Pergamon

S0346-251X(96)00001-2

System, Vol. 24, No. 2, pp. 163-172, 1996
Copyright © 1996 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0346-251X/96 $15.00 + 0.00

INVALIDATING VALIDITY CLAIMS FOR THE ACTFL ORAL RATING SCALE

GLENN FULCHER

University of Surrey, Guildford, Surrey GU2 5XH, U.K.

Dandonoli and Henning [*Foreign Language Annals* **23**(1), 11–22 (1990)] and Henning [*System* **20**(3), 365–372 (1992)] aim to present evidence for the validity of the American Council on the Teaching of Foreign Languages (ACTFL) rating scales [*ACTFL Proficiency Guidelines* (1982, 1986)]. This paper examines the evidence presented in these publications and argues that it is not sufficient for a claim for validity to be made. Lantolf and Frawley [*ADFL Bulletin* **23**(2), 34–37 (1992)] indicate that they had "some concern about the experimental design and statistics" of the studies conducted by Dandonoli and Henning; the basis for their concern is investigated. Copyright © 1996 Elsevier Science Ltd

THE ACTFL RATING SCALE: A BRIEF DESCRIPTION

The ACTFL *Guidelines* (ACTFL, 1982, 1986) has become the basis for a widespread approach to the teaching and testing of foreign languages within the U.S. The rating scales that are used for the assessment of oral proficiency have grown out of the Foreign Services Institute (FSI) and Interagency Language Roundtable (ILR) rating scales. The history of this development is well documented (Sollenberger, 1978; Liskin-Gasparro, 1984a,b; Lowe, 1983, 1985, 1987; Barnwell, 1987; Clark, 1988).

The ACTFL oral rating scale was specifically designed with a larger number of levels than its predecessors in order to discriminate more accurately between students in U.S. non-government settings at the level below 2/2+. This change followed recommendations made as a result of studies by Carroll (1967) and ETS (Liskin-Gasparro, 1984b). These studies had shown that college language majors could not progress beyond this level. The ACTFL rating scale, unlike the ILR, was designed for use in schools and colleges, rather than in the U.S. government context. The argument for this change was that it was unreasonable to subject students to hours of language tuition from which there would be no evidence of progress on tests.

The ACTFL scale is marked by an increase in the amount of information that is provided for the rater (in comparison with the ILR rating scale). However, the lack of detailed explanation of the terms used or potential exponents in actual speech continues to be a mark of the prose descriptions. No empirical evidence is available to confirm that new criteria introduced into the rating scale such as "discourse", "interactive" or "communicative" strategies discriminate between the students at the proposed ability levels where these terms occur on the scale. Indeed, Lowe (1985: p. 16) states that "the use of the system remains implicit", which seems to imply that the descriptors, as they stand, must be interpreted by each reader, although rater training is essential for any practical use of the scale. Finally, it is also noted that the mixture of linguistic and non-linguistic criteria is increased in the ACTFL scale, indicating task type and topic area that may be dealt with at a given level of ability.

CRITICISMS OF THE ACTFL APPROACH TO TESTING

For many commentators, the most important failing of the ACTFL is, therefore, the fact that none of the scales have any "empirical underpinning" (Lantolf and Frawley, 1985, 1988; Pienemann, Johnstone and Brindley, 1988). Similarly, Valdman (1988: p. 121) argues that: "...it is fair to say that although the OPI may be experientially based, its theoretical underpinnings are shaky and its empirical support scanty".

One example of this is the criticism of Pienemann, Johnstone and Brindley (1988: p. 219) of the concept of "weaknesses" in language skills or knowledge that is used in the scale. They argue that:

> Such descriptions are so vague and general as to be utterly unhelpful in distinguishing any second language learner from another. If "areas of weakness" can be construed to mean areas in which learners' usage does not confirm to the standard, then every language learner conforms to this description. Numerous research studies have shown that learners do not suddenly "learn" a structure and begin to use it correctly 100% of the time . . . Even the most advanced of second language learners will therefore display "weaknesses" in the areas cited.

