DOCUMENT RESUME

ED 386 045 FL 023 208

AUTHOR Brindley, Geoff
TITLE Task-Centred Assessment in Language Learning: The Promise and the Challenge.
PUB DATE 94
NOTE 24p.; In: Bird, Norman, Ed., And Others. Language and Learning. Papers presented at the Annual International Language in Education Conference (Hong Kong, 1993); see FL 023 205.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Competency Based Education; Evaluation Criteria; Foreign Countries; Language Teachers; *Language Tests; Second Language Instruction; Second Language Learning; *Second Languages; *Task Analysis; Teacher Education; *Test Construction; Test Reliability; *Test Validity

ABSTRACT

It is argued that while task-centered assessment of second language learning has the strong support of teachers and learners, focuses on language as a tool rather than as an end in itself, and fosters learning, there remain some problems with its use. Further work must be done to develop assessment criteria that reflect current theories of language learning and use. Consistency in application of these criteria must then be assured. On a practical level, adoption of task-centered assessment will require that teachers and learners become accustomed to thinking of language tasks not only as activities but also as indicators of progress and achievement, and learners will need to understand the criteria for evaluation of performance. In turn, this will require closer examination of the components of language tasks and a raising of learners' awareness of how language functions to achieve particular communicative purposes. In addition, reporting of task performance is complex, qualitative, and multidimensional rather than standardized and uniform. Contains 56 references. (MSE)

TASK-CENTRED ASSESSMENT IN LANGUAGE LEARNING: THE PROMISE AND THE CHALLENGE

GEOFF BRINDLEY

TASK-CENTRED ASSESSMENT IN LANGUAGE LEARNING: THE PROMISE AND THE CHALLENGE

Geoff Brindley

1. Introduction

The advent of task-centred language teaching has brought with it various forms of assessment which are aimed at providing information on how well learners are able to mobilise language to achieve meaningful communicative goals. As Mendelsohn (1989) states:

> I believe that the goal of testing today ... is to see what someone **can do** with the language.

This type of 'can-do' assessment has a number of positive features:

- Teachers' and learners' attention becomes more focused on language as a tool for communication rather than on language knowledge as an end in itself (Shohamy 1992).
- Assessment is integrated into the learning process through the use of attainment targets which are directly linked to course content and objectives (Griffin and McKay 1992).
- Learners are able to obtain useful diagnostic feedback on their progress and achievement since explicit performance criteria are provided against which they can compare their performances. This fosters collaborative learning and encourages self-assessment (Brindley 1989).
- Better communication between users of assessment information and educational institutions can be established through the use of various forms of outcome reporting which are couched in performance terms and are hence intelligible to non-specialists (Griffin and Nix 1991).

However, despite these attractive features of task-centred assessment (TCA), it is not without its problems. Concerns have been have expressed regarding the validity of some forms of TCA (Bachman 1990), the feasibility of achieving reliability (Swain 1993), and the practical constraints, particularly the financial costs, involved in its implementation (Shohamy 1993). Some applied linguists have argued that a lot more research into the theoretical foundations of TCA is needed before this kind of assessment can be automatically endorsed as a viable alternative to more traditional forms of assessment (McNamara 1990).

In this paper I want to examine some of the issues and problems which have been raised in relation to TCA in the context of language learning. I will deal with these under the broad areas briefly mentioned above: *validity*, *reliability* and *practicality*. In the first part I shall examine some of the controversies surrounding the notion of authenticity as it relates to the validity of TCA. I will then discuss the important question of the construct validity of TCA by looking at various ways in which the criteria for judging task performance have been established and then suggest ways in which these criteria might be made more accountable to empirical data derived from studies of language acquisition and use. Since the success of TCA depends very much on expert human judgement, the second part of the paper will focus on some of problems which have been encountered in attempting to ensure reliability in TCA and look at the potential contribution of recent advances in measurement technology to addressing these problems. Finally, since TCA can only work 'on the ground' if the conditions exist for its implementation at a system level; in the final part of the paper I shall explore some of the practical ramifications of introducing TCA systems into educational institutions.

2. Defining Task-Centred Assessment

As various authors have pointed out, the term 'task' is used in a variety of ways in the language learning literature, ranging from very broad definitions that accentuate the 'real-world' nature of tasks:

> ... the hundred and one things people *do* in everyday life, at work, at play, and in between. Tasks are the things people will tell you to do if you ask them and they are not applied linguists (Long 1985:89).

to those that focus primarily on the role of language in the classroom:

> A range of work plans which have the overall purpose of facilitating language learning-from the simple and brief exercise type to more complex and lengthy activities such as group problem-solving or simulations and decision-making (Breen 1987:23).

For the purposes of this discussion I will propose the following definition of task-centred assessment which is sufficiently general to cover both in-class and out-of-class situations and can thus be applied to the assessment of proficiency acquired independently of the curriculum or to curriculum-based achievement:

> Task-centred language assessment is the process of evaluating, in relation to a set of explicitly stated criteria, the quality of the communicative performances elicited from learners as part of goal-directed, meaning-focused language use requiring the integration of skills and knowledge.

This definition draws on definitions of task-centred language learning as enunciated by, *inter alia*, Crookes (1986), Nunan (1989) and Swales (1990), and attempts to incorporate some key notions of criterion-referenced communicative assessment, namely:

- the need for explicitness in stating the criteria according to which learners' performances are to be judged (Brindley 1991).
- the centrality of communicative goals as a starting point in syllabus design and assessment (Nunan 1993).
- the view of language proficiency as encompassing both knowledge and ability for use (Bachman 1990).

