PASAA
Vol. 25
December 1995

Testing for Communicative Proficiency: A Scenario-Based Approach

Jane Davies and Gavin Hibbs
Institute of Foreign Affairs; Ministry of Foreign Affairs

Abstract

This article gives an account of a presentation given at Chulalongkorn University Language Institute *Forum* on 6 September, 1995, reporting on work in progress on a current Institute of Foreign Affairs project which concerns the delivery of a training and testing package to implement new Ministry policy to upgrade the communication skills of its support staff in the workplace.

Under the new policy, MFA support staff must now pass a test of English proficiency in order to be considered for posting to work in Royal Thai Embassies overseas. The article focuses on the issues surrounding the design of an appropriate testing procedure, outlining the steps in the preparatory research and clarifying the concept and rationale of the ultimate choice of a "scenario-based" approach.

While the content of this particular programme is genre-specific, the issues raised concerning how best to enable transfer of language training from classroom to real-life workplace proficiency, and how such transfer can be evaluated, are increasingly of general concern in a wide variety of training contexts, as higher and higher levels of English skills come to be more and more in demand in the competitive job market.

I. Background to the Project

The Institute of Foreign Affairs, the training centre for Ministry of Foreign Affairs personnel, is currently involved in a project to deliver a training and testing package to implement a new Ministry policy to upgrade the communication skills of its support staff in the workplace. The team responsible for this project is made up of three foreign teachers, the writers of the paper, together with Mr. Frank Thorne.

Under the new policy, MFA support staff must now pass a test of English proficiency in order to be considered for posting to work in Royal Thai Embassies overseas, with the aim of improving overall efficiency in the foreign missions. The key

challenge, therefore, in designing the test, was to develop an instrument to effectively assess whether or not support staff would be able to perform adequately in English at their post.

In other words, a test was needed, not to find out what candidates knew *about* English, but rather what they could effectively *do* in English: that is, a test which would determine how well they could mobilize their linguistic resources in order to carry out those duties and responsibilities which would be required of them *in* English. To use the Widdowson (1978) terminology, the test had to be a communicative test of language *use*, rather than *usage*.

The practical, skills-orientated message of the test would be underpinned by a comprehensive, preparatory training package, a design task which presents its own related set of challenges and problems. However, it is the test itself which will form the focus of this paper.

II. Communicative testing

The concept of communicative testing is hardly new. As early as 1968 — nearly 30 years ago — Spolsky was affirming the need to "test a person's ability to perform in a specified sociolinguistic setting" (Spolsky 1968). As the communicative approach to language teaching gathered momentum in the 1970's, so the need — if not the supply — grew for new, more appropriate tests to measure the learning outcome of this teaching, assessing achievement in terms of communication skills acquired rather than the ability to handle standard objective exercises.

Earlier tests had concentrated on evaluating linguistic competence: knowledge of grammar, structure, and the system of the language, biased towards analytical skills of input-processing rather than the dynamic skills of output production. Such tests discriminate between a clearly defined "right" and "wrong" in the language, and have the advantage of being thus very easy and reliable to mark.

However, as Morrow (1979) puts it, "knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to *use* the language" (emphasis added).

Savignon (1972) even earlier had agreed that "grammatical competence was not by itself a good predictor of communicative skills."

As long as the language teaching too had focused on these discrete points of formal accuracy, there had been at least a kind of validity to these tests in that they were assessing what had been retained from what had been taught. However, while the emphasis of the teaching gradually switched to a more communicative approach, changes in testing procedures generally failed to keep pace. Increasingly, a mismatch grew up between teaching and testing aims: while teaching focused more on the *communicative* competence side, tests tended to carry on testing *linguistic* competence. Given this confusingly mixed message, our students, conditioned over centuries to attach inordinate importance to examinations, can hardly be blamed for looking to the test as the signal of where the true learning agenda lies. The backwash effect of the University Entrance examination on language teaching and learning attitudes here in Thailand is a good case in point.

It seems to me that the reasons why testing methods have not changed commensurately with teaching approaches are closely bound up with the tensions between the three key, conflicting demands of test design: *reliability*, *validity* and *efficiency*.

III. Juggling the variables: reliability, validity, efficiency

Test design involves a delicate balancing act between three key elements:

- **Reliability** — can we rely on the results of this test?
- **Validity** — does the test measure what it is supposed to measure?
- **Efficiency** — is the test practical in design/to administer?