Another issue in the empirical foundation of the scale's construction to which attention has been drawn is the confusion of linguistic and non-linguistic criteria in the scale descriptors (Bachman and Savignon, 1986; Bachman, 1988; Matthews, 1990), as it makes validation studies extremely difficult. That is, one cannot distinguish between test method facets and traits when conducting validation studies. This may be one reason that the literature on the ACTFL rating scale consistently fails to provide adequate evidence of construct validity.

From this brief review, it would seem fair to conclude that without a sound empirical basis for initial rating scale development, it makes little sense to investigate the validity of an oral rating scale *post hoc*, when results cannot be related to initial hypotheses and constructs, and other confounding factors may have been introduced to the picture. Jarvis (1986: p. 21) is probably correct when he argues that: "After-the-fact inquiry is unacceptable and has historically degenerated into little more than validation of flawed systems."

Despite lack of empirical evidence, the model of language learning assumed by the ACTFL/ETS/ILR rating scales (often shortened to AEI) has become the basis for a whole approach to language teaching and testing in the U.S. known as the Proficiency Movement (Higgs, 1984). The wide acceptance of the principles of the movement has led some authors to make a strong claim for the validity of the AEI rating scales, based on what is essentially face validity. Thus, for example, the most common defence for the validity of the AEI oral tests and rating scales is that of experience. Liskin-Gasparro (1984a) uses this argument, but its strongest expression is found in Lowe (1986: p. 392), written in response to criticisms by Kramsch (1986) and Bachman and Savignon (1986). Lowe writes: "The essence of the AEI proficiency lies not in verbal descriptions of it, but in its thirty-year-long tradition of practice—making training in AEI proficiency testing a desideratum." This would appear to be an overt acceptance of the criticisms that this approach to oral testing has no basis in theory or evidence. Bachman (1988: p. 163) is surely correct when he says that having had years of experience in working with a rating scale "in no way constitutes evidence for validity".

AN EVALUATION OF RECENT EMPIRICAL EVIDENCE

Recently, studies by Dandonoli and Henning (1990) and Henning (1992) have attempted to present evidence that counter the claim that there is no empirical evidence to support the validity of the ACTFL rating scale.

They specifically attempt to deal with the criticisms that the ACTFL OPI does not take into account test method facets (Bachman and Savignon, 1986), that there is no evidence to suggest that the ACTFL rating scale has any of the properties of a measuring instrument, and that results cannot be generalized across languages. It initially needs to be stated, however, that although Henning (1992) refers to the work of Dandonoli and Henning (1990) as the only study published by ACTFL researchers providing empirical evidence for its reliability and validity, Henning (1992) is a summary of the 1990 study. The status of this study is, therefore, crucial for current views on the validity of the ACTFL testing system.

Dandonoli and Henning (1990) designed a multitrait–multimethod study (MTMM) to investigate claims that test method and trait were confounded, following the methodology and procedures of Campbell and Fiske (1959). A Rasch analysis (Rasch, 1960) was used to investigate the extent to which the ACTFL bands could be said to represent an acquisition hierarchy, or rather an instrument possessing the basic qualities of a measurement scale. Finally, a correlational study of the relationship between scores awarded by trained and untrained raters was conducted to discover to what extent naive raters were capable of interpreting the scale descriptors. These studies were carried out with data from learners of English and French. In what follows, this paper deals only with the results from the English tests to investigate problems with methodology and the tenuous nature of the evidence upon which validity is claimed.

A multitrait–multimethod study

The results of the MTMM study are set out in Table 1. In Table 1 the abbreviation TM stands for test methods, MC for multiple choice, OE for open ended questions, and A and B for two independent raters.

Henning and Dandonoli (1990) claim, on the basis of this study, that heterotrait–monomethod discriminant validity is achieved entirely by the speaking and reading traits, and in the case of writing and listening, achieving discriminant validity is only

Table 1. MTMM results for English (Dandonoli and Henning, 1990)

