Scales continue to employ the native speaker norm without explicating what this abstract notion represents. Perhaps the in-depth training that OPI raters receive enables them to operationalize and render the abstract native speaker norm concrete. Such evidence, however, is not available. It is unlikely, given published evidence on rater behavior, that a uniform interpretation of native speech by OPI raters is achieved. Research is thus needed to replace the vague native speaker norm with explicit criteria deemed critical for evaluating speakers' performances at different levels.

Classroom Impact

Since its inception, ACTFL's central goal has been to change curricular and instructional practices (Omaggio, 1986). Lowe (1987) maintained that the model of language

acquisition, as represented in the ACTFL and ILR scale descriptors, may exhibit its "ultimate utility ... beyond testing per se in its effect on curriculum" (p. 47). Omaggio declared more boldly that, given the experiential grounding of the ACTFL Guidelines and the ILR scales, "teachers can amend their expectations for students' linguistic and communicative development to conform to reality" (p. 35).

In a recent paper, however, Liskin-Gasparro (2001) recanted many of the proficiency proponents' early assertions. She acknowledged that the empirical bases of the ACTFL Guidelines are shaky and that claims about measuring conversational ability and making predictions of learners' performance in real-life have not been borne out. Nevertheless, she argued that the ACTFL activities have been a catalyst for change, especially in terms of classroom instructional practices. She writes that "a kind of folk pedagogy has emerged that associates with proficiency (and, by extension, with the ACTFL Guidelines) all manner of teaching practices that are considered communicative, educationally progressive, and culturally authentic" (p. 9). If classroom impact or "washback" was the goal of ACTFL and the proficiency movement, a fundamental question to ask is whether this goal has been achieved.

Many researchers and practitioners would undoubtedly acknowledge that in the last two decades, the ACTFL products have influenced or perhaps even shaped second language instruction in the United States. Such acknowledgement, however, amounts to no more than anecdotal evidence. Empirical research is needed to support anecdotal observations. Research should be conducted to document the extent and type of impact. Such research could help ACTFL improve its ability to systematically foster its products and services to the benefit of instructional practices.

Washback or impact investigations are a relatively new area of research (see Wall, 1997). ACTFL, with its extensive experience, connections with educators, and access to various data (e.g., classroom enrollment, organization membership, teacher certification, second language policies) is well positioned to take the lead in expanding the knowledge base in this area.

Conclusion

More than a quarter of a century ago, Jones (1975) wrote that very little was known about the FSI, as "no validation studies have been made" (p. 4). This remains largely true today for OPI tests. The initial studies that provided some empirical basis for claims made about the OPI and the ACTFL Guidelines have not been entirely convincing. Studies such as those by Dandonoli and Henning (1990) and Henning (1992) have been questioned on a variety of instrumentation and methodological grounds (Fulcher, 1996). Other studies (e.g., Henry, 1996, and Thompson, 1995) presented findings that failed to support the hierarchy proposed by the Guidelines. Nevertheless, such studies have provided useful groundwork and direction for further research.

In this article, we have argued that ACTFL has a responsibility to its many stakeholders to initiate a comprehensive, focused research program to collect appropriate and meaningful evidence that documents the quality of OPI practices and ratings. We have highlighted a number of high-priority areas for research, including delimiting OPI purposes, researching the features of the interview discourse, documenting rater/interlocutor behavior, explicating the use of the native speaker criterion, and investigating the impact on language pedagogy. These areas should be incorporated into a larger research agenda that would outline the investigations needed to support the claims that ACTFL makes about the interpretations and uses of OPI ratings.

As we have argued, our call for this research agenda is supported by the code of practice in the testing community, namely, the *Standards for Educational and Psychological Testing* (AERA et al., 1999). Moreover, evidence established by this research agenda will enable ACTFL professionals, especially those involved with LTI, to formulate coherent arguments in defense of their practices in a variety of arenas.

Notes

1. The concept of the native speaker (NS) is employed differentially at different proficiency levels. At the top of the scale, the NS concept is used to describe the ideal against which the test candidate's performance is to be judged. In the ILR scale, the NS concept appears as a description of the type of communicative ability expected of test takers at Level 5. While earlier versions of the ILR scale used the term "native speaker," that concept was narrowed in the 1985 version and the label was changed to read "well-educated native speaker." This attempt at better explicating the criterion of NS is commendable. Nevertheless, it would still be important to empirically define and validate this "anchor point" of the ILR scale. (As explained above in this article, the ACTFL scale does not describe ability at this high level.). In the lower ranges of both the ILR and ACTFL scales, the term NS describes not the speaker, but the type of listener, who would likely be able to comprehend the speech of the test candidate. These "common sense" definitions have yet to be empirically validated.

