

Using observation checklists to validate speaking-test tasks

Barry O’Sullivan *The University of Reading*, **Cyril J. Weir** *University of Surrey, Roehampton* and **Nick Saville** *University of Cambridge Local Examinations Syndicate*

Test-task validation has been an important strand in recent revision projects for University of Cambridge Local Examinations Syndicate (UCLES) examinations. This article addresses the relatively neglected area of validating the match between intended and actual test-taker language with respect to a blueprint of language functions representing the construct of spoken language ability. An observation checklist designed for both *a priori* and *a posteriori* analysis of speaking task output has been developed. This checklist enables language samples elicited by the task to be scanned for these functions in real time, without resorting to the laborious and somewhat limited analysis of transcripts. The process and results of its development, implications and further applications are discussed.

I Background to the study

This article reports on the development and use of observation checklists in the validation of the Speaking Tests within the University of Cambridge Local Examinations Syndicate (UCLES) ‘Main Suite’ examination system (see Figure 1). These checklists are intended to

ALTE Level 1	ALTE Level 2	ALTE Level 3	ALTE Level 4	ALTE Level 5
Waystage User	Threshold User	Independent User	Competent User	Good User
CAMBRIDGE Level 1	CAMBRIDGE Level 2	CAMBRIDGE Level 3	CAMBRIDGE Level 4	CAMBRIDGE Level 5
<i>Key English Test (KET)</i>	<i>Preliminary English Test (PET)</i>	<i>First Certificate in English (FCE)</i>	<i>Certificate in Advanced English (CAE)</i>	<i>Certificate of Proficiency in English (CPE)</i>
BASIC	INTERMEDIATE	ADVANCED		

Figure 1 The Cambridge/ALTE five-level system

Address for correspondence: Barry O’Sullivan, Testing and Evaluation Unit, School of Linguistics and Applied Language Studies, The University of Reading, PO Box 241, Whiteknights, Reading RG6 6WB, UK; email: b.e.osullivan@reading.ac.uk

Table 1 Format of the Main Suite Speaking Test

Part	Participants	Task format
1	Interviewer–candidate	Interview: Verbal questions
2	Candidate–candidate	Collaborative task: Visual stimulus; Verbal instructions
3	Interviewer–candidate–candidate	Long turns and discussion: Written stimulus; Verbal questions

provide an effective and efficient tool for investigating variation in language produced by different task types, different tasks within task types, and different interview organization at the proficiency levels in Figure 1. As such, they represent a unique attempt to validate the match between intended and actual test-taker language with respect to a blueprint of language functions representing the construct of spoken language ability in the UCLES tests of general language proficiency, from PET to CPE level (for further information related to the different tests in the ‘Main Suite’ battery, see the individual handbooks produced by UCLES). Beyond this study, the application of such checklists has clear relevance for any test of spoken interaction.

The standard Cambridge approach in testing speaking is based on a paired format involving an interlocutor, an additional examiner and two candidates. Careful attention has been given to the tasks through which the spoken language performance is elicited in each different part. The format of the Main Suite Speaking Tests (with the exception of the Level 1 KET test) is summarized in Table 1.

II Issues in validating tests of oral performance

In considering the issue of the validity of a performance test¹ of speaking, we need a framework that describes the relationship between the construct being measured, the tasks used to operationalize that construct and the assessment of the performances that are used to make inferences to that underlying ability.

There have been a number of models that have attempted to portray the relationship between a test-taker’s knowledge of, and ability to use, a language and the score they receive in a test designed to evaluate that knowledge (e.g., Milanovic and Saville, 1996; McNamara, 1996; Skehan, 1998; Upshur and Turner, 1999).

¹By performance tests we are referring to direct tests where a test-taker’s ability is evaluated from their performance on a set task or tasks.

Milanovic and Saville (1996) provide a useful overview of the variables that interact in performance testing and suggest a conceptual framework for setting out different avenues of research. The framework was influential in the revisions of the Cambridge examinations during the 1990s, including the development of KET and CAE exams and revisions to PET, FCE and, most recently, CPE (for a summary of the UCLES approach, see Saville and Hargreaves, 1999).

The Milanovic and Saville framework is one of the earliest, and most comprehensive of these models (reproduced here as Figure 2). This framework highlights the many factors (or facets) that must be considered when designing a test from which particular inferences are to be drawn about performances; all of the factors represented in the model pose potential threats to the reliability and validity of these inferences. From this model, a framework can be derived, through which a validation strategy can be devised for Speaking Tests such as those produced by UCLES.

The essential elements of this framework are:

- the test-taker;
- the interlocutor/examiner;
- the assessment criteria (scales);
- the task;
- the interactions between these elements.

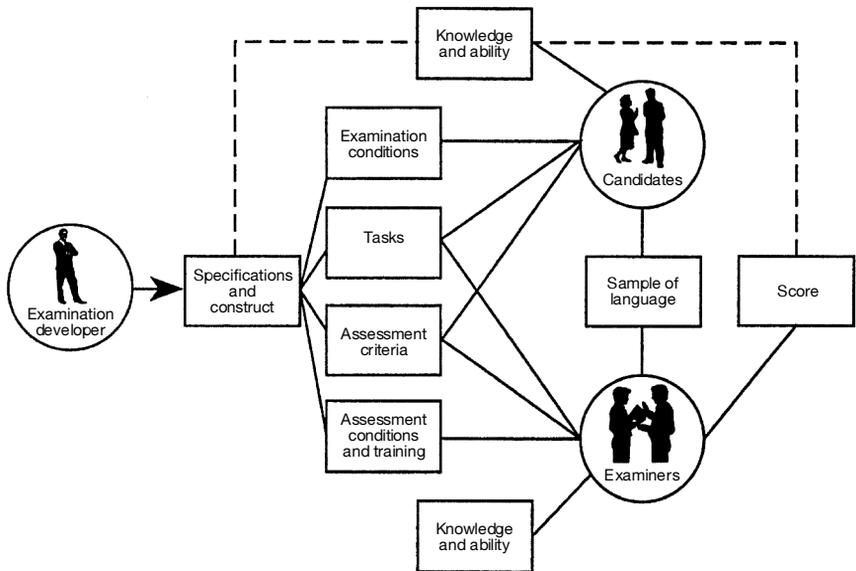


Figure 2 A conceptual framework for performance testing
 Source: adapted from Milanovic and Saville, 1996: 6