The great advantage of standardised tests is that they are *reliable* and *efficient*. Where there is only one best answer, right and wrong, different markers can be relied upon to get the same results on

any day in any situation, and can mark large numbers of papers quickly and easily. However, if our goal is to assess communicative proficiency, then we cannot say these tests are *valid*: they are measuring linguistic knowledge not communication skills.

To be *valid* we need to give the candidates authentic communication tasks, but then assessing these *reliably* becomes problematic. Where appropriateness rather than formal correctness is the criterion, marking necessarily becomes subjective and may vary greatly from one marker to another, setting up an opposition between reliability and validity.

> While one can have test reliability without test validity, a test can only be valid if it is reliable. There is thus sometimes said to be a reliability-validity tension. It is sometimes necessary to sacrifice a degree of reliability in order to enhance validity. If, however, validity is lost to increase reliability we finish up with a test which is a reliable measure of something other than what we wish to measure. (Weir 1990: 33)

Additionally, if we solve this problem of reliability, to gain maximum validity, a larger sample of test items is required in order to cover the whole range of communication skills in which the candidate may be involved in the real-life situation. This means that the test will necessarily become longer and therefore less *practical* to administer, a concern of crucial importance if we are talking about large numbers of candidates. As Weir (1990: 34) points out: "A valid and reliable test is of little use if it does not prove to be a practical one."

Faced with these conflicting choices, it is easy to see why reliability and efficiency have so consistently prevailed over concerns of greater validity in test design: communicative tests are quite simply much harder to deal with: "The increased per-capita cost of using communicative tests in large-scale testing operations may severely restrict their use" (Weir 1990: 35).

However, problematic as it may be, we clearly have a professional duty to attempt to develop test formats and evaluation criteria that go some way at least to balancing these needs for reliability, validity and efficiency in the assessment of communicative skills.

We at the Institute felt that to really fulfil our brief as well as to initiate an attitudinal shift in trainees regarding the relative importance between practical language skills and analytical grammatical knowledge, we had to try to design a communicative test which would require candidates to perform the types of task they would have to do at post, possibly sacrificing a little reliability and efficiency in the process, but gaining in terms of validity and authenticity, a view supported by Weir: ". . . if a choice has to be made, validity, after all, is the more important" (Weir 1990: 33).

It is worth stressing at this point the somewhat privileged position at the Institute. There are three of us working closely as a team, and the numbers of candidates we have to deal with are not overly large: about 100 or so twice a year. We were thus in a position to be able to handle a loss in efficiency in order to have a long and valid test, and to choose validity with minimum loss of reliability by training ourselves to use the ESU Band Scale criteria references for assessing written and spoken output.

We fully realise the institutional constraints that may militate against the choice of this course of action in other contexts. Even the British Council here in Bangkok, where teacher/student ratios are relatively low, and where an oral component is included in its placement test system, does not routinely test oral communication in its end-of-course assessment, largely for practical reasons. It is difficult — though stimulating — to envisage how a test of spoken communication could be incorporated into high school, entrance and university examination systems. One of the most well-known public exams is of course TOEFL, which makes no concessions towards communication validity. A number of public examinations, such as IELTS, PET and TEEP in Britain, as well as TOEIC in the U.S.A., have been working to resolve this dilemma. However, it seems there is still a great deal of important work left to do.

IV. Features of communicative testing

Having established the desirability — notwithstanding the difficulties — of communicative tests, it is as well to examine in more detail the

nature of how such tests might be composed. Weir (1990) states unequivocally: "Only direct tests which simulate relevant authentic communicative tasks can claim to mirror actual communicative interaction." Authenticity is thus the key word here. Whilst not wishing to criticise the university Entrance examination unfairly, having as it does a virtually impossible task to perform, we might like to consider, as an example, the activity whereby students have to fill in the gaps in dialogues. On the surface, this is a step towards recognising the spoken language dimension. However, the task could by no means be considered authentic, as it does not reflect the way in which actual discourse is processed. In a real dialogue, participants have to process aural input, formulate an oral, rather than written response with appropriate non-verbal and paralinguistic features, under time pressure with no option to monitor, go back and check, revise and so on. The gap-fill dialogue is unable to assess interactive skills requiring negotiation of meaning and fulfilment of personal communication goals. Neither can it claim to be an authentic piece of spoken language, being specially made up by the examiners for the sole purpose of testing.

Weir (1990) expresses this most succinctly:

> It is held that the performance tasks candidates are faced with in communicative tests should be representative of the type of task they might encounter in *their own real-life situations* and should correspond to *normal language use where an integration of communication skills is required with little time to reflect on or monitor their language input or output.* (emphasis added).