TCA, of course, is not new in language learning. As far as proficiency assessment is concerned, the oral interview and various kinds of writing tasks have been a standard part of tests of communicative language proficiency for some time. And with the advent of the task-based syllabus, assessments and diagnostic feedback based on teachers' observations of task performance have become a feature of classroom practice (Brindley 1989). What is relatively new, however, is the weight which is now starting to be attached to assessment of learners' ongoing task performance as a factor in final assessment for purposes of certification. Another comparatively recent change in the assessment landscape is the degree of explicitness and rigour with which teachers are now being required to document and justify their assessments (Barrs 1992). What was once informal and formative, in other words, is becoming a high-stakes business. For this reason, TCA is being increasingly called to demonstrate its validity, reliability and practicality in much the same way as standardized pencil-and-paper tests have been in the past, and it is to the first of these issues that I now turn.

3. Validity Issues in Task-Centred Assessment

3.1 Validity and authenticity

One of the most attractive aspects of adopting the communicative task as a unit of teaching and assessment in language learning is that it enables the teacher to focus on communicative activities resembling the authentic use of language such as listening to news broadcasts and picking out the main points, reading a TV guide to find out what's on, defending a point of view, writing letters to a pen-friend in a foreign country, etc. These tasks can readily be turned into assessment activities as long as they are accompanied by a set of assessment criteria which describe what the learner must do in order to demonstrate that he or she is able to perform the task successfully.

It might seem reasonable enough to assume that assessments based on communicative tasks such as these are valid by definition since they attempt to replicate 'real life' language use situations, which is ultimately what communicative language teaching and assessment are concerned with. However such assumptions have been questioned by many writers in both general education and language assessment on a number of grounds. In the first place, an assessment activity is by its very nature an artificial situation: no matter how 'life-like' the task is, people still know they are being assessed under special conditions. As Spolsky (1985:36) comments:

> ... We are forced to the conclusion that testing is not authentic language behaviour, that examination questions are not real, however much like real-life questions they seem, and that an examinee needs to learn the special rules of examinations before he or she can take part in them successfully.

A second problem with 'authentic' assessment tasks is the difficulty of generalizing from a one-off performance to other situations of language use. Commenting on the tendency of some language testers to claim validity on the basis that their tests reflect real-life settings, Skehan (1984:208) comments:

> This viewpoint confuses naturalness of setting with sufficiency. A large part of the problem in testing is in sampling a sufficiently wide range of language to be able to generalize to new situations. Merely making an interaction 'authentic' does not guarantee that the sampling of language involved will be sufficient, or the basis for wide-ranging and powerful predictions of language behaviour in other situations.

In a similar vein, Bachman (1990) has cautioned against the acceptance of authentic-looking 'direct' tests such as the oral interview as automatically valid measures of ability. He points out (1990:309) that such tests confuse the observation of a performance with the ability itself and are limited in their generalizability beyond the specific context in which testing takes place. Bachman proposes a somewhat different approach to authenticity by suggesting that authenticity lies not only in the surface resemblance between assessment tasks and real-world behaviour but also in the extent to which different areas of language skills and knowledge are sampled in the task. This is what he refers to as *interactional authenticity*:

> In summary the IA (interaction/ability) approach views authenticity as residing in the interaction between the test taker, the test task and the testing context. (Bachman 1990:317)

In order to construct valid 'authentic' tests of communicative language ability Bachman argues that we have to construct or select tests or test tasks that reflect

our knowledge of the nature of language abilities and language use. Both the development and selection of authentic language tests is thus based on a theoretical framework that includes the language abilities of the test taker and the characteristics of the testing context.

Whether or not the framework one adopts is that proposed by Bachman (1990), or some other model, the implications for developers of tests and assessment tasks are clear: some kind of conceptualization of communicative language ability is needed which can serve as a starting point for deciding which abilities to sample and which test methods are most appropriate to tap these abilities. Thus in an oral test we would need to start with an idea of what we understand by 'speaking ability', what its components are, and which of these components we want to tap, drawing on the best of our knowledge of the nature of the ability or abilities in question. This would lead to a kind of sampling frame which enabled us to see which ability components were being sampled by which tasks using which methods. Shohamy (1993) provides an example of how this might be done in the development of a testing program used to provide diagnostic feedback to learners in Hebrew programs in the United States and Canada. Each of the sub-tests is based on current theories and understandings of the skill concerned and encompasses a wide range of tasks, text-types and item formats. Another example of this broad sampling approach is *Access:* a recently developed Australian Government English language proficiency test for prospective immigrants in which the Oral Interaction sub-test incorporates a variety of task types, topics and language functions (Wigglesworth and O'Loughlin 1993).

Summing up this discussion on authenticity, then, I would argue, along with Bachman and others, that authenticity in the sense of surface resemblance to target language use situations is a necessary but not sufficient condition for test and task validity. In addition, we need a principled way of specifying the abilities that assessment tasks are tapping and a sampling frame which enables us to obtain a complex and multidimensional picture of the way in which these abilities are being assessed through a range of different tasks using different assessment methods. One of the advantages of this multidimensional approach in the context of the classroom is that it naturally lends itself to profiling approaches which enable teachers to build up samples of different types of student work which reflect progress over a period of time.