The subject of this study, the task, has been explored from a number of perspectives. Briefly, these have been:

- Task/method comparison (quantitative): involving studies in which comparisons are made between performances on different tasks or methods (Clark, 1979; 1988; Henning, 1983; Shohamy, 1983; Shohamy *et al.*, 1986; Clark and Hooshmand, 1992; Stansfield and Kenyon, 1992; Wigglesworth and O'Loughlin, 1993; Chalhoub-Deville, 1995a; O'Loughlin, 1995; Fulcher, 1996; Lumley and O'Sullivan, 2000; O'Sullivan, 2000).
- Task/method comparison (qualitative): as above but where qualitative methods are employed (Shohamy, 1994; Young, 1995; Luoma, 1997; O'Loughlin, 1997; Bygate, 1999; Kormos, 1999).
- Task performance (method effect): where aspects of the task are systematically manipulated; e.g., planning time, pre- or post-task operations, etc. (Foster and Skehan, 1996; 1999; Wigglesworth, 1997; Mehnert, 1998; Ortega, 1999; Upshur and Turner, 1999).
- Native speaker/Nonnative speaker comparison: where native speaker performance on specific tasks is compared to nonnative speaker performance on the same tasks (Weir, 1983; Ballman, 1991).
- Task difficulty/classification: where an attempt has been made to classify tasks in terms of their difficulty (Weir, 1993; Fulcher, 1994; Kenyon, 1995; Robinson, 1995; Skehan, 1996; 1998; Norris *et al.*, 1998).

The central importance of the test task has been clearly recognized; however, in terms of test validation, there is one question that has, to date, remained largely unexplored. Although there has been a great deal of debate over the validation of performance tests through analysis of the language generated in the performance of language elicitation tasks (LETs) (e.g., van Lier, 1989; Lazaraton, 1992; 1996), attention has not been drawn to the one aspect of task performance that would appear to be of most interest to the test designer. That is, when tasks are performed in a test event, how does that performance relate to the test designer's predictions or expectations based on their definition or interpretation of the construct? After all, no matter how reliably the performance is scored, if it does not match the expectations of the test designer (in other words represent the constructs which are to be tested), then the inferences that the test designer hopes to draw from the evaluated performance will not be valid.

Cronbach went to the heart of the matter (1971: 443): 'Construction of a test itself starts from a theory about behaviour or mental organization derived from prior research that suggests the ground plan for the test.' Davies (1977: 63) argued in similar vein: 'it is, after

all, the theory on which all else rests; it is from there that the construct is set up and it is on the construct that validity, of the content and predictive kinds, is based.' Kelly (1978: 8) supported this view, commenting that: 'the systematic development of tests requires some theory, even an informal, inexplicit one, to guide the initial selection of item content and the division of the domain of interest into appropriate sub-areas.'

Because we lack an adequate theory of language in use, *a priori* attempts to determine the construct validity of proficiency tests involve us in matters that relate more evidently to content validity. We need to talk of the communicative construct in descriptive terms and, as a result, we become involved in questions of content relevance and content coverage. Thus, for Kelly (1978: 8) content validity seemed 'an almost completely overlapping concept' with construct validity, and for Moller (1982: 68): 'the distinction between construct and content validity in language testing is not always very marked, particularly for tests of general language proficiency.'

Content validity is considered important as it is principally concerned with the extent to which the selection of test tasks is representative of the larger universe of tasks of which the test is assumed to be a sample (see Bachman and Palmer, 1981; Henning, 1987: 94; Messick, 1989: 16; Bachman, 1990: 244). Similarly, Anastasi (1988: 131) defined content validity as involving: 'essentially the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured.' She outlined (Anastasi, 1988: 132) the following guidelines for establishing content validity:

- 1) 'the behaviour domain to be tested must be systematically analysed to make certain that all major aspects are covered by the test items, and in the correct proportions';
- 2) 'the domain under consideration should be fully described in advance, rather than being defined after the test has been prepared';
- 3) 'content validity depends on the relevance of the individual's test responses to the behaviour area under consideration, rather than on the apparent relevance of item content.'

The directness of fit and adequacy of the test sample is thus dependent on the quality of the description of the target language behaviour being tested. In addition, if the responses to the item are invoked Messick (1975: 961) suggests 'the concern with processes underlying test responses places this approach to content validity squarely in the realm of construct validity'. Davies (1990: 23) similarly notes: 'content validity slides into construct validity'.

Content validation is, of course, extremely problematic given the difficulty we have in characterizing language proficiency with sufficient precision to ensure the validity of the representative sample we include in our tests, and the further threats to validity arising out of any attempts to operationalize real life behaviours in a test. Specifying operations, let alone the conditions under which these are performed, is challenging and at best relatively unsophisticated (see Cronbach, 1990). Weir (1993) provides an introductory attempt to specify the operations and conditions that might form a framework for test task description (see also Bachman, 1990; Bachman and Palmer, 1996).

The difficulties involved do not, however, absolve us from attempting to make our tests as relevant as possible in terms of content. Generating content related evidence is seen as a necessary, although not sufficient, part of the validation process of a speaking test. To this end we sought to establish in this study an effective and efficient procedure for establishing the content validity of speaking tests. As well as being useful in helping specify the domain to be tested we would argue that the checklist discussed below would enable the researcher to address how predicted vs. actual task performance can be compared.

III Methodological issues

While it is relatively easy to rationalize the need to establish that the LETs used in performance tests are working as predicted (i.e., in terms of language generated), the difficulty lies in how this might best be done.