References

Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In J. R. Frith (ed.), *Measuring spoken language proficiency* (pp. 1–6). Washington, DC: Georgetown University Press.

Adams, M. L., & Frith, J. R. (1979). *Testing kit: French and Spanish.* Washington DC: Foreign Services Institute, U.S. Department of State.

Agard, F. & Dunkel, H. (1948). *An investigation of second language teaching.* Chicago: Ginn and Company.

American Council on the Teaching of Foreign Languages. (1982). *ACTFL provisional proficiency guidelines.* Yonkers, NY: ACTFL.

American Council on the Teaching of Foreign Languages. (1986). *ACTFL proficiency guidelines.* Yonkers, NY: ACTFL.

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

Angiolillo, P. (1947). *Armed forces foreign language teaching.* New York: Vanni.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bachman, L. F. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *Modern Language Journal, 70,* 380–90.

Barnwell, D. (1989). Naïve NSs and judgments of oral proficiency in Spanish. *Language Testing, 6,* 152–63.

Barnwell, D. (1996). *A history of foreign language testing in the United States.* Tempe, AZ: Bilingual Press.

Breiner-Sanders, K., Lowe Jr., P., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines—speaking, revised. *Foreign Language Annals, 33,* 13–18.

Brennan, R. (2001). *Generalizability theory.* New York: Springer-Verlag Inc.

Brindley, G. (1991). Defining language ability: the criteria for criteria. In S. Anivan (ed.), *Current developments in language testing* (pp. 139–64). Singapore: Regional Language Center.

Brown, A. (1995). The effect of rater variables in the development of occupation-specific language performance test. *Language Testing, 12,* 1–15.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20,* 1–25.

Carroll, J. B. (1967). The foreign language attainments of language majors in the senior year: A survey conducted in U.S. colleges and universities. *Foreign Language Annals, 1,* 131–51.

Chalhoub-Deville, M. (1995a). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12,* 16–33.

Chalhoub-Deville, M. (1995b). A contextualized approach to describing oral language proficiency. *Language Learning, 45,* 251–81.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14,* 3–22.

Chalhoub-Deville, M. (2001). Task-based assessment: a link to second language instruction. In M. Bygate, P. Skehan, & M. Swain (eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 210–28). Harlow, UK: Longman.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing,* 20, 369–83.

Chalhoub-Deville, M. & Deville, C. (in press). A look back at and forward to what language testers measure. In E. Hinkel (ed.), *Handbook of research in second language teaching and learning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Clark, J. L. (1975). Theoretical and technical considerations in oral proficiency testing. In R.L. Jones & B. Spolsky (eds.), *Testing language proficiency* (pp. 10–28). Arlington, VA: Center for Applied Linguistics.

Clark, J. L. (1980). Toward a common measure of speaking proficiency. In J. R. Frith (ed.), *Measuring spoken language proficiency* (pp. 15–26). Washington, DC: Georgetown University Press.

Clark, J. L. (1988). *The proficiency-oriented testing movement in the United States and its implications for instructional program design and evaluation.* Monterey, CA: Defense Language Institute.

Clifford, R.T. (1981). Convergent and discriminant validation of integrated and unitary language skills: The need for a research model. In A.S. Palmer, P.J.M. Groot, & G.A. Trosper (eds.), *The construct validation of tests of communicative competence* (pp. 62–70). Washington, DC: Teachers of English to Speakers of Other Languages.

Dandonoli, P. & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals, 23,* 11–22.

Fulcher, G. (1996). Invalidating validity claims for the ACTFL Oral Rating Scale. *System, 24,* 163–72.

Fulcher, G. (1997). Testing speaking. In D. Corson & C. Clapham (eds.), *Encyclopaedia of language and education, vol. 7: Language testing and assessment* (pp. 75–86). Dordrecht: Kluwer.

Fulcher, G. (2003). *Testing second language speaking.* London: Longman/Pearson Education.

Fulcher, G., & Bamford, R. (1996). I didn't get the grade I need. Where's my solicitor? *System, 24,* 437–48.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64,* 428–33.

Hadden, B. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning, 41,* 1–24.

Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language Learning, 33,* 315–32.

Henning, G. (1992b). The ACTFL oral proficiency interview: validity evidence. *System, 20,* 365–72.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *Modern Language Journal, 80,* 309–26.

Johnson, M. (2001). *The art of non-conversation.* New Haven, CT: Yale University Press.