3.2 Defining assessment criteria

The use of a wide variety of task-types places the onus on test developers (or teachers if they are responsible for assessment) to specify the characteristics of tasks with sufficient precision that they can be assessed. This means not only describing the tasks but also identifying a set of key criteria according to which learners' performance can be rated or scored. A range of different approaches has been adopted to task specification and identification of rating criteria. In this section I want to briefly discuss some of these approaches and to suggest ways in which

data from studies of second language acquisition and use might be drawn on to inform task descriptors and rating criteria.

3.2.1 'Expert judgement' approaches

One commonly used way of producing criteria for proficiency testing is to ask expert judges to identify and sometimes to weight the key features of learner performance which are to be assessed. Experienced teachers tend to be the audience most frequently consulted in the development and refinement of criteria and performance descriptions (eg. Griffin and Nix 1991; Griffin and McKay 1992). In some cases they may be asked to generate the descriptors themselves by describing key indicators of performance at different levels of proficiency. In others, test developers may solicit comments and suggestions from teachers for modification of existing descriptors on the basis of their knowledge and experience.

The idea of using teachers' expert judgement appeals to logic and common sense. However it also brings with it certain difficulties. The first of these is that teachers' observations of language are bound to be influenced by the personalized constructs of language ability with which they operate. This is recognized by Griffin and McKay (1992:20), who adopted what they refer to as a 'bottom up consultative approach' to develop scales of ESL development for primary and secondary education in Australia and who write that:

> Limitations of this approach include the difficulties involved in obtaining appropriate descriptions of language behaviour from practitioners. It is often the case that practitioners' observations are limited by a lack of knowledge of theoretical models, by inadequate observation skills and/or an inability to articulate descriptions of independent student language behaviour. The developer of the scales has to make decisions about the need to use the imprecise language of the practitioner, and perhaps lose some of the definitive nature of the theoretical model, or to use a specialist terminology and run the risk of practitioner misinterpretation and rejection.

If expert opinion is to have any currency as a method of developing criteria, then one would expect that a given group of expert judges would concur, first on the criteria which make up the behavioral domain being assessed and second, on the allocation of particular performance features to particular levels. (Obtaining data in this way would be an integral part of construct validation). One would also expect that the group would be able to agree on the extent to which a test item was testing a particular skill and the level of difficulty represented by the item. (Agreement would constitute evidence for content validity).

Studies aimed at investigating how expert judgements are made, however, cast some doubt on the ability of expert judges to agree on any of these issues. Alderson and Lukmani (1989), for example, in an examination of item content in EFL

reading tests, found that judges were unable to agree not only on what particular items were testing but also on the level of difficulty of items or skills and the assignment of these to a particular level. Devenney (1989) who investigated the evaluative judgements of ESL teachers and students of ESL compositions, found both within-group and between-group differences in the criteria which were used. He comments:

> Implicit in the notion of interpretive communities are these assumptions: (1) a clear set of shared evaluative criteria exists, and (2) it will be used by members of the interpretive community to respond to text. Yet this did not prove to be the case for either ESL teachers or students

On the basis of findings such as these it would clearly not be advisable to rely solely on teachers' expert judgement as a basis for determining assessment criteria. This is not to suggest that teachers do not have a role in identifying behavioral indicators and tasks that will be used to assess their students. However, as Griffin and McKay (1992:21) note, the data they provide needs to be cross-checked against theoretical research and other published data.

3.2.2 Rating scales

Another - and possibly the easiest - way to define criteria and descriptors for language assessment is to use those already in existence. There is no shortage of models and examples. Literally thousands of rating scales, band scales and performance descriptors are used throughout the world to describe aspects of language performance in a global way. These are frequently used as generalized criteria against which task performance can be rated.

A number of objections, have been raised, however, to some of the more commonly used proficiency rating scales, such as the ACTFL scale, used to certify foreign language teachers in the United States. These are discussed in detail by Brindley (1991) and North (1993) and will not be reiterated in detail here. However, in the context of TCA some of the more pertinent objections could be summarized as follows:

- The scales are not based on studies of second language use and as such, have no empirical support (Lantolf and Frawley 1988).
- The logic of the way levels are arrived at is essentially circular-'the criteria are the levels and vice-versa' (Lantolf and Frawley 1985:340). They cannot therefore be criterion-referenced in the accepted sense since there is no external standard against which the testee's behaviour may be compared.
- It is very difficult to specify relative degrees of mastery of a particular skill with sufficient precision to distinguish clearly between levels. This is

illustrated by Alderson's (1991:81-2) comment on the development of the IELTS Speaking scales:

> For some criteria, for example pronunciation or grammatical accuracy, the difference in levels came down to a different choice of quantifiers and we were faced with issues like is 'some' more than 'a few' but fewer than 'several' or considerable' or 'many'. How many is 'many'?

In addition to these problems, one of the major shortcomings in using generalized rating criteria is that they are too general to be applied to particular tasks. They do not, in other words, describe the qualities of individual task performances, nor do they describe what constitutes an acceptable standard of performance which is what is required in TCA (Pollitt 1991). Thus though rating scales of the general kind may be helpful in providing broad information for reporting purposes, they are of less use in assisting raters to make judgements relative to particular tasks and their construct validity continues to be surrounded by doubt.