In other words, the test tasks should be as similar as possible to the real-life tasks candidates are to be required to perform. While acknowledging, along with Davies (1978), that full authenticity may be merely a "chimera," Skehan's (1988) view is pragmatically optimistic: "While such tests may not replicate exactly the performance conditions of a specific task in the target situation they are *likely to replicate to some degree conditions of actual performance*" (emphasis added). Perhaps this is the best we can do: to travel hopefully towards a more communicative test even if we know we will never quite arrive.

The principle features of communicative tests are summarised in Figure 1 below.

Figure 1 *Features of Communicative Tests (Adapted from Weir 1983a, 1990; Hawkey 1982; Morrow 1977, 1979)*

- Interaction between participants emphasised — candidates both receive and produce language in register appropriate to role and setting
- Unpredictability of language use allowed for
- Purposeful nature of communication reflected — task fulfils some communicative function
- Authenticity of tasks, contexts and domains relevant to candidate's real-life situation
- Integration of 4 skills.
- Appropriateness of language use emphasised rather than accuracy
- Realistic discourse processing reflected
- Qualitative, rather than quantitative assessment of productive abilities — use of rating scales vs. scores

V. A SCENARIO-BASED APPROACH

Any communication event in real life takes place in a context: based on our knowledge of who we are speaking to, about what, when, where and why, we are able to draw on the linguistic resources at our disposal to make choices about what is and is not appropriate to say. Hymes (1972) defines this as the *sociolinguistic context.* If knowledge of one of these elements is missing, decisions about language appropriacy become problematic.

For this reason, we believe that if we are truly to test a candidate's ability to perform in real-life contexts, where each of the variables in the sociolinguistic context would generally be known, it is only fair that this sociolinguistic context should be clearly specified at every stage of the test process also. As Weir (1990) reminds us, "language cannot be meaningful if it is devoid of context." Oller (1979) too underlines the importance of contextualisation to the selection process: "The higher the level at which language is contextualised, the more effective language perception, processing and acquisition are likely to be."

Additionally, apart from facilitating the examinee's decision-making process about language appropriacy, if we as examiners are to pass judgement on the degree of appropriacy achieved, we have to be as clear as possible about the precise elements of the sociolinguistic context. Even a minor shift in any one of the elements will almost certainly signify a need for a different language choice appropriate to the new situation.

We have coined the term "scenario" as a convenient label for this sociolinguistic-context specification in both testing and training.

Both the generally accepted meaning of the term "scenario" as well as the more detailed, genre-specific one used by di Pietro (1987), seem to combine elements appropriate to our purposes. The *Longman Dictionary of Contemporary English* defines "scenario" as "a description of a possible course of action/events," which is a useful starting point, but for our needs requires the further refinement offered by di Pietro (1987).

While diverging from us in a number of ways in his application of the term to language teaching, di Pietro gives three useful definitions:

> a realistic happening involving the unexpected and requiring the use of language to be resolved;
>
> a label for real-life happenings that entail the unexpected and require the use of language to resolve them;
>
> a strategic interplay of roles and functions to fulfil personal agendas within a shared context.

This last definition we find particularly appropriate, emphasising as it does all the key aspects of the communication process outlined in (IV) above: interaction, role awareness, purposefulness and shared context.

Our own working definition includes by implication the dimensions stressed by di Pietro, while emphasising Hymes' insights: "a clearly-defined, authentic, relevant sociolinguistic context specifying who is talking to whom, about what, when, where and why."

VI. The pre-overseas posting English test [*POPET*]

Having stressed the importance of authenticity in contextualisation of test tasks, it was crucial to find out what these might be in our particular case. Again, we are in a somewhat privileged position at the Institute, in that it is possible to specify quite precisely what kinds of activities support staff not only *are* involved in, but also what kinds of activities they ideally *ought* to be involved in given a sufficient level of English proficiency. In many teaching institutions, the ultimate target situation may be far more vague and undefinable, making the task of test design even more challenging.

We sent out a questionnaire to Heads of Section at Royal Thai Embassies to find out their desired profile of the kinds of tasks and duties support staff should ideally be able to perform in English. To be as minimally directive as possible, we left the questionnaire very open-ended, asking Heads of Section simply to list in as much detail as possible the various scenarios they would like their support staff to take charge of, both in the workplace and in their personal life, under the headings: administration, protocol, consular, socialising, personal and miscellaneous other duties.