UCLES EFL (English as a foreign language) routinely collects audio recordings and carries out transcriptions of its Speaking Tests. These transcripts are used for a range of validation purposes, and in particular they contribute to revision projects for the Speaking Tests, for example, FCE which was revised in 1996, and currently the revision of the International English Language Testing System (IELTS) Speaking Test, in addition to the CPE revision project.

In a series of UCLES studies focusing on the language of the Speaking Tests, Lazaraton has applied conversational analysis (CA) techniques to contribute to our understanding of the language used in pair-format Speaking Tests, including the language of the candidates and the interlocutor. Her approach requires a very careful, fine-tuned transcription of the tests in order to provide the data for analysis (see Lazaraton, 2000). Similar qualitative methodologies have been applied by Young and Milanovic (1992) – also to UCLES data – by Brown (1998) and by Ross and Berwick (1992), amongst others.

While there is clearly a great deal of potential for this detailed analysis of transcribed performances, there are also a number of drawbacks, the most serious of which involves the complexity of the transcription process. In practice, this means that a great deal of time and expertise is required in order to gain the kind of data that will answer the basic question concerning validity. Even where this is done, it is impractical to attempt to deal with more than a small number of test events; therefore, the generalizability of the results may be questioned.

Clearly then, a more efficient methodology is required that allows the test designer to evaluate the procedures and, especially, the tasks in terms of the language produced by a larger number of candidates. Ideally this should be possible in 'real' time, so that the relationship of predicted outcome to specific outcome can be established using a data set that satisfactorily reflects the typical test-taking population. The primary objective of this project, therefore, was to create an instrument, built on a framework that describes the language of performance in a way that can be readily accessed by evaluators who are familiar with the tests being observed. This work is designed to be complementary to the use of transcriptions and to provide an additional source of validation evidence.

The FCE was chosen as the focus of this study for a number of reasons:

- It is 'stable', in that it is neither under review nor is due to be reviewed.
- It represents the middle of the ALTE (and UCLES Main Suite) range, and is the most widely subscribed test in the battery.
- It offers the most likelihood of a wide range of performance of any Main Suite examination: as it is often used as an 'entry-point' into the suite, candidates tend to range from below to above this level in terms of ability.
- Like all of the other Main Suite examinations, a database of recordings (audio and video) already existed.

IV The development of the observation checklists

Weir (1993), building on the earlier work of Bygate (1988), suggests that the language of a speaking test can be described in terms of the informational and interactional functions and those of interaction management generated by the participants involved. With this as a starting point, a group of researchers at the University of Reading were commissioned by UCLES EFL, to examine the spoken language, second language acquisition and language testing literatures to come up with an initial set of such functions (see Schegloff *et al.*,

1977; Schwartz, 1980; van Ek and Trim, 1984; Bygate, 1988; Shohamy, 1988; 1994; Walker, 1990; Weir, 1994; Stenström, 1994; Chalhoub-Deville, 1995b; Hayashi, 1995; Ellerton, 1997; Suhua, 1998; Kormos, 1999; O’Sullivan, 2000; O’Loughlin, 2001).

These were then presented as a draft set of three checklists (Appendix 1), representing each of the elements of Weir’s categorization. What follows in the three phases of the development process described below (Section VI), was an attempt to customize the checklist to more closely reflect the intended outcomes of spoken language test tasks in the UCLES Main Suite. The checklists were designed to help establish which of these functions resulted, and which were absent.

The next concern was with the development of a procedure for devising a ‘working’ version of the checklists to be followed by an evaluation of using this type of instrument in ‘real’ time (using tapes or perhaps live speaking tests).

V The development model

The process through which the checklists were developed is shown in Figure 3. The concept that drives this model is the evaluation at each level by different stakeholders. At this stage of the project these stakeholders were identified as:

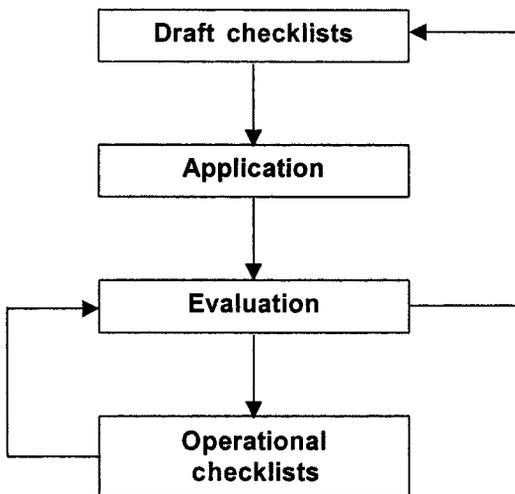


Figure 3 The development model

- the consulting 'expert' testers (the University of Reading group);
- the test development and validation staff at UCLES;
- UCLES Senior Team Leaders (i.e., key staff in the oral examiner training system).

All these individuals participated in the application of each draft. It should also be noted that a number of drafts were anticipated.

VI The development process

In order to arrive at a working version of the checklists, a number of developmental phases were anticipated. At each phase, the latest version (or draft) of the instruments was applied and this application evaluated.

Phase 1

The first attempt to examine how the draft checklists would be viewed, and applied, by a group of language teachers was conducted by French (1999). Of the participants at the seminar, approximately 50% of the group reported that English (British English, American English or Australian English) was their first language, while the remaining 50% were native Greek speakers.

In their introduction to the application of the Observation Checklists (OCs), the participants were given a series of activities that focused on the nature and use of those functions of language seen by task designers at UCLES to be particularly applicable to their EFL Main Suite Speaking Tests (principally FCE, CAE and CPE). Once familiar with the nature of the functions (and where they might occur in a test), the participants applied the OCs in 'real' time to an FCE Speaking Test from the 1998 Standardization Video. This video featured a pair of French speakers who were judged by a panel of 'expert' raters (within UCLES) to be slightly above the criterion ('pass') level.