3.2.3 Genre-Based Approaches to TCA

One way of obtaining detailed assessment information at the level of the individual task is represented by genre-based approaches to assessment which derive from the analysis of spoken and written genres within the framework of systemic-functional linguistic theory (Halliday 1985). Within this approach, the genres (such as *argument, describing a procedure*, etc.) are carefully described in terms of their structural organization and linguistic features. These features are then used as the basis for the implementation of a teaching-learning cycle and also serve as the criteria for assessment of overall task performance. In the context of primary English as a Second Language teaching, teachers have reportedly found this way of specifying tasks useful since it links the assessment criteria directly to what is being taught and focuses their attention on ways in which students are learning to make meaning (Mincham 1992). At the same time it offers the opportunity for learners to obtain diagnostic feedback on the extent to which they have met the criteria in the task, since the assessment checklist allows for differing levels of achievement to be recorded, ranging from 'very competent' to 'not yet'.

Genre-based approaches offer a way of describing and assessing language task performance that is underpinned by a powerful linguistic theory. Descriptions of genres provide explicit and testable hypotheses concerning the language demands of different text-types. The first step in a potentially valuable research agenda would be to determine to what extent the absence or presence of particular linguistic features in a text can be related to task difficulty. If a systematic relationship were shown to exist (as evidence from studies by Shohamy and Inbar (1991) and Pollitt and Hutchinson (1987) would seem to suggest), such information could be of great assistance in informing the rather vague descriptions of task and

text characteristics which are used to rate task performance ('can understand more abstract texts', 'can make simple requests', etc.).

However, as Murray (forthcoming) points out, the genre approach relies on the availability of very comprehensive descriptions of different oral and written genres and 'full descriptions of the structures of most oral and written genres have yet to be developed.' From the practical point of view the amount of work involved in filling in individual checklists for large numbers of students could also prove quite daunting for teachers.

A variant on this approach is found in the Certificate in Spoken and Written English (Hagan et al. 1993), a competency-based curriculum framework used within the Australian Adult Migrant English Program for immigrant learners of English. The Certificate sets out outcome statements in the form of language competency specifications which describe the elements of the language performance in question, the criteria by which the performance is to be judged and the range of variables which obtain in the assessment situation (e.g. the amount of assistance the learner may receive). Though underpinned by the same systemic-functional theory of language, what distinguishes the assessment system used in the Certificate from the approach previously described is that here the performance criteria are mandatory. Before they can be awarded the competency, learners must demonstrate evidence of each of the performance criteria. Thus for the competency 'can negotiate complex/problematic exchanges', the performance criteria to be demonstrated are as follows:

> **Competency 4. Can negotiate complex/problematic spoken exchanges for personal business and community purposes**
>
> Achieves purpose of exchange and provides all essential information accurately.
>
> Uses appropriate staging for text, e.g. opening and closing strategies.
>
> Provides and requests information as required.
>
> Provides and requests goods and services as required.
>
> Explains circumstances, causes, consequences and proposes solutions as required.
>
> Sustains dialogue, e.g. using feedback, turn taking, seeking clarification and understands statements and requests of the interlocutor. (Hagan et al. 1993:76)

It is interesting to contrast this set of performance criteria with the following example taken from the Royal Society of Arts Practical Skills Profile Scheme

which provides a method for assessing work-related and non-vocational courses in Communication, Numeracy and Process Skills.

Profile Sentence : C12 Participate effectively in negotiation

Performance criteria (the student has demonstrated the ability to:)

Define own preferred outcome in given situation.

State own needs/wishes clearly in language appropriate to listeners.

Express disagreement sympathetically.

Consider sympathetically suggestions of others.

Suggest new ideas to solve temporary difficulties.

Contribute to discussion freely and clearly without dominating the meeting. (Royal Society of Arts 1987.41)

This comparison illustrates graphically how the perspective of the test designer can influence the way that tasks are described and hence assessed. In the first case, it is primarily language that is the object of assessment (staging of discourse, conversational strategies, information giving, etc.). In the second, it is communication skills, in a more general sense, and non-linguistic, social and affective factors (intentionality, sympathy, empathy) which have a much greater role.

This raises a fundamental question that needs to be asked about TCA:

If task fulfilment is the principal criterion (as it usually is) for assessment, then to what extent should non-linguistic factors be taken into account? (For example, Clark and Scarino (1993:32) include as one of the generic criteria for judgement of performance in the Hong Kong Targets and Target Related Assessment (TTRA) curriculum 'effectiveness of the product in relation to the purpose and context expressed in the task').

It is interesting, and perhaps significant, to note in the context of this discussion that disciplines outside applied linguistics interpret 'communication' or 'communicative competence' quite differently and hence employ different criteria for assessment. Communication theorists, for example, accentuate criteria such as *empathy, behavioral flexibility* and *interaction management* (Wiemann and Backlund 1980) and emphasise the role of non-verbal aspects of communication. In other fields, such as organisational management, communicative ability is seen

very much in terms of 'getting the job done' and the success of communication is thus judged primarily in relation to how well the outcomes are achieved rather than on specific linguistic features (Brindley 1989:122-23). McNamara (1990:32) makes this point in relation to doctor-patient communication, noting that in the medical profession 'there is a concern for the communication process in terms of its outcomes.' He comments (1990:47) that 'sociolinguistic approaches to 'communicative ability' are indeed narrow, and narrowly concerned with language rather than communicative behaviour as a whole.' If TCA is concerned with task outcomes, then perhaps it is time for language testers who are concerned with certifying people's ability to get things done with language to reconsider the position they have conventionally taken with respect to these factors-that they are not part of communicative competence and are therefore not the object of assessment.