Unfortunately the results were not back in time. Such are the constraints we operate under in the real world. We had believed that training would precede the testing, but in fact it turned out that the test was to come first, to be used as a predictor of language proficiency at post, rather than as the achievement test we had envisaged. Ironically, our "authentic" scenarios had to come largely from our own imaginations, though we were also able to get a certain amount of advice from local Ministry sources.[1]

As previously mentioned, our main aim was to attain maximum validity. For this reason, the test we designed is indeed very long, but we felt that in trying to gain an overall skills profile, it was crucial to test over a wide range of types of tasks. The test may also appear rather complex, for much the same reasons.

Up to this point, it may have been implied that communicative testing is synonymous with oral testing, but in fact, we set out to integrate the skills as far as possible. All four skills were tested com-

[1] Though it was pleasing to note, when the questionnaire results came through, that our projections of realistic scenarios proved to be remarkably appropriate. See Section VII below.

municatively in the context of scenarios. For the reading component, examinees had to read a fax addressed to their Head of Section, extract the key information and work out what the writer wanted to know in reply. For the writing component, examinees were given relevant extra information and then asked to draft a suitable reply to the same fax they had just read.

For the listening section, examinees had to listen to recorded messages on an Embassy telephone answering machine, make notes of the key information and decide what action needed to be taken in each case.

The oral test was the most complex component. One of the key aspects of the support staff role overseas is as representatives of Thailand. In this connection, it is crucial that at each interaction, face-to-face as well as over the phone, in the office as well as at official social functions, a positive image of Thailand and Thai people is being presented, so we needed a test which would cover the range of types of interaction.

The first part aims to assess the social English dimension, consisting of a general "conversation," with candidates answering general questions about themselves, their life and their work on a one-to-one basis with the examiner, who tries to push the talk to the candidate's language ceiling, while striving to keep as far as possible a "conversational" rather than an inquisitorial, interview-style tone.

Inevitably, however, the talk tends to be examiner-led, with the examiner assuming the dominant interactional role, taking responsibility for turn-taking, topic negotiation and shifts. This section fails to test candidate's ability to get information, or to give detailed, accurate factual information, to take the initiative in the interaction, or to pursue their own communication goals. As a testing instrument, therefore, while it has its uses, it also has limitations, necessitating further tasks to get a true global profile.

For these reasons, we have added two further sections. Firstly, candidates have to answer a telephone call at the Embassy and deal with an enquiry. Here, language appropriacy is vital in order to project a polite, efficient and informative image of the Embassy. This part of the test also checks ability to transmit factual information accurately, as the enquiry may concern requirements for visa extensions, malaria precautions or export regulations. Candidates are given the necessary background information to study five minutes beforehand in each case, but in written language format. This section of the test allows them to demonstrate their skills for structuring and explaining this information in oral mode, as well as their ability to negotiate meaning. This section also addresses the additional challenge of oral performance in a different channel – over the phone – where candidates have to cope without the aid of non-verbal clues.

The third section of the oral test requires the examinee to perform some realistic task related to Embassy work, such as making arrangements for Embassy guests at a travel agent's, organising car hire for a visiting delegation, or inquiring about accommodation for Embassy personnel at an estate agent's. Candidate's are given ten minutes to study the scenario and familiarise themselves with the task requirements before joining the examiner, who takes the complementary role in the role-play. In this section, candidates have to deal with the unpredictability of face-to-face interaction, while fulfilling the communication goal specified. They are forced to take the initiative in the role-play in order to get the information they require to complete the task, checking their understanding, reformulating their questions and asking for clarification. This we find to be the most challenging part of the test.

In order to try to avoid rehearsed responses as a result of candidates "leaking" to their succeeding colleagues what the oral scenarios were about, we designed a menu of three possible scenarios – of equivalent levels of difficulty – in both the telephone and face-to-face sections. The examiner would choose one from this menu for each candidate.

To gain a little in efficiency, each stage of the oral test allows for a cut-off point. Candidates had to attain at least a Band 3 on the *ESU* scale (see Appendix One) in the first section in order to get the chance to proceed to Section 2. Similarly, only candidates achieving a Band 4 on the next section were allowed to proceed to the final section. We were very fortunate in having very supportive ad-

ministrative staff to help with the logistical organisation of this three-stage procedure.

The reading, writing and listening tests were taken by all 107 candidates simultaneously. The oral testing was done by three teachers over two days: about eighteen people each to process through a potential maximum of three test instruments.

For the receptive skills, an objective scoring system was used and then converted to *ESU* Band scales. The productive skills, speaking and writing, were assessed according to the relevant *ESU* scales, giving an overall skills profile, though the question of relative skills weighting in the final aggregate score is an issue difficult to resolve.