		Speaking		Writing		Listening		Reading	
	TM	*A*	*B*	*A*	*B*	MC	OE	MC	OE
Speaking	A	1.00							
	B	**0.97**	1.00						
Writing	A	0.85	0.86	1.00					
	B	0.95	0.88	**0.87**	1.00				
Listening	MC	0.80	0.79	0.77	0.80	1.00			
	OE	0.76	0.71	0.74	0.77	**0.84**	1.00		
Reading	MC	0.83	0.79	0.80	0.82	0.92	0.85	1.00	
	OE	0.76	0.75	0.78	0.79	0.86	0.76	**0.92**	1.00

half successful. For example, the correlation coefficient between rater *A* and rater *B* on speaking is 0.97. This should be greater than the correlation coefficients between rater *A* on writing and speaking (0.85), and rater *B* on writing and speaking (0.88). That is, the scores from different raters should be more highly correlated when they are rating the same trait than the correlation between scores awarded by the same rater on different traits. The evidence presented in Table 1 is taken to be a claim for the existence of four independent traits, two of which are not completely separable from other traits.

However, when considering the data provided in Table 1, the high value of all the correlations is particularly noticeable. It is suspected that the magnitude of the correlation coefficients does not justify a conclusion of convergent or divergent validity. Consistently high correlation coefficients in an MTMM matrix usually indicate that at least three (if not all four) of the traits do not exist. However, before analysing the correlation matrix again, it is necessary to note a number of design faults in the study.

In this study two raters were used. These two raters were defined as method facets for the skill areas of speaking and writing, and the skill modalities (reading, writing, speaking and listening) were defined as traits. Method facets in the reading and listening tests were test formats (multiple choice: MC, and open-ended questions: OE). This is a highly questionable practice. Individual raters are not normally treated in this way. The methods must be tests or radically different test formats. How using one rater instead of another can be said to constitute a different test is rather puzzling. The two raters who were used in the study were also trained and accredited ACTFL/ETS raters; it is thus quite possible that the results obtained by Dandonoli and Henning are an artefact of the rigorous training of the raters used. The correlation of 0.97 is high, and this could be expected to be considerably lower if the raters were untrained, and if real test method facets were introduced into the design. In the latter case, this would involve using two different sets of tasks and elicitation procedures, but asking the raters to award scores to the students on the same rating scale. These faults in the design of the study make the evidence presented very difficult to interpret.

A maximum likelihood confirmatory factor analysis

It should also be noted that Dandonoli and Henning do not follow up this analysis with a maximum likelihood (ML) confirmatory factor analysis (Jöreskog and Sörbom, 1969), which allows researchers to specify the theoretical model with which they are working and test to what extent the data fit the model. In cases where strong claims to validity are being made, this procedure should be used as a matter of course. As such a study was not conducted and, as has already been noted, the magnitude of the correlation coefficients in the matrix does suggest that there may not be four factors at work, an ML analysis was conducted.

The correlation matrix provided by Dandonoli and Henning (1990) was analysed with LISREL 7.16 (see Jöreskog, 1989), using a four factor model, which is made clear in the path diagram in Fig 1. Readers not familiar with LISREL analysis are advised to consult Long (1983). In Fig. 1, x refers to an actual "test", so that $x1$ is the speaking test using rater A and $x2$ is the speaking test using rater B, etc. δ (delta) is the measurement error associated with the x variable (test), and λ (lambda) is the hypothesized relationship between the x variable and the ξ (xi) variable. In this study it is elements of the λ matrix that can be "fixed" or "freed" to create the model that is tested against the data. ξ refers to a latent trait or underlying construct that can account for test scores. The ϕ (phi) matrix represents the interrelationship of ξ variables. It should also be noted that in this analysis, it was possible for each of the factors to be intercorrelated, as no element in the ϕ matrix was fixed. This allows the researcher to test the claim that four traits can account for the data, but that these traits may be intercorrelated. Indeed, the correlation matrix suggests that this is probably the case.