3.2.4 Towards data-based assessment criteria

I want to conclude this section on criteria for assessment by making some suggestions as to how criteria might be developed that are more consistent with current understandings of language acquisition and use.

First we need to compare data derived from studies of language in use conducted within a variety of theoretical paradigms with the descriptions of language skills and abilities that are used for assessment. Such research is important since what little work has been done tends to indicate major discrepancies between what actually happens and what test developers think happens. For example, Fulcher (1987) demonstrates that the criteria for fluency used in the ELTS Interview Assessment Scale do not reflect what happens in real conversational exchanges and that native speakers would, in fact, not meet the criteria. He recommends that data drawn from discourse analysis should be used to inform the constructs used in tests of oral performance. In a similar vein, Chalhoub-Deville (1993:20) in a study of the rating patterns of three groups of native speakers of Arabic assessing learners' oral performance concludes that 'research on L2 oral performance is needed ... that derives scales empirically according to the given tasks and audiences.'

At the same time, a lot more information is also needed about the cognitive demands that different types of tasks make on learners. In this regard, research evidence suggests that tasks that may appear to be of similar overall complexity may make different processing demands on learners (Bialystok 1991:121).

In the light of this and similar findings (see, for example, Snow et al. 1991; Chalhoub-Deville 1993), it is important to establish a principled way of describing and evaluating tasks. In order to throw more light on the question of task demands, as Bialystok (1991) points out, it will be necessary to undertake task analysis in different situations of language use. Bialystok suggests that task demands can be described in terms of two processing dimensions, namely *analysis of linguistic knowledge* and *control of processing*:

> Thus, the demands imposed upon language learners by various language uses can be described more specifically in terms of the demands placed upon each of these skill components, and the proficiency of learners can be described more specifically by reference to their mastery of each of the skill components (Bialystok 1991:64).

Once the qualitative characteristics of the tasks are known, banks of tasks can be developed and trialed. Using Rasch techniques, the tasks can then be calibrated on a common scale and related to defined bench-mark levels of performance. Griffin and McKay (1992) provide a description of how this can be done in a principled way.

Second, if assessment tasks are going to be closely related to 'real-world' tasks, then it is necessary to gather more information on those criteria that are used by other people outside the language classroom to make judgements on learners' task performance (Brindley 1991). This is particularly important in the case of learners in second language contexts. After all it is not teachers' judgements of students' language ability that will decide whether they manage to communicate in the 'real world.' Outside the classroom it is the lay person's impression of people's communicative effectiveness that will determine the extent to which learners' communicative goals are achieved. There is increasing evidence to suggest that non-teacher native speakers use quite different criteria in judging language performance from those used by teachers (Shohamy et al. 1992: Chalhoub-Deville 1993). If it is the judgement of these non-teachers that determines learners' communicative acceptability, it is necessary to investigate how these judgements are made, what criteria are used, and perhaps attempt to take these criteria into account in the construction of the instruments that are used to assess proficiency and achievement.

4. Reliability in TCA

4.1 The problem of human judgement

TCA relies heavily on teachers' subjective judgements of language performance. In the interests of fairness to learners, it is important that these judgements are seen to be reliable. As more and more rating tools are developed to assess productive task performance, teachers will need to be trained to interpret and apply assessment instruments in a consistent way. Rater training involving familiarization with the rating criteria and practice in applying them to samples of performances across a range of ability levels has long been standard practice with proficiency rating scales and it has been claimed that high levels of inter-rater agreement can be obtained in this way (e.g. Dandonoli and Henning 1991).

However the feasibility of obtaining inter-rater reliability with respect to language performance has come increasingly under question. Research in language testing has shown that despite training, 'significant and substantial differences between raters

persist' and that rater behaviour can change significantly over time (Lumley and McNamara 1993). North (1993:45) in a comprehensive survey of the whole field of subjective judgements in rating concludes that 'judge severity is relatively impervious to training and that people rate in different ways.' This, of course, is hardly surprising given the complexity of the interaction between the language behaviour being rated, the personal characteristics of both the rater and the candidate, and aspects of the setting in which the rating takes place. However, it leaves the language tester in a dilemma: if variability in rater behaviour is the norm, then what - if anything - can be done to reduce the error in rater judgements?

4.2 The promise of new measurement technology

One possible solution to this problem is offered by recent advances in measurement technology in the form of multi-faceted Rasch analysis and its accompanying software package known as FACETS (Linacre 1988).

The Rasch model is one of a family of techniques known as latent trait theory or item response theory (IRT) which have been developed by psychometricians over the last three decades or so. One of the strengths of the theory is that it allows candidate ability and item difficulty to be estimated independently and reported on a common scale, thus avoiding many of the problems associated with sample-dependent classical measurement techniques (Henning 1987). The multi-faceted Rasch model extends previous Rasch models to include rater characteristics. It provides an estimate of candidates' ability based on the probability of a candidate obtaining a particular score on a particular task given the ability of the candidate, the difficulty of the item (in the case of language assessment this might be the rating category such as *fluency* or *cohesion*), the harshness of the rater, and the effect of any additional facets (Linacre 1989). The program adjusts candidate ability estimates to take account of raters' tendency to rate either harshly or leniently.