Of the 37 participants, 32 completed the task successfully, that is, they attempted to make frequency counts of the items represented in the OCs. Among this group, there appear to be varying degrees of agreement as to the use of language functions, particularly in terms of the specific number of observations of each function. However, when the data are examined from the perspective of agreement on whether a particular function was observed or not (ignoring the count, which in retrospect was highly ambitious when we consider the lack of systematic training in the use of the questionnaires given to the teachers who attended), we find that there is a striking degree of

agreement on all but a small number of functions (Appendix 2). Note here that, in order to make these patterns of behaviour clear, the data have been sorted both horizontally and vertically by the total number of observations made by each participant and of each item.

From this perspective, this aspect of the developmental process was considered to be quite successful. However, it was apparent that there were a number of elements within the checklists that were causing some difficulty. These are highlighted in the table by the tram-lines. Items above the lines have been identified by some participants, in one case by a single person, while those below have been observed by a majority of participants (in two cases by all of them). For these cases, we might infer a high degree of agreement. However, the middle range of items appears to have caused a degree of confusion, and so are highlighted here, i.e., marked for further investigation.

Phase 2

In this phase, a much smaller gathering was organized, this time involving members of the development team as well as the three UK-based UCLES Senior Team Leaders. In advance of this meeting all participants were asked to study the existing checklists and to exemplify each function with examples drawn from their experiences of the various UCLES Main Suite examinations. The resulting data were collated and presented as a single document that formed the basis of discussion during a day-long session. Participants were not made aware of the findings from Phase 1.

During this session many questions were asked of all aspects of the checklist, and a more streamlined version of the three sections was suggested. In addition to a number of participants making a written record of the discussions, the entire session was recorded. This proved to be a valuable reminder of the way in which particular changes came about and was used when the final decisions regarding inclusion, conflation or omission were being made. Although it is beyond the scope of this project to analyse this recording, when coupled with the earlier and revised documents, it is in itself a valuable source of data in that it provides a significant record of the developmental process.

Among the many interesting outcomes of this phase were the decisions either to rethink, to reorganize or to omit items from the initial list. These decisions were seen to mirror the results of the Phase 1 application quite closely. Of the 13 items identified in Phase 1 as being in need of review (7 were rarely observed, indicating a high degree of agreement that they were not, in fact, present, and 6 appeared to be confused with very mixed reported observations), 7

were recommended for either omission or inclusion in other items by the panel, while the remaining 6 items were identified by them as being of value. Although no examples of the latter had appeared in the earlier data, the panel agreed that they represented language functions that the UCLES Main Suite examinations were intended to elicit. It was also decided that each item in this latter group was in need of further clarification and/or exemplification. Of the remaining 17 items:

- two were changed: the item 'analysing' was recoded as 'staging' in order to clarify its intended meaning, while it was decided to separate the item '(dis)agreeing' into its two separate components;
- three were omitted: it was argued that the item 'providing non-personal information' referred to what was happening with the other items in the informational function category, while the items 'explaining' and 'justifying/supporting' were not functions usually associated with the UCLES Main Suite tasks and no occurrences of these had been noted.

We would emphasize that, as reported in Section IV above, the initial list was developed to cover the language functions that various spoken language test tasks might elicit. The development of the checklists described here reflects an attempt to customize the lists, in line with the intended functional outcomes of a specific set of tests.

We are, of course, aware that closed instruments of this type may be open to the criticism that valuable information could be lost. However, for reasons of practicality, we felt it necessary to limit the list to what the examinations were intended to elicit, rather than attempt to operationalize a full inventory. Secondly, any functions that appeared in the data that were not covered by the reduced list would have been noted. There appeared to be no cases of this.

The data from these two phases were combined to result in a working version of the checklists (Appendix 3), which was then applied to a pair of FCE Speaking Tests in Phase 3.

Phase 3

In the third phase, the revised checklists were given to a group of 15 MA TEFL students who were asked to apply them to two FCE tests. Both of these tests involved a mixed-sex pair of learners, one pair of approximately average ability and the other pair above average. Before using the observation checklists (OCs), the students were asked first to attempt to predict which functions they might expect to find. To help in this pre-session task, the students were given details of the FCE format and tasks.

Unfortunately, a small number of students did not manage to complete the observation task, as they were somewhat overwhelmed with the real-time application of the checklists. As a result only 12 sets of completed checklists were included in the final analysis.

Prior to the session, the group was given an opportunity to have a practice run using a third FCE examination. While this 'training' period, coupled with the pre-session task, was intended to provide the students with the background they needed to apply the checklists consistently, there was a problem during the session itself. This problem was caused by the failure of a number of students to note the change from Task 3 to Task 4 in the first test observed. This was possibly caused by a lack of awareness of the test itself and was not helped by the seamless way in which the examiner on the video moved from a two-way discussion involving the test-takers to a three-way discussion. This meant that a full set of data exists only for the first two tasks of this test. As the problem was noticed in time, the second test did not cause these problems. Unlike the earlier seminar, on this occasion the participants were asked only to record each function when it was first observed. This was done as it was felt that the earlier seminar showed that, without extensive training, it would be far too difficult to apply the OCs fully in 'real' time in order to generate comprehensive frequency counts. We are aware that a full tally would enable us to draw more precise conclusions about the relative frequency of occurrence of these functions and the degree of consensus (reliability) of observers.

Against this we must emphasize that the checklists, in their current stage of development, are designed to be used in real time. Their use was therefore restricted to determining the presence or absence of a particular function. Rater agreement, in this case, is limited to a somewhat crude account of whether a function occurred or did not occur in a particular task performance. We do not, therefore, have evidence of whether the function observed was invariant across raters.

The results from this session are included as Appendix 4. It can be seen from this table that the participants again display mixed levels of agreement, ranging from a single perceived observation to total agreement. As with the earlier session, it appears that there is relatively broad agreement on a range of functions, but that others appear to be more difficult to identify easily. These difficulties appear to be greatest where the task involves a degree of interaction between the test-takers.