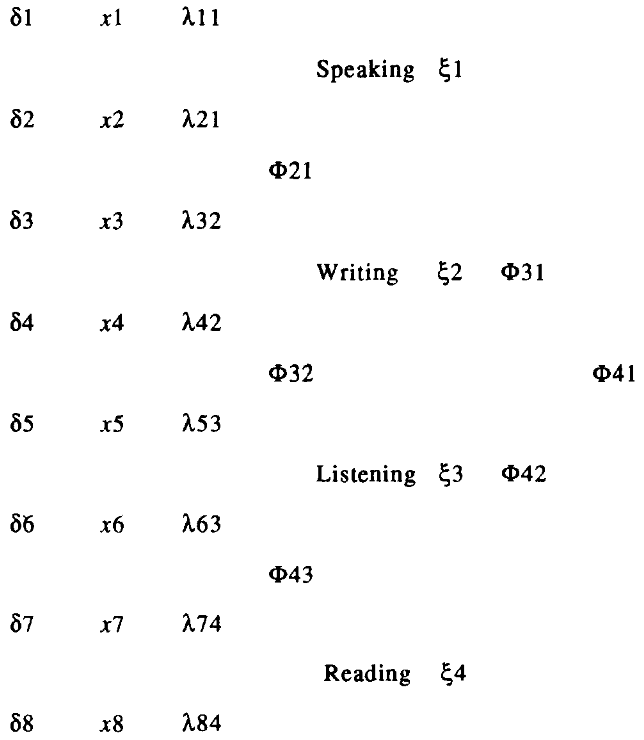

Fig. 1. Path diagram for a four factor model of the Dandonoli and Henning (1990) data.

An ML solution allows the researcher to investigate whether data fit a hypothesized model. The model is a statement of what patterns and relationships the researcher would expect to find in the data, if the model is an adequate theoretical account of what the test is supposed to test. A model may fit the data "more" or "less" well. In every case, through the analysis of the modification index, it may be seen which parameters of the model reflected in the path diagram fit least well, and there may be some theoretical justification for freeing those parameters and improving the model fit. However, if the initial estimate of model fit is extremely poor, LISREL will not proceed with an analysis, although it will provide some information to suggest why the model is not appropriate.

In this analysis, the X^2 statistic was 104.67, $P = .000$. If a X^2 statistic is significant, this indicates that the model does not fit the data. In this case, the misfit is enormous, and the data failed the "admissibility test". The admissibility test "detects very bad models after 10 iterations where conducting further iterations would be useless" (Jöreskog and Sörbom, 1989: p. 296). This test detects (in this case) whether ΛX has full column ranks and no rows of only zeros, and that the ϕ and $\theta\delta$ (theta delta) matrices are positive definite. The ϕ matrix has already been described. δ in the path diagram represents the measurement error in the test. $\theta\delta$ is the covariance of the measurement error for each test. These figures "can become negative if the data are unfavourable relative to the assumed model" (Jöreskog and Sörbom, 1989: p. 123). In the case of this model, $\theta\delta$ became negative for the speaking factor. The model was so poor that further analysis was terminated.

We may only conclude that although the initial MTMM results presented in Table 1 look promising, further empirical investigation shows that direct interpretation of the correlation matrix would be unwise. The reason for this is primarily that the correlation coefficients in the matrix are all exceptionally high. Under such circumstances, one is bound to get convergent validity and almost certain not to achieve divergent validity. Interpreting small correlational differences as significant, in terms of divergent validity, should be avoided.

Even a cursory look at the correlation matrix in Table 1 suggests that this is very "much of a muchness", and that differences between correlation coefficients may not be real. To investigate "real" differences, it would be necessary to calculate confidence intervals for the correlation coefficients. However, it is suspected that this would not produce much additional information. It is when correlations in a matrix are uniformly high, as is the case with the matrix in Table 1, that the researcher should suspect that the matrix is singular. It is when a matrix is singular that a LISREL analysis will produce negative entries on the $\theta\delta$ matrix, which means that it is not possible to interpret the model statistically.

The Rasch analysis

The Rasch study was designed to show that the rating scale does represent a true continuum—a basic requirement of a measuring instrument. The results are presented in Table 2.

In Table 2, N = the number of students who were rated at each level of the rating scale, and the mean is the average ability level (in logits) of those people. The standard error is

Table 2. Means and standard errors of estimated person abilities on a single test of speaking

Proficiency	*N*	Mean	Standard error
Novice low	1	na	na
Novice mid	6	–3.50	0.51
Novice high	2	–1.25	0.41
Intermediate low	24	–2.16	0.35
Intermediate mid	27	–0.70	0.16
Intermediate high	29	–0.05	0.13
Advanced	14	1.36	0.18
Advanced plus	8	2.32	0.29
Superior	7	3.97	0.53

the error expressed in logits, and these figures should be no larger than the difference between the mean of one level and the mean of another. If this is the case, then it suggests that the levels are discrete.