The use of FACETS can assist in the analysis of ratings of task performance in a number of ways:

- Since FACETS accepts variability and compensates for rater severity, it is not necessary to try to achieve complete agreement between raters. As long as raters are internally consistent, there is no need for raters whose rating patterns appear to be deviant to be excluded.

- It enables reports to be provided to raters showing their tendency towards severity and leniency. It also shows how each rater is using the steps on the scale.

- Through a technique known as bias analysis, it enables the interactions between different aspects or 'facets' of the rating situation to be modelled and examined, e.g. it is possible to see whether a certain rater is rating more or

less harshly on a particular task or rating category (Lumley and McNamara 1993).

- It enables the rating categories used in assessing oral or written task performance to be subjected to scrutiny. If a rating category does not fit the underlying model, indicating an inconsistent pattern in scoring, it is flagged by the program as 'misfitting'. This allows rating criteria to be monitored and revised as necessary (Wigglesworth and O'Loughlin 1993).

I am not suggesting that new measurement technology can answer all the problems that will inevitably arise on the ground from the fact that raters will not always agree on the quality of task performance. It would be unrealistic to expect that many educational institutions would invest in the necessary training and expense associated with multi-faceted Rasch analysis. However there would appear to be no reason why co-operative research ventures could not be undertaken between educational institutions wishing to monitor the way in which subjective judgements are being made on task performance and on institutions with the necessary expertise in the use of the technology.

On a day-to-day level, institutions will still have to find ways of trying to achieve a common understanding and definition of different standards of learner performance. In this regard the collection and analysis of 'bench-mark' performance samples, accompanied by regular moderation sessions, has proved a useful way of focusing raters' attention on key aspects of task performance at different levels (Griffin and McKay 1992). Another way of trying to accommodate for rater severity without the benefit of technology is outlined by North (1993:45). He describes a procedure for oral assessment using two assessors, one who knows the class in question (high sensitivity) and one who is familiar with the whole range of the level (low sensitivity). Ratings are carried out independently using both holistic and analytical marking 'with negotiation over grades between the two assessors as a final step to adjust for severity' (ibid).

5. Practicality

Finally I would like to turn to the rather crucial issue of practicality. Though it is widely agreed that TCA has significant benefits, it is also likely to be affected by a number of pressures and constraints, including financial cost, time, expertise and demands for external accountability. To what extent can TCA be made to work, given these constraints?

5.1 Financial cost/time

There is no doubt that TCA is extremely time-consuming and by extension, expensive. Eliciting individual performances is much more difficult and time-

intensive than administering pencil-and-paper tests. Commenting on the introduction of performance assessment in general education in the United States, O'Neil (1992:18) reports that 'some experts say performance assessments are likely to be at least two or three times more expensive per student.' Worthen (1992:452) suggests that 'the labour intensity of scoring and the need to observe performance over extended periods are primarily responsible for the high costs of performance assessment.'

Nuttall (1992:56) notes that teachers who administered the 1991 Standard Assessment Tasks in the UK, although they had learned new things about children's attainment 'found the tasks to be demanding of their time and energies; invariably, to prepare, administer and grade them required an average of 44 hours.'

On a similar note, Barrs (1992:55) comments that a common concern voiced about the implementation of the detailed observational recording system used with the Primary Language Record in the UK was the sheer amount of time necessary to document many student performances on an ongoing basis:

> Keeping detailed observational records of up to thirty children seems just too difficult.

She observes, however, that this aspect gradually became more manageable but noted that:

> ... it does seem to be the case that it takes a full school year to "learn the forms", to internalise the ways of observing that they encapsulate and to see the full value of this kind of recording (Barrs 1992:56).

These experiences would indicate that TCA has to be seen as a long-term rather than a short-term investment.

5.2 Teacher development

TCA is demanding not only in terms of time but also of teacher skill and a considerable investment in teacher development is necessary if teachers and learners are to obtain the maximum benefit from its use. In this context it should be noted that changing teaching and assessment practices or adopting new tools is no different to introducing a new curriculum or a new textbook. The introduction of TCA is an exercise in change management which by definition means trying to plan for the implementation of whatever change is proposed as a result of the professional development activities that are offered. If educational administrators are concerned with the long-term effects of what they do, then they need to be aware of this. Workshop participants who return to their institutions full of enthusiasm for new assessment ideas or tools cannot automatically be expected to

apply their ideas. If they are to change their assessment practices or systems, they require support in terms of time, funding, resources and sometimes skilled support personnel (Fullan 1982).

The importance of providing this support cannot be overestimated. If teachers are not given adequate assistance in understanding and implementing new modes of assessment, the whole purpose of the introduction of TCA may be undermined. This is graphically illustrated by Shohamy (1993) in a study of the impact of three different types of language tests on teaching and learning in Israel. She reports on the introduction of a form of TCA in the Israel-EFL oral test, which is part of the national matriculation examination taken by high school students at the end of twelfth grade. Although the rationale of the test was to place greater emphasis on oral proficiency and improve students' speaking skills, she found that teachers perceived oral language 'exclusively in terms of testlike activities.' Thus when asked to define 'oral language' teachers often gave answers such as 'It is a role play' or 'It is an interview'.' She concludes (1993:15) that 'in terms of the nature of the test effect, in all three cases the results showed the instruction became more testlike and that this was most likely a result of teachers not having been trained to teach the new areas being tested' and adds that 'when teaching and testing become synonymous, the tests become the new de facto curriculum.'