Phase 4

In this phase a transcription was made of the second of the two interviews used in Phase 3, since there was a full set of data available for

this interview. The OCs were then 'mapped' on to this transcript in order to give an overview from a different perspective of what functions were generated (it being felt that this map would result in an accurate description of the test in terms of the items included in the OCs). This mapping was carried out by two researchers, who initially worked independently of each other, but discussed their finished work in order to arrive at a consensus.

Finally the results of Phases 2 and 3 were compared (Appendix 5). This clearly indicates that the checklists are now working well. There are still some problems in items such as 'staging' and 'describing', and feedback from participants suggests that this may be due to misunderstandings or misinterpretations of the gloss and examples used. In addition, there are some similar difficulties with the initial three items in the interactional functions checklist, in which the greatest difficulties in applying the checklists appear to lie.

VII Discussion and initial conclusions

The results of this study appear to substantiate our belief that, although still under development for use with the UCLES Main Suite examinations, an operational version of these checklists is certainly feasible, and has potentially wider application, *mutatis mutandis*, to the content validation of other spoken language tests. Further refinement of the checklists is clearly required, although the developmental process adopted here appears to have borne positive results.

1 Validities

We would not wish to claim that the checklists on their own offer a satisfactory demonstration of the construct validity of a spoken language test, for, as Messick argues (1989: 16): 'the varieties of evidence supporting validity are not alternatives but rather supplements to one another.' We recognize the necessity for a broad view of 'the evidential basis for test interpretation' (Messick, 1989: 20). Bachman (1990: 237) similarly concludes: 'it is important to recognise that none of these [evidences of validity] by itself is sufficient to demonstrate the validity of a particular interpretation or use of test scores' (see also Bachman, 1990: 243). Fulcher (1999: 224) adds a further caveat against an overly narrow interpretation of content validity when he quotes Messick (1989: 41):

the major problem is that so-called content validity is focused upon test forms rather than test scores, upon instruments rather than measurements . . . selecting content is an act of classification, which is in itself a hypothesis that needs to be confirmed empirically.

Like these authors, we regard as inadequate any conceptualization of validity that does not involve the provision of evidence on a number of levels, but would argue strongly that without a clear idea of the match between intended content and actual content, any comprehensive investigation of the construct validity of a test is built on sand. Defining the construct is, in our view, underpinned by establishing the nature of the actual performances elicited by test tasks, i.e. the true content of tasks.

2 Present and future applications of observational checklists

Versions of the checklists require a degree of training and practice similar to that given to raters if a reliable and consistent outcome is to be expected. This requires that standardized training materials be developed alongside the checklists. In the case of these checklists, this process has already begun with the initial versions piloted during Phase 3 of the project.

The checklists have great potential as an evaluative tool and can provide comprehensive insight into various issues. It is hoped that, amongst other issues, the checklists will provide insights into the following:

- the language functions that the different task-types (and different sub-tasks within these) employed in the UCLES Main Suite Paper 5 (Speaking) Tests typically elicit;
- the language that the pair-format elicits, and how it differs in nature and quality from that elicited by interlocutor-single candidate testing;
- the extent to which there is functional variation across the top four levels of the UCLES Main Suite Spoken Language Test.

In addition to these issues, the way in which the checklists can be applied may allow for other important questions to be answered. For example, by allowing the evaluator multiple observations (stopping and starting a recording of a test at will), it will be possible to establish whether there are quantifiable differences in the language functions generated by the different tasks; i.e., the evaluators will have the time they need to make frequency counts of the functions.

While the results to date have focused on *a posteriori* validation procedures, these checklists are also relevant to task design. By taking into account the expected response of a task (and by describing that response in terms of these functions) it will be possible to explore predicted and actual test task outcome. It will also be a useful guide for item writers in taking *a priori* decisions about content coverage. Through this approach it should be possible to predict more accurately

linguistic response (in terms of the elements of the checklists) and to apply this to the design of test tasks – and of course to evaluate the success of the prediction later on. In the longer term this will lead to a greater understanding of how tasks and task formats can be manipulated to result in specific language use. We are not claiming that it is possible to predict language use at a micro level (grammatical form or lexis), but that it is possible to predict informational and interactional functions and features of interaction management – a notion supported by Bygate (1999).

The checklists should also enable us to explore how systematic variation in such areas as interviewer questioning behaviour (and interlocutor frame adherence) affects the language produced in this type of test. In the interview transcribed for this study, for example, the examiner directed his questions very deliberately (systematically aiming the questions at one participant and then the other). This tended to stifle any spontaneity in the intended three-way discussion (Task 4), so occurrences of Interactional and Discourse Management Functions did not materialize to the extent intended by the task designers. It is possible that a less deliberate (unscripted) questioning technique would lead to a less interviewer-oriented interaction pattern and allow for the more genuine interactive communication envisaged in the task design.

Perhaps the most valuable contribution that this type of validation procedure offers is its potential to improve the quality of oral assessment in both low-stakes and high-stakes contexts. By offering the investigator an instrument that can be used in real time, the checklists broaden the scope of investigation from limited case study analysis of small numbers of test transcripts to large scale field studies across a wide range of testing contexts.

Acknowledgements

We would like to thank Don Porter and Rita Green for their early input into the first version of the checklist. In addition, help was received from members of the ELT division in UCLES, in particular from Angela French, Lynda Taylor and Christina Rimini, from a group of UCLES Senior Team Leaders and from MA TEFL students at the University of Reading. Finally, we would like to thank the editors and anonymous reviewers of *Language Testing* for their insightful comments and helpful suggestions for its improvement. The faults that remain are, as ever, ours.

