2+
Will understand gist of questions, occasionally misinterprets intent. May miss some detail.

2
May grasp gist and follow general directions but may misinterpret intent of questions, will miss some detail.

2-
Some repetition and/or rephrasing necessary in order to get gist - may still miss some details.

1+
Has problem getting gist and details even with repetition and rephrasing.

RESEARCH IN TESTING

Some Priority Areas for Research in Oral Language Testing

Glenn Fulcher
University of Surrey

Any attempt to isolate priority areas for research in the field of oral testing must of necessity be somewhat subjective. The areas which are highlighted here are based on a recently completed PhD in the construct validation of oral tests at the University of Lancaster, supervised by Professor Charles Alderson (Fulcher, 1993).

1. **Establishing divergent validity for hypothesised constructs**
Very little is known about what constitutes 'speaking ability', even though many rating scales currently in use in oral tests assume that there are component competencies. The descriptors of holistic rating scales often merely fudge the issue by referring to a number of abilities within one band, assuming that these come in neatly packaged bundles. If only ...

What is needed is a clear linguistic definition of at least two component traits which may be operationalised in rating scales (or some other scoring method) which can be studied through the use of Multitrait-Multimethod designs (MTMM), and Maximum Likelihood Confirmatory Factor Analysis (ML). However, the main problem in conducting such studies with data from oral tests is the effect of cross-contamination: raters giving a score on one scale and then reproducing that across other scales. This must be dealt with in future studies.

It is suggested that one way of overcoming this problem is to write rating scale specifications, which would contain as full a description of the trait to be measured as possible and an indication of the number of bands to be used in the rating scale. Two rating scales

would then be produced for each trait by separate scale writers on the basis of the specifications.

Sets of raters, using the different rating scales on different occasions, would then rate a small number of videos of students doing the same task. This would, of course, introduce a time facet and a potential order effect. To control for this each team of raters would have to start with a different rating scale.

A Partial Credit model could be used to assess the rating pattern of the raters on the two scales, and an MTMM/ML study could be devised in which the methods were rating scales. It would be hypothesised that scores on different rating scales of the same trait would exhibit convergent validity and that scores on rating scales of different traits would exhibit divergent validity, as the design of the data collection technique would presumably reduce the effect of cross-contamination. It should however be pointed out that this design is not strictly speaking an MTMM design, as *each method is a rating scale*. This must be borne in mind when considering convergent validity, but should not affect the study of divergent validity.

Another approach to this problem may be to use recently developed semi-direct tests of oral proficiency in conjunction with a rating procedure which did not require raters to make any judgements at all. For example, the taped responses of students to a completely standardised elicitation technique could be transcribed and rated on the basis of purely linguistic criteria hypothesised to be related to two or more distinct traits. This method would not be as powerful as that described above, as it would of necessity have to concentrate only on surface elements of speech, but it would at least give some insight into whether or not simple counts of the occurrence of surface elements said to be exponents of the separate traits would result in the students being awarded different scores on those traits. This type of design is, in fact, very much like that used by Pienemann et al (1988) in the construction of their oral observation forms.

Yet another approach to the problem would be to conduct an introspective study. In this design a large number of raters would be used to rate a small number of students taking two or three oral tests on a number of component rating scales. However, prior to doing this, all raters would be interviewed about their conception of the nature of the traits on which they would be asked to rate. These interviews would have to be recorded, transcribed, and the conceptual baggage which each rater brings to the rating process thoroughly documented. Immediately after this the raters would be asked to rate the students, and the rating patterns in practice could then be related directly to the introspective data in order to see to what extent prior views of raters affected their use of the rating scale. It might also be possible to arrange for the raters to introspect as they were rating, commenting on their interpretation of the rating scale, and to arrange a retrospective study in which raters were shown their actual rating pattern and asked to compare this to what they thought they were doing at the time. Such a study would throw a great deal of light onto the rating process which at the present time we know very little about.

If at least two component traits could be demonstrated to have convergent and divergent validity, the latter being the more difficult to achieve, research could then be conducted into other hypothesised traits in order to build a more comprehensive theoretical model of speaking ability.

2. Task and Topic

We do not know a great deal about task difficulty. Much more research is needed into task difficulty along the lines of Stansfield and Kenyon (1992), but using scores from actual oral tests rather than relying on the judgements of teachers. Further research in this is important, as research has demonstrated that task difficulty level is related to test anxiety. If it is possible for tasks to be scaled on a difficulty continuum, appropriate tasks may be selected for students of certain predicted ability levels in order to control for test anxiety.