It appears from these results that the ACTFL rating scale successfully orders students on an instrument that has the appropriate properties of a measurement instrument, with the exception of novice high and intermediate low. Indeed, this is what Dandonoli and Henning (1990) claim. However, in this Rasch analysis of speaking ability there appear to be two problems. The first is that no method facets appear to be specified, as is usually the case with such studies (McNamara and Adams, 1991). The second is that, as has already been noted, the data for the study were generated by trained ACTFL/ETS raters. The results and hence the scale could have been created by test method facets, the influence of which have not been measured. Alternatively, the results could be an artefact of the rigorous rater training (or "cloning"), which is known to have an effect upon test results (Alderson, 1991). Or it could be a combination of the two. The data may indicate that a measurement scale has been created, but we just do not know from the evidence that has been presented what the underlying trait is.

Table 3. Spearman rank order correlation means, ranges and standard deviations between four native raters and two trained ACTFL raters for English

Mean correlation	Range	Standard deviation
0.934	0.904–1.000	0.045

Inter-rater reliability using naive judges

Finally, Dandonoli and Henning (1990) and Henning (1992) provide Spearman rank correlation coefficients between the two trained ACTFL/ETS raters and four untrained native speakers of English.

Dandonoli and Henning (1990) acknowledge that this was not in the original design of the study. It may, therefore, be somewhat unfair to subject the results to criticism. However, there are no descriptions of the precise nature of the study in the literature, other than that the selection of tapes for the study was made: "such that one tape . . . was reliably judged to be at each proficiency level" (Dandonoli and Henning, 1990: p. 20). If the tapes used represented samples of speech from students who were well spread out over an ability continuum (even intuitively defined) and also limited in number, such high rank order correlations would not be surprising. The study also fails to provide data concerning the degree to which, once rank ordered, the native and trained raters agreed on the band that described the performance of each of the sample tapes (see Barnwell, 1989, for similar studies on a principled basis).

Evidence that shows that trained and untrained raters can rank order four or five tapes of students at arguably different ability levels would not appear to constitute evidence for validity.

CONCLUSIONS

Although Henning (1992) significantly reduces the claims made for the ACTFL rating scale, there is evidence to suggest that the limited aims of providing "some limited research evidence related to the reliability, construct and criterion-related validity, scalability and generalizability of ratings obtained according to the ACTFL Oral Proficiency Interview" (p. 369) have still to be achieved.

The implication for research into the use of rating scales in the assessment of speaking is that it is vitally important for the researchers to consider construct validity at the test development stage of the process, rather than as a *post hoc* activity. This would ideally involve theory construction and empirical investigation into the relationship between data and theory. As Lantolf and Frawley (1992. p. 36) aptly state: "We think that it is better to make predictions than to wait for lucky breaks".

Acknowledgement—I would like to thank Dr P. K. Marriott of the Department of Mathematics and Computing Science, University of Surrey, for useful discussions on the results and interpretation of the LISREL analysis described in this study.

REFERENCES

ACTFL (1982) *ACTFL Provisional Proficiency Guidelines.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

ACTFL (1986) *ACTFL Proficiency Guidelines.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

ALDERSON, J. C. (1991) Dis-sporting life. Response to Alistair Pollitt's paper: "Giving students a sporting chance: assessment by counting and judging". In Alderson, J. C. and North, B. (eds) *Language Testing in the 1990s*, pp. 60–70. London: Modern English Publications and the British Council.

ALDERSON, J. C. and NORTH, B. (eds) (1991) *Language Testing in the 1990s.* London: Modern English Publications and the British Council.

BACHMAN, L. (1988) Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition* **10**(2), 149–164.

BACHMAN, L. and SAVIGNON, J. (1986) The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal* **70**(4), 380–390.

BARNWELL, D. (1987) Oral proficiency testing in the United States. *British Journal of Language Teaching* **25**(1), 35–42.

BARNWELL, D. (1989) Naive native speakers and judgements of oral proficiency in Spanish. *Language Testing* **6**(2), 152–163.