VIII References

- Anastasi, A.** 1988: *Psychological testing*. 6th edition. New York: Macmillan.
- Bachman, L.F.** 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L.F. and Palmer, A.S.** 1981: The construct validation of the FSI oral interview. *Language Learning* 31, 67–86.
- 1996: *Language testing in practice*. Oxford: Oxford University Press.
- Ballman, T.L.** 1991: The oral task of picture description: similarities and differences in native and nonnative speakers of Spanish. In Teschner, R.V., editor, *Assessing foreign language proficiency of undergraduates*. AAUSC Issues in Language Program Direction. Boston: Heinle and Heinle, 221–31.
- Brown, A.** 1998: *Interviewer style and candidate performance in the IELST oral interview*. Paper presented at the Language Testing Research Colloquium, Monterey, CA.
- Bygate, M.** 1988: *Speaking*. Oxford: Oxford University Press.
- 1999: Quality of language and purpose of task: patterns of learners' language on two oral communication tasks. *Language Teaching Research* 3, 185–214.
- Chalhoub-Deville, M.** 1995a: Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12, 16–33.
- 1995b: A contextualized approach to describing oral language proficiency. *Language Learning* 45, 251–81.
- Clark, J.L.D.** 1979: Direct vs. semi-direct tests of speaking ability. In Briere, E.J. and Hinofotis, F.B., editors, *Concepts in language testing: some recent studies*. Washington DC: TESOL.
- 1988: Validation of a tape-mediated ACTFL/ILR scale based test of Chinese speaking proficiency. *Language Testing* 5, 187–205.
- Clark, J.L.D. and Hooshmand, D.** 1992: 'Screen to Screen' testing: an exploratory study of oral proficiency interviewing using video teleconferencing. *System* 20, 293–304.
- Cronbach, L.J.** 1971: Validity. In Thorndike, R.L., editor, *Educational measurement*. 2nd edition. Washington DC: American Council on Education, 443–597.
- 1990: *Essentials of psychological testing*. 5th edition. New York: Harper & Row.
- Davies, A.** 1977: The construction of language tests. In Allen, J.P.B. and Davies, A., editors, *Testing and experimental methods. The Edinburgh Course in Applied Linguistics*, Volume 4. London: Oxford University Press, 38–194.
- 1990: *Principles of language testing*. Oxford: Blackwell.
- Ellerton, A.W.** 1997: Considerations in the validation of semi-direct oral testing. Unpublished PhD thesis, CALS, University of Reading.
- French, A.** 1999: *Language functions and UCLES speaking tests*. Seminar in Athens, Greece. October 1999.

- Foster, P.** and **Skehan, P.** 1996: The influence of planning and task type on second language performance. *Studies in Second Language Acquisition* 18, 299–323.
- 1999: The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research* 3, 215–47.
- Fulcher, G.** 1994: Some priority areas for oral language testing. *Language Testing Update* 15, 39–47.
- 1996: Testing tasks: issues in task design and the group oral. *Language Testing* 13, 23–51.
- 1999: Assessment in English for academic purposes: putting content validity in its place. *Applied Linguistics* 20, 221–36.
- Hayashi, M.** 1995: Conversational repair: a contrastive study of Japanese and English. MA Project Report, University of Canberra.
- Henning, G.** 1983: Oral proficiency testing: comparative validities of interview, imitation, and completion methods. *Language Learning* 33, 315–32.
- 1987: *A guide to language testing*. Cambridge, MA: Newbury House.
- Kelly, R.** 1978: On the construct validation of comprehension tests: an exercise in applied linguistics. Unpublished PhD thesis, University of Queensland.
- Kenyon, D.** 1995: An investigation of the validity of task demands on performance-based tests of oral proficiency. In Kunnan, A.J., editor, *Validation in language assessment: selected papers from the 17th Language Testing Research Colloquium, Long Beach*. Mahwah, NJ: Lawrence Erlbaum, 19–40.
- Kormos, J.** 1999: Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing* 16, 163–88.
- Lazaraton, A.** 1992: The structural organisation of a language interview: a conversational analytic perspective. *System* 20, 373–86.
- 1996: A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE). In Milanovic, M. and Saville, N., editors, *Performance testing, cognition and assessment: selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*. Studies in Language Testing 3. Cambridge: University of Cambridge Local Examinations Syndicate, 18–33.
- 2000: *A qualitative approach to the validation of oral language tests*. Studies in Language Testing, Volume 14. Cambridge: Cambridge University Press.
- Lumley, T.** and **O'Sullivan, B.** 2000: The effect of speaker and topic variables on task performance in a tape-mediated assessment of speaking. Paper presented at the 2nd Annual Asian Language Assessment Research Forum, The Hong Kong Polytechnic University.
- Luoma, S.** 1997: Comparability of a tape-mediated and a face-to-face test of speaking: a triangulation study. Unpublished Licentiate Thesis, Centre for Applied Language Studies, Jyväskylä University, Finland.

- McNamara, T.** 1996: *Measuring second language performance*. London: Longman.
- Mehnert, U.** 1998: The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition* 20, 83–108.
- Messick, S.** 1975: The standard problem: meaning and values in measurement and evaluation. *American Psychologist* 30, 955–66.
- 1989: Validity. In Linn, R.L., editor, *Educational measurement*. 3rd edition. New York: Macmillan.
- Milanovic, M. and Saville, N.** 1996: Introduction. *Performance testing, cognition and assessment*. Studies in Language Testing, Volume 3. Cambridge: University of Cambridge Local Examinations Syndicate, 1–17.
- Moller, A. D.** 1982: A study in the validation of proficiency tests of English as a Foreign Language. Unpublished PhD thesis, University of Edinburgh.
- Norris, J. D., Brown, J. D., Hudson, T. and Yoshioka, J.** 1998: *Designing second language performance assessments*. Technical Report 18. Honolulu, HI: University of Hawaii Press.
- O'Loughlin, K.** 1995: Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing* 12, 217–37.
- 1997: The comparability of direct and semi-direct speaking tests: a case study. Unpublished PhD Thesis, University of Melbourne, Melbourne.
- 2001: *An investigatory study of the equivalence of direct and semi-direct speaking skills*. Studies in Language Testing 13. Cambridge: Cambridge University Press/UCLES.
- Ortega, L.** 1999: Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition* 20, 109–48.
- O'Sullivan, B.** 2000: Towards a model of performance in oral language testing. Unpublished PhD dissertation, CALS, University of Reading.
- Robinson, P.** 1995: Task complexity and second language narrative discourse. *Language Learning* 45, 99–140.
- Ross, S. and Berwick, R.** 1992: The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14, 159–76.
- Saville, N. and Hargreaves, P.** 1999: Assessing speaking in the revised FCE. *ELT Journal* 53, 42–51.
- Schegloff, E., Jefferson, G. and Sachs, H.** 1977: The preference for self-correction in the organisation of repair in conversation. *Language* 53, 361–82.
- Schwartz, J.** 1980: The negotiation for meaning: repair in conversations between second language learners of English. In Larsen-Freeman, D., editor, *Discourse analysis in second language research*. Rowley, MA: Newbury House.
- Shohamy, E.** 1983: The stability of oral language proficiency assessment in the oral interview testing procedure. *Language Learning* 33, 527–40.
- 1988: A proposed framework for testing the oral language of