Similarly, research is needed into the type of language and discourse produced by certain tasks. Silverman (1976), MacPhail (1985), van Lier (1989), Perrett (1990) and Lazaraton (1992), among others, have conducted research into the discourse of the Oral Proficiency Interview, but little research has been conducted into the discourse produced by other task types. For example, to my knowledge no similar study exists into the discourse produced by a group oral discussion, let alone a comparative study of the discourse produced on a group task compared with the discourse produced by the same students taking other task types.

Even with regard to the Oral Proficiency Interview (OPI) there is growing evidence from studies of interaction between native and non-native speakers in the Second Language Acquisition literature that knowledge of topic is a determiner of power relations in situations where the native speaker would otherwise be the dominant party by default (Zuengler, 1993). It has often been claimed that in the OPI the learner merely responds to the questions of the interviewer, thus creating 'unnatural conversations'. One way of overcoming this is to vary the task type, but research is needed into the effects of leading the learner (particularly in ESP tests) to believe (probably correctly) that s/he is the subject specialist and that the interviewer is an interested layman. This may alter the nature of the discourse produced, and remove some of the immediate criticisms which have been levelled at the OPI for lack of 'authenticity' (Fulcher, forthcoming).

Little research has been conducted which takes into account task and topic and any possible interaction effects between the two as test method facets.

I would suggest a research project in which one task is designed with one standardised elicitation procedure, but with two or more different topics. A sample of students would be asked to take the two or more tests which would then differ only in terms of topic, with each of the students rated on a single rating scale. It would be hypothesised that any difference in the scores would thus be due to

the effect of topic. Prior to carrying out this study, it would be recommended that each of the students be interviewed about their background interests, learning history, and any other factors which might influence their ability to talk about the topics chosen for the tests. The results of the tests could then be related directly to the background of each student in the sample, making it possible to isolate the nature of any topic effect in oral testing.

3. **Specificity and the length of rating scales**

Writers from the time of Lado (1961) to Baker (1989), Hughes (1989) and Matthews (1990) have argued that rating scales must be as specific and detailed as possible. On the other hand, Porter (1991) argues that they should be short and as simple as possible, to be judged on the basis of what they leave out rather than what they include. Yet, no research has been conducted on the optimum length of rating scales for practical use; the views expressed regarding length are merely the opinions of the writers in the light of their own personal experiences of rating.

Research is therefore needed into the length of rating scales. It is suggested that this could be done by taking one extremely detailed rating scale and producing from it a series of other rating scales along a cline of decreasing specificity by editing out parts of the scale, until it becomes just a title with a list of band numbers. Students would all do the same task, and teams of raters would each independently rate them according to one of the scales. Each team's reliability coefficients and rating patterns would be compared using a Partial Credit model. It would be assumed that somewhere along the cline the evidence would suggest that optimum reliability had been achieved, and thus empirical evidence would be available to suggest that a certain degree of specificity was preferable to others.

4. **Rating scale construction**

In reviews of rating scales, such as the American Council on the Teaching of Foreign Languages (ACTFL) scale, it is frequently suggested that they confound linguistic and nonlinguistic criteria for

measurement. This leads to the confounding of trait with test method, making validation extremely difficult if not impossible (Bachman and Savignon, 1986; Bachman, 1988). Dandonoli and Henning (1990) and Henning (1992) have attempted to show that this is not the case. However, the criticisms have not been answered in full.

If confounding trait and method in rating scale descriptors leads to an inability to establish validity, it may be hypothesised that in a G-Study the Equivalent Forms Generalizability Coefficient would be extremely low, and that in a Rasch Partial Credit analysis the outfit statistics would be fairly high. Proving the opposite to be true, provided that the research design did not bias the outcome, would constitute a reasonable defence for the ACTFL rating scales, but these studies remain to be carried out.