CAMPBELL, D. T. and FISKE, D. W. (1959) Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin* **56**(2), 81–105.

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**(2), 131–151.

CLARK, J. L. D. (1988) *The Proficiency-Oriented Testing Movement in the United States and its Implications for Instructional Program Design and Evaluation,* Mimeo: Defense Language Institute.

DANDONOLI, P. and HENNING, G. (1990) An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals* **23**(1), 11–22.

HENNING, G. (1992) The ACTFL oral proficiency interview: validity evidence. *System* **20**(3), 365–372.

HIGGS, T. V. (ed) (1984) *Teaching for Proficiency, the Organizing Principle.* Lincolnwood, IL: National Textbook Company.

JARVIS, G. A. (1986) Proficiency testing: a matter of false hopes? *ADFL Bulletin* **18**(1), 20–21.

JÖRESKOG, K. G. (1989) *LISREL 7: A Guide to the Program and Application.* New York: SPSS Inc.

JÖRESKOG, K. G. and SÖRBOM, D. (1969) A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34**(2), 183–202.

JÖRESKOG, K. G. and SÖRBOM, D. (1989) *LISREL 7.* Mooresville, IN: Scientific Software Ltd.

KRAMSCH, C. J. (1986) From language proficiency to interactional competence. *Modern Language Journal* **70**(4), 366–372.

LANTOLF, J. P. and FRAWLEY, W. (1985) Oral proficiency testing: a critical analysis. *Modern Language Journal* **69**(4), 337–345.

LANTOLF, J. P. and FRAWLEY, W. (1988) Proficiency: understanding the construct. *Studies in Second Language Acquisition* **10**(2), 181–195.

LANTOLF, J. P. and FRAWLEY, W. (1992) Rejecting the OPI – again: a response to Hagen. *ADFL Bulletin* **23** (2), 34–37.

LISKIN-GASPARRO, J. (1984a) The ACTFL proficiency guidelines: gateway to testing and curriculum. *Foreign Language Annals* **17**(5), 475–489.

LISKIN-GASPARRO, J. (1984b) The ACTFL proficiency guidelines: a historical perspective. In Higgs, T. V. (ed.) *Teaching for Proficiency, the Organizing Principle,* pp. 11–42. Lincolnwood, IL: National Textbook Company.

LONG, J. S. (1983) *Confirmatory Factor Analysis: A Preface to LISREL.* London: Sage.

LOWE, P. (1983) The ILR oral interview: origins, applications, pitfalls, and implications. *Die Unterrichtspraxis* **16**, 230–244.

LOWE, P. (1985) The ILR proficiency scale as a synthesizing research principle: the view from the mountain. In James, C. J. (ed.) *Foreign Language Proficiency in the Classroom and Beyond,* pp. 9–53. Lincolnwood, IL: National Textbook Company.

LOWE, P. (1987) Interagency language roundtable proficiency interview. In Alderson, J. C., Krahnke, K. J. and Stansfield, C. (eds) *Reviews of English Language Proficiency Tests,* pp. 43–47. Washington D.C.: Teachers of English to Speakers of Other Languages.

LOWE, P. (1986) Proficiency: panacea, framework, or process? A reply to Kramsch, Schulz, and particularly BACHMAN and SAVIGNON. *Modern Language Journal* **70**(4), 391–397.

MATTHEWS, M. (1990) The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations. *English Language Teaching Journal* **44**(2), 117–121.

MCNAMARA, T. F. and ADAMS, R. J. (1991) Exploring rater behaviour with Rasch techniques. Paper presented at the 13th Language Testing Research Colloquium, Educational Testing Service, Princeton NJ, 21–23 March.

PIENEMANN, M., JOHNSTON, M. and BRINDLEY, G. (1988) Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition* **10**, 217–234.

RASCH, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Denmarks Paedagogiske Institut.

SOLLENBERGER, H. E. (1978) Development and current use of the FSI oral interview test. In Clark, J. L. D. (ed.) *Direct Testing of Speaking Proficiency: Theory and Application*, pp. 1–12. Princeton: Education Testing Service.

VALDMAN, A. (1988) The assessment of foreign language oral proficiency: introduction. *Studies in Second Language Acquisition* **10**(2), 121–128.