- second/foreign language learners. *Studies in Second Language Acquisition* 10, 165–79.
- 1994: The validity of direct versus semi-direct oral tests. *Language Testing* 11, 99–123.
- Shohamy, E., Reves, T. and Bejarano, Y.** 1986: Introducing a new comprehensive test of oral proficiency. *ELT Journal* 40, 212–20.
- Skehan, P.** 1996: A framework for the implementation of task based instruction. *Applied Linguistics* 17, 38–62.
- 1998: *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Stansfield, C.W. and Kenyon, D.M.** 1992: Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System* 20, 347–64.
- Stenström, A.** 1994: *An introduction to spoken interaction*. London: Longman.
- Suhua H.** 1998: A communicative test of spoken English for the CET 6. Unpublished PhD Thesis, Shanghai Jiao Tong University, Shanghai.
- Upshur, J.A. and Turner, C.** 1999: Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing* 16, 82–111.
- van Ek, J.A. and Trim J.L.M.,** editors, 1984: *Across the threshold*. Oxford: Pergamon.
- van Lier, L.** 1989: Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly*, 23, 489–508.
- Walker, C.** 1990: Large-scale oral testing. *Applied Linguistics* 11, 200–19.
- Weir, C.J.** 1983: Identifying the language needs of overseas students in tertiary education in the United Kingdom. Unpublished PhD thesis, University of London.
- 1993: *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall.
- Wigglesworth, G.** 1997: An investigation of planning time and proficiency level on oral test discourse. *Language Testing* 14, 85–106.
- Wigglesworth, G. and O'Loughlin, K.** 1993: An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English. *Melbourne Papers in Language Testing* 2, 56–67.
- Young, R.** 1995: Conversational styles in language proficiency interviews. *Language Learning* 45, 3–42.
- Young, R. and Milanovic, M.** 1992: Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition* 14, 403–24.

Appendix 1 Items included in initial draft checklists (with short gloss)

Informational functions

Providing personal information	<ul style="list-style-type: none"> • Give information on present circumstances? • Give information on past experiences? • Give information on future plans?
Providing nonpersonal information	Give information which does not relate to the individual?
Elaborating	Elaborate on an idea?
Expressing opinions	Express opinions?
Justifying opinions	Express reasons for assertions s/he has made?
Comparing	Compare things/people/events?
Complaining	Complain about something?
Speculating	Hypothesize or speculate?
Analysing	Separate out the parts of an issue?
Making excuses	Make excuses?
Explaining	Explain anything?
Narrating	Describe a sequence of events?
Paraphrasing	Paraphrase something?
Summarizing	Summarize what s/he had said?
Suggesting	Suggest a particular idea?
Expressing preferences	Express preferences?

Interactional functions

Challenging	Challenge assertions made by another speaker?
(Dis)agreeing	Indicate (dis)agreement with what another speaker says? (apart from 'yeah'/'no' or simply nodding)
Justifying/Providing support	Offer justification or support for a comment made by another speaker?
Qualifying	Modify arguments or comments?
Asking for opinions	Ask for opinions?
Persuading	Attempt to persuade another person?
Asking for information	Ask for information?
Conversational repair	Repair breakdowns in interaction?
Negotiating meaning	<ul style="list-style-type: none"> • Check understanding? • Attempt to establish common ground or strategy? • Respond to requests for clarification? • Ask for clarification? • Make corrections? • Indicate purpose? • Indicate understanding/uncertainty?

Managing interaction

Initiating	Start any interactions?
Changing	Take the opportunity to change the topic?
Reciprocity	Share the responsibility for developing the interaction?
Deciding	Come to a decision?
Terminating	Decide when the discussion should stop?

Appendix 2 Phase 1 results (summarized)

	Participants
Make excuses	
Terminate	
Conversational repair	
Summarize	
Complain	
Paraphrase	
Persuade	
Change topic	
Challenge	
Qualify	
Ask for info	
Suggest	
Narrate	
Reciprocate	
Analyse	
Elaborate	
Initiate	
Provide nonpersonal information	
Explain	
Justify opinions	
Negotiate meaning	
Decide	
(Dis) agree	
Justify/Support	
Ask for opinions	
Express preferences	
Speculate	
Compare	
Provide nonpersonal information	
Express opinion	

Appendix 3 Operational checklist (used in Phase 3)

Informational functions

Providing personal information	<ul style="list-style-type: none"> • Give information on present circumstances • Give information on past experiences • Give information on future plans
Expressing opinions	Express opinions
Elaborating	Elaborate on, or modify an opinion
Justifying opinions	Express reasons for assertions s/he had made
Comparing	Compare things/people/events
Speculating	Speculate
Staging	Separate out or interpret the parts of an issue
Describing	<ul style="list-style-type: none"> • Describe a sequence of events • Describe a scene
Summarizing	Summarize what s/he has said
Suggesting	Suggest a particular idea
Expressing preferences	Express preferences