5. **The development and trialling of new methods of investigating validity**

Fulcher (1993:327) suggests a statistic which makes it possible to attach a validity coefficient to a band within a rating scale rather than to the rating scale as a whole. The assumptions behind the development of this statistic (and conditions for its use) are that (a) band descriptors should not refer to test method facets, only traits; (b) the traits which have been operationalised in the band descriptors are stable in that they do not change radically in the learner's competence from moment to moment or day to day - that is, the variable competence model of second language acquisition (Tarone, 1983, 1988, 1990) while supported by some researchers in testing (Skehan, 1987) is rejected; (c) a valid rating scale which measures traits which are generalizable to other contexts (as opposed to performance on one task or test) should provide the same results *across task types* (data collection methods).

The theoretical adjunct of this statistic, and the assumptions relevant to its interpretation, is that a construct should be *coherent*. Its description in the bands of the rating scale should be such that it remains possible to use it as a data collection instrument irrespective

of the context in which this is done. Variation in scores depending on task undertaken by the learners indicates that no underlying construct is being measured, merely performance at a particular moment in time. Whether we call this 'variable capability' or not is irrelevant; in testing terms it means lack of ability to generalize test scores beyond the immediate testing situation.

Such approaches to validity need to be scrutinised both in terms of the theoretical assumptions they make in the light of research being conducted in fields such as Second Language Acquisition, and in terms of their practical benefits in assessing and improving oral rating scales. This statistic has only been applied to three rating scales (Fulcher, 1993), and only 3 bands of one rating scale were discovered to meet the criterion of coherence validity. Results of its use in the study of other rating scales would be necessary before assessing its utility as another tool in the armoury of the testing researcher.

Conclusion

It is clear that the suggestions contained in this article in no way cover all the research which is needed in this field, but it is hoped that one assessment of priority areas may provide others with ideas for their own research projects.

References

Bachman, L.F. 1988. "Problems in Examining the Validity of the ACTFL Oral Proficiency Interview." *Studies in Second Language Acquisition* 10, 2, 149-164.

Bachman, L.F. and Savignon, S.J. 1986. "The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview." *Modern Language Journal* 70, 4, 380-390.

Baker, D. 1989. *Language Testing: A Critical Survey and Practical Guide.* Edward Arnold.

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.

Fulcher, G. 1993. *The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language*. Unpublished PhD thesis: University of Lancaster.

Fulcher, G. (forthcoming) "Testing tasks: issues in task design and the group oral."

Henning, G. 1992. "The ACTFL Oral Proficiency Interview: Validity Evidence." *System* 20, 3, 365-372.

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge University Press.

Lado, R. 1961. *Language Testing*. Longman.

Lazaraton, A. 1992. "The Structural Organization of a Language Interview: A Conversational Analytic Perspective." *System* 20, 3, 373-386.

van Lier, L. 1989. "Reeling, Writhing, Drawling, Stretching and Fainting in Coils: Oral Proficiency Interviews as Conversation." *TESOL Quarterly*, 23, 3, 489-508.

MacPhail, J. 1985. *Oral Assessment Interviews: Suggestions for Participants*. Unpublished MA dissertation, University of Lancaster.

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.

Perrett, G. 1990. "The Language Testing Interview: A Reappraisal." In de Jong, J.H.A.L. and Stevenson, D.K. (eds) *Individualizing the Assessment of Language Abilities*. Philadelphia: Multilingual Matters, 225-238.

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.

Porter, D. 1991. "Sex, Status and Style in the Interview." In Caudery, T. (ed) *New Thinking in TEFL*. Denmark: University of Aarhus Press, 117-128.

Silverman, D. 1976. "Interview talk: bringing off a research instrument." In Silverman, D. and Jones, J. (eds) *Organizational Work: the language of grading, the grading of language*. London: Collier Macmillan.

Skehan, P. 1987. "Variability and Language Testing." In Ellis, R. (ed) *Second Language Acquisition in Context*. Prentice Hall, 195-206.

Stansfield, C. and Kenyon, D.M. 1992. "Comparing the Scaling of Speaking Tasks by Language Teachers and by the ACTFL Guidelines." Paper presented at the 14th annual Language Testing Research Colloquium, Vancouver, BC.

Tarone, E. 1983. "On the variability of interlanguage systems." *Applied Linguistics* 4,2, 146-163.

Tarone, E. 1988. *Variation in Interlanguage*. Edward Arnold.

Tarone, E. 1990. "On Variation in Interlanguage: A Response to Gregg." *Applied Linguistics* 11, 4, 392-400.

Zuengler, J. 1993. "Encouraging Learners' Conversational Participation: The Effect of Content Knowledge." *Language Learning* 43, 3, 403-432.