Jones, R. L. (1975). Testing language proficiency in the United States government. In R.L. Jones & B. Spolsky (eds.), *Testing language proficiency* (pp. 1–9). Arlington, VA: Center for Applied Linguistics.

Jones, R. L. (1985). Some basic considerations in testing oral proficiency. In Y. P. Lee (ed.), *New directions in language testing* (pp. 77–84). New York: Pergamon Institute of English.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 19*, 5–17.

Kaulfers, W. V. (1944). War-time developments in modern language achievement tests. *Modern Language Journal, 70*, 366–72.

Kleinman, L., & Faley, R. H. (1985). The implications of professional and legal guidelines for court decisions involving criterion-related validity: A review and analysis. *Personnel Psychology, 38*, 803–33.

Kramsch, C. J. (1986). From language proficiency to interactional competence. *Modern Language Journal, 70*, 366–72.

Lantolf, J. P. & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal, 69*, 337–45.

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System, 20*, 37–86.

Liskin-Gasparro, J. E. (1984a). The ACTFL proficiency guidelines: Gateway to testing and curriculum. *Foreign Language Annals, 17*, 475–89.

Liskin-Gasparro, J. E. (1984b). The ACTFL proficiency guidelines: A historical perspective. In T. V. Higgs (ed.), *Teaching for proficiency: The organizing principle* (pp. 11–42). Lincolnwood, IL: National Textbook Co.

Liskin-Gasparro, J. E. (2001). L2 speaking as proficiency. Paper presented at the annual meeting of AAAL–LTRC/AAAL Joint Colloquium. St. Louis, MO.

Lowe, P. (1983). The ILR oral interview: Origins, applications, pitfalls, and implications. *Die Unterrichtspraxis, 16*, 230–44.

Lowe, P. (1985). The ILR proficiency scale as a synthesizing research principle: The view from the mountain. In C. J. James (ed.), *Foreign language proficiency in the classroom and beyond* (pp. 9–53). Lincolnwood, IL: National Textbook Co.

Lowe, P. (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and particularly Bachman and Savignon. *Modern Language Journal, 70*, 391–97.

Lowe, P. (1987). Interagency language roundtable proficiency interview. In J. C. Alderson, K. J. Krahnke, & C.W. Stansfield (eds.), *Reviews of English language proficiency tests* (pp. 43–47). Alexandria, VA: TESOL.

McDonough, M. W., & Wolf, W. C. (1988). Court actions which helped define the direction of the competence-based testing movement. *Journal of Research and Development in Education, 21*, 37–43.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*, 52–76.

North, B. (1993a). *Scales of language proficiency: A survey of some existing systems.* Strasbourg: Council of Europe, CC-Language (94) 24.

North, B. (1993b). *The development of descriptors on scales of language proficiency.* Washington, DC: National Foreign Language Center.

Omaggio, A. C. (1986). *Teaching language in context: Proficiency-oriented instruction.* Boston: Heinle & Heinle.

Perrett, G. (1990). The language testing interview: A reappraisal. In J. H. de Jong and D. K. Stevenson (eds.), *Individualizing the assessment of language abilities* (pp. 225–38). Philadelphia, Multilingual Matters.

Quinones, J. (no date). Independent rating in oral proficiency interviews. Central Intelligence Agency.

Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning, 33*, 527–39.

Silverman, D. (1976). Interview talk: bringing off a research instrument. In D. Silverman & J. Jones (eds.), *Organizational work: The language of grading, the grading of language* (pp. 133–50). London: Collier Macmillan.

Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. Clark (ed.), *Direct testing of speaking proficiency: Theory and application* (pp. 1–12). Princeton, NJ: Educational Testing Service.

Strength through wisdom: A critique of U.S. capability. (1979). *A report to the president from the president's commission on foreign language and international studies.* Washington, DC: U.S. Government Printing Office. [Reprinted in *Modern Language Journal*, 1984, *64*, 9–57.]

Swender, E. (ed.). (1999). *ACTFL oral proficiency interview tester training manual.* Yonkers, NY: ACTFL.

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*, 407–22.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489–508.

Wall, D. (1997). Impact and washback in language testing. In D. Corson & C. Clapham (eds.), *Encyclopaedia of language and education, vol. 7: Language testing and assessment* (pp. 291-302). Dordrecht: Kluwer.

Wilds, C. (1979). The measurement of speaking and reading proficiency in a foreign language. In M. L. Adams & J. R. Frith (eds.), *Testing kit: French and Spanish* (pp. 1–12). Foreign Services Institute, U.S. Department of State.