5.3 External accountability

One of the main difficulties in introducing TCA in general education has been its public acceptability. Nuttall (1992) reports that the introduction of performance assessment into the school curriculum in the UK has alarmed some people who are used to standard pencil-and-paper assessment and who feel that the new kinds of assessment are less rigorous:

Unfortunately the minority who take this view have the ear of Prime Minister John Major, who has decided that the amount of performance assessment must be significantly curtailed. His exact words were:

> By testing I do not mean some weird experiment in a corner. (i.e. a reference to the Standard Assessment Tasks). What I mean is pencil-and-paper testing for a classroom so people have a measure of how they are doing-see if there is a problem so that you can put it right.

Nuttall (1992:57) reports that the day after the Prime Minister's pronouncement, the development contracts for Grade 2 Standard Assessment Tasks were cancelled and new tenders called for the development of pencil-and-paper tests.

Hopefully the justification for the use of TCA in language assessment is more self-evident. However Nuttall's comments are a sobering reminder that whenever new forms of assessment are introduced they need to be accompanied by very

clearly presented and accessible statements explaining their purpose, use and justifying their financial costs. The latter is a particularly important factor where large-scale oral assessment is concerned.

6. Conclusion

In this paper I have identified a number of key issues relating to the validity, reliability and practicality of task-centred assessment and I have suggested ways in which some of the potential difficulties associated with its implementation and use might be addressed.

Overall, experience in general education in a number of countries seems to indicate that TCA can be made to work. This is, I suspect, because it has the strong support of both teachers and learners - it takes assessment out of the realm of something which is done *to* students into the realm of something that can be done *with* them. In the words of Broadfoot:

> In place of ubiquitous competition and external judgement, assessment is harnessed to teaching to provide explicit statements of curriculum goals; to equip pupils with the skills to set their own goals and review progress towards them; to make pupils jointly responsible with teachers for both formative and summative reviewing and reporting (Broadfoot 1988:5).

However, a number of challenges remain. In the first place, in relation to the validity question, as I have indicated at various points throughout this paper, a lot more work needs to be done in order to develop assessment criteria which reflect current theories of language learning and language use. Second, as far as reliability is concerned, if TCA is to have public credibility, the problem remains of trying to ensure consistency in the application of assessment criteria. While new measurement technology may provide some solutions to this problem, this technology will only be available to a few. As some testers have suggested, one of the consequences of adopting TCA may well be learning to rethink the notion of reliability:

> ... if we indeed value clinical judgement and a diversity of opinions among appraisers (such as certainly occurs in professional settings or post-secondary education), we will have to revise our notions of high-agreement reliability as a cardinal symptom of a useful and viable approach to scoring student performance. We will have to find a previously uncharted course between insisting on uniform judgements and mayhem. Possibly, we will have to seek other sorts of evidence that responsible judgement is unfolding-that participants agree on the relevant categories for describing performance, that scores fall within a certain range, or that recipients can make thoughtful use of the range of opinions offered to them (Wolf et al. 1991:63).

Finally, at a practical level, the adoption of TCA has a number of major consequences at all levels of an educational system. For teachers and learners it means that they will need to become accustomed to thinking of language tasks not only as activities but also as indicators of progress and achievement. Learners will thus need to understand the criteria according to which their performances will be judged. This, in turn, will necessitate a closer examination of the components of language tasks and a raising of learners' awareness of how language functions to achieve particular communicative purposes. As far as 'consumers' of assessment information are concerned, the reporting of task performance means that they may have to be persuaded to accept assessments that are complex, qualitative and multidimensional, rather than uniform and standardized. This will not be easy and will necessitate close co-operation and continuing dialogue between all of the stakeholders involved in language programs.

There is no doubt that task-centred assessment in language learning is firmly established at the level of the classroom where it has demonstrated the potential to bring about significant improvements in the quality of learning (Shohamy 1993). As it moves into high stakes areas such as certification and selection, however, it remains to be seen whether the momentum will continue. In this regard, we can only hope that the value of TCA will become as evident to those outside the classroom as it is to those within it.

References

Alderson, J.C. & Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. *Reading in a Foreign Language*, 5(2).

Alderson, J.C. (1991). Bands and scores. In J.C. Alderson & B. North (eds.), *Language Testing in the 1990s*. London: Macmillan.

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

Barrs, M. (1992). The primary language record: what we are learning in the U.K. In C. Bouffler, (ed.) *Literacy evaluation*. Sydney: Primary English Teaching Association.

Bialystok, E. (1991). Achieving proficiency in a second language: A processing description. In R. Phillipson, R. Kellerman, L. Selinker & M. Sharwood Smith, M. (eds.) *Foreign / Second language pedagogy research*. Clevedon, Avon: Multilingual Matters.

Broadfoot, P. (1988). *Introducing profiling*. London: Macmillan.

Breen, M. P. (1987). Learner contributions to task design. In C.N. Candlin & D. Murphy (eds.) *Language learning tasks*. Englewood Cliffs: Prentice Hall.

Brindley, G. (1986). *The assessment of second language proficiency: Issues and approaches*. Adelaide: National Curriculum Resource Centre.

Brindley, G. (1989). *Assessing achievement in the learner-centred curriculum*. Sydney: National Centre for English Language Teaching and Research.

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

Broadfoot, P. (1988). *Introducing profiling*. London: Macmillan.