Interactional functions

Agreeing	Agree with an assertion made by another speaker (apart from 'yeah' or nonverbal)
Disagreeing	Disagree with what another speaker says (apart from 'no' or nonverbal)
Modifying	Modify arguments or comments made by other speaker or by the test-taker in response to another speaker
Asking for opinions	Ask for opinions
Persuading	Attempt to persuade another person
Asking for information	Ask for information
Conversational repair	Repair breakdowns in interaction
Negotiating meaning	<ul style="list-style-type: none"> • Check understanding • Indicate understanding of point made by partner • Establish common ground/purpose or strategy • Ask for clarification when an utterance is misheard or misinterpreted • Correct an utterance made by other speaker which is perceived to be incorrect or inaccurate • Respond to requests for clarification

Managing interaction

Initiating	Start any interactions
Changing	Take the opportunity to change the topic
Reciprocating	Share the responsibility for developing the interaction
Deciding	Come to a decision

Appendix 4 Summary of Phase 2 observation

	Tape 1				Tape 2			
	Task 1	Task 2	Task 3	Task 4	Task 1	Task 2	Task 3	Task 4
<i>Informational functions</i>								
Providing personal information								
Present	12 (G)	1 (L)	1 (L)	1 (L)	12 (G)		1 (L)	4 (L)
Past	10 (G)			4 (S)	12 (G)			
Future	11 (G)		3 (L)	6 (S)	12 (G)			
Expressing opinions	12 (G)	11 (G)	9 (G)	8 (G)	11 (G)	10 (G)	10 (G)	11 (G)
Elaborating	9 (G)	11 (G)	9 (G)	7 (G)	3 (L)	9 (G)	7 (S)	12 (G)
Justifying opinions	10 (G)	7 (G)	9 (G)	7 (G)	4 (L)	8 (S)	6 (S)	8 (S)
Comparing	11 (G)	8 (G)	1 (L)	6 (S)	3 (L)	12 (G)	7 (S)	5 (S)
Speculating	7 (S)	11 (G)	8 (G)	3 (L)	7 (S)	10 (G)	10 (G)	5 (S)
Staging		6 (S)	1 (L)		3 (L)	6 (L)		
Describing								
Sequence of events	1 (L)	1 (L)			3 (L)	1 (L)		4 (L)
Scene	5 (S)	9 (G)	2 (S)	2 (S)		10 (G)	2 (S)	3 (S)
Summarizing	1 (L)	1 (L)	1 (L)	1 (L)	3 (L)	1 (L)	1 (L)	1 (L)
Suggesting	1 (L)	2 (L)	1 (L)		3 (L)		2 (L)	
Expressing preferences	12 (G)	11 (G)	6 (S)	8 (G)	11 (G)	10 (G)	5 (S)	12 (G)
<i>Interactional functions</i>								
Agreeing	6 (S)		9 (G)	2 (L)			10 (G)	4 (L)
Disagreeing			9 (G)	4 (S)			2 (L)	6 (S)
Modifying		1 (L)	5 (S)	4 (S)			7 (S)	1 (L)
Asking for opinions	1 (L)		8 (G)	2 (L)			11 (G)	
Persuading			2 (L)				2 (L)	
Asking for information			2 (L)	1 (L)			5 (S)	
Conversational repair						5 (S)	4 (L)	1 (L)
Negotiating meaning								
Check meaning			2 (L)	4 (S)				4 (L)
Understanding			5 (S)				3 (L)	3 (L)
Common ground			2 (L)				2 (L)	1 (L)
Ask clarification			2 (L)				1 (L)	2 (L)
Correct utterance			3 (L)			1 (L)		
Respond to required clarification	4 (S)						1 (L)	
<i>Managing interaction</i>								
Initiating			8 (G)	1 (L)			10 (G)	5 (S)
Changing			8 (G)				7 (S)	
Reciprocating			7 (G)				9 (G)	1 (L)
Deciding			3 (L)	1 (L)		1 (L)	2 (L)	

Notes: The figures indicate the number of students that complete the task in each case. L: Little agreement; S: Some agreement; G: Good agreement. For Tasks 3 and 4 in the first tape observed, the maximum was 9; for all others the maximum was 12. This is because 3 of the 12 MA students did not complete the task for these last 2 tasks. This was not a problem during the observation of the second tape, so for all the maximum figures are 12.

Appendix 5 Transcript results and observation checklist results

<i>Informational functions</i>	Task 1	Task 2	Task 3	Task 4
<i>Providing personal information</i>				
Present	T G		L	T L
Past	T G			
Future	T G			
Expressing opinions	T G	T G	T G	T G
Elaborating	L	T G	T S	T G
Justifying opinions	L	T S	T S	T S
Comparing	L	T G	T S	S
Speculating	T S	T G	T G	S
Staging	T L	T S		
<i>Describing</i>				
Sequence of events	T L	L		L
Scene		T G	L	L
Summarizing	T L	L	L	L
Suggesting	L		L	
Expressing preferences	T G	T G	S	T G
<i>Interactional functions</i>				
Agreeing			T G	T L
Disagreeing			T	S
Modifying			T S	T L
Asking for opinions			T G	
Persuading			L	
Asking for information			S	
Conversational repair		T S	T L	L
<i>Negotiating meaning</i>				
Check meaning				L
Understanding			L	L
Common ground			L	L
Ask clarification			L	T L
Correct utterance		L		
Respond to required clarification			L	
<i>Managing interaction</i>				
Initiating			T G	T S
Changing			T S	
Reciprocating			T G	L
Deciding		L	L	

Notes: T indicates that this function has been identified as occurring in the transcript of the interaction. L, S and G indicate the degree of agreement among the raters using the checklists in real time (L: Little agreement; S: Some agreement; G: Good agreement).

Copyright of Language Testing is the property of Arnold Publishers and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.