Chalhoub-Deville, M. (1993). Performance assessment and the components of the oral construct across different tasks and rater groups. Paper presented at Language Testing Research Colloquium, Arnhem, August.

Clark, J. & Scarino, A. (1993). The integration of language and content in TTRA: implications for schools. In N. Bird, J. Harris & M. Ingham (eds.) *Language and content*. Hong Kong: Government Printer.

Crookes, G. (1986). *Task classification: A cross-disciplinary review*. University of Hawaii at Manoa: Centre for Second Language Classroom Research, Social Science Research Institute.

Dandonoli, P. & Henning, G. (1991). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals*, 23(1):11-22.

Devenney, R. (1989). How ESL teachers and peers evaluate and respond to student writing. *RELC Journal*, 20(1).

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELTJ* 41(4).

Fullan, M. (1982). *The meaning of educational change*. Toronto: Ontario Institute for Studies in Education.

Griffin, P. & Nix, P. (1991). *Educational assessment and reporting*. Sydney: Harcourt Brace Jovanovich.

Griffin, P., & McKay, P. (1992). Assessment and reporting in ESL Language and Literacy in Schools project. In P. McKay (ed.). *ESL development: Language and literacy in schools. Volume II: Documents on bandscale development and*

language acquisition. Melbourne: National Languages and Literacy Institute of Australia.

Hagan, P., Hood, S., Jackson, E., Jones, M., Joyce, H. & Manidis, M. (1993). *Certificate in spoken and written English*. Sydney: NSW Adult Migrant English Service and the National Centre for English Language Teaching and Research.

Halliday, M.A.K. (1985). *An introduction to functional grammar*. London: Edward Arnold.

Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Lantolf, J. & W. Frawley (1985). Oral proficiency testing: a critical analysis. *Modern Language Journal*, 69(2).

Lantolf, J. & W. Frawley (1988). Proficiency: defining the construct. *Studies in Second Language Acquisition*, 10(2).

Linacre, J. M. (1988). *FACETS: A computer program for the analysis of multi-faceted data*. Chicago: MESA Press.

Linacre, J. M. (1989). *Multi-faceted measurement*. Chicago: MESA Press.

Long, M. H. & G. Crookes (eds.) (1993). *Tasks in a pedagogical context*. Clevedon, Avon: Multilingual Matters.

Long, M.H. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam & M. Pienemann (eds.) *Modelling and assessing second language acquisition*. Clevedon, Avon: Multilingual Matters.

Lumley, T. & McNamara T.F. (1993). Rater characteristics and rater bias: implications for training. Paper presented at Language Testing Research Colloquium, University of Cambridge, August 1-2.

McNamara, T.F. (1990). Assessing the language proficiency of overseas-qualified health professionals. Ph. D. Thesis, University of Melbourne.

Mendelsohn, D. (1989). Testing should reflect teaching. *TESL Canada Journal*, 7(1).

Mincham, L. (1992). Assessing the English language needs of ESL students. In C. Bouffler (ed.) *Literacy evaluation*. Sydney: Primary English Teaching Asociation.

Murray, D. (forthcoming). Using portfolios to assess writing. To appear in *Prospect*, 9(2).

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

Nunan, D. (1989). *Designing tasks for the communicative classroom.* Cambridge: Cambridge University Press.

Nunan, D. (1993). Task-based syllabus design: selecting, grading and sequencing tasks. In M.H. Long & G. Crookes (eds.).

Nuttall, D. (1992). Performance assessment: the message from England. *Educational Leadership*, May.

O'Neil, J. (1992). Putting performance assessment to the test. *Educational Leadership*, May.

Pienemann, M., Johnston, M. & Brindley, G. (1988). Constructing an acquisition-based assessment procedure. *Studies in Second Language Acquisition*, 10(2).

Pollitt, A. (1991). Response to Charles Alderson's paper: 'Bands and scores'. In J.C. Alderson & B. North (eds.) *Language testing in the 1990s.* London: Macmillan, pp. 87-94.

Pollitt, A. & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing* 4(1):72-92.

Royal Society of Arts (1987). *Practical skills profile scheme.* London: Royal Society of Arts.

Shohamy, E. (1992). New modes of assessment: the connection between testing and learning. In E. Shohamy & R. Walton (eds.). *Language assessment for feedback: Testing and other strategies.* Dubuque, Iowa: Kendall/Hunt Publishing Company.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning.* Washington, D.C.: The National Foreign Language Center.

Shohamy, E. & Inbar, O. (1991). Validation of listening comprehension tests: the effect of text and question type. *Language Testing*, 8(1):23-40.

Shohamy, E., Gordon, C. & Kraemer R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, 76(2).

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing*, 1(2).

Snow, C.E., Cancino, H., de Temple, J. & Schley, S. (1991). Giving formal definitions: a linguistic or metalinguistic skill. In E. Bialystok (ed.) *Language*

processing in bilingual children. Cambridge: Cambridge University Press. pp. 90-112.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing*, 2(1).

Swain, M. (1993). Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? *Language Testing* 10(2).

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Wigglesworth, G., & 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(1).

Wiemann, J, & Backlund, P. (1980). Communicative competence. *Review of Educational Research*, 50(1).

Wolf, D., Bixby, J., Glenn, J. & Gardner, H. (1991). To use their minds well: New forms of student assessment. *Review of Research in Education* 17.

Worthen B. (1993). Critical issues that will determine the future of alternative assessment. *Phi Delta Kappa*, February.