Bearing in mind that the aim of the test was to determine whether candidates could function adequately at post, we had determined that an *ESU* Band 4 was probably the minimum required coping level. At this level, examinees have the confidence to keep the channels of communication open and not give up, even though their understanding and their language production may still be characterised by a great many inaccuracies and inappropriacies, hesitations and false starts. The test therefore only aimed to discriminate between Bands 1-5 and not between any of the levels above.

This is not to say, however, that support staff are encouraged to see Band 5 as the final culmination point of their language learning. On the contrary, it is very much hoped during the upcoming training programmes to instill an awareness of independent learning, ensuring ongoing progress throughout the four years of posting and beyond.

VII. Outcomes and issues for the future.

For security reasons, the actual test formats and results curves from our pilot test version in April 1995 must remain confidential. Additionally, since we are still very much in the evolutionary stage of development of this test, it is difficult to draw any hard and fast conclusions at this point. However, certain implications arising are worth discussion.

The actual questionnaire results which came back after we had piloted the test not only confirmed the *validity* of our own ideas for authentic scenarios but also provided a rich resource for future scenarios for both testing and training. A very clear picture emerged of the key role the support staff are expected to play in the smooth-running of the missions overseas, and just how versatile their communication skills need to be to meet these expectations. It seems to us with hindsight that we may need to revise our threshold level of minimum competence required from *ESU* Band 4 to Band 5, and review the level of difficulty — though not the actual formats — of the test accordingly.

The first section of the oral test worked effectively in this pilot version as a discriminator to screen out candidates who would not be able to cope with the telephone and face-to-face scenarios. However, the artificiality and limitations of this section have already been discussed, and in future we might consider contextualising this section of the test also, to allow more scope for equality of interactional roles between the examiner and examinee. A possible scenario might be, for example, initiating a conversation with a guest at a Royal Thai Embassy official function.

As well as being pleased — and pleasantly surprised — at how *efficiently* the test worked logistically, we also found that it discriminated effectively between the levels. From looking at the kinds of problems candidates had with the tasks and studying the patterns of inappropriacy they produced in their language output, we were further able to make use of the test as a diagnostic tool for assessing training needs and as a basis for ideas for training programme materials.

At the moment an understanding gap exists between examiners' aims and candidate awareness of what is required. The pass/fail mentality reinforced over time from school days persists in spite of attempts to make the test more achievement-orientated. For example, examinees who were told they did not need to proceed to the next stage of the oral test perceived this as having "failed" the test and suffered damaged morale and self-esteem as a result. The great advantage of

Figure 2 Pre-overseas Posting English Test [POPET] Profile

Test	Description	Scoring
Part I: Reading	candidates extract key information from letter/fax/ memo/order form, etc.	Max 20, converted to ESU Scale Band 4
Part II: Writing	candidates given specific scenario/data and required to write a fax/memo/letter	ESU Scale, max Band 5
Part III: Listening	candidates extract key information from recorded answerphone messages	Max 20, converted to ESU Scale Band 4
Part IV: Speaking	1. candidates interviewed by assessor to talk generally about self/work/Thailand	ESU Scale - candidates attaining less than Band 3 do not proceed
	2. candidates deal with a phone-call typical of the sort received at representations abroad: scenario (A), (B) or (C)	ESU Scale - candidates achieving less than Band 4 do not proceed
	3. candidates deal with a face-to-face scenario typical to overseas work: scenario (A), (B) or (C)	ESU Scale
Aggregate Score		Average ESU Scale Band

Figure 3 Interpretation of Scores

Band 5 and over:	Candidate can function effectively in English at post, though further ESP skills training recommended
Band 4:	Candidate can function adequately in English at post, though further ESP skills training needed
Band 3:	Candidate cannot function adequately in English at post, substantial further general English and ESP skills training required
Band 2:	Candidate cannot function in English at post, extensive further general English training required
Band 1:	Candidate cannot function at all in English, comprehensive general English training required

criterion-referencing as opposed to scoring should be to get away from these artificial yardsticks, assessing performance on a positive continuum of accretion rather than negatively as a deficit from some idealised model. Some work remains to be done to bridge this awareness gap and address this emotional dimension.

The oral testing procedure was, it must be admitted, extremely tiring for the examiners. To be involved in the role-play as a "performer" while at the same time acting as assessor can be a confusing conflict of roles. We cannot be certain yet to have standardised sufficiently among the three of us on the team to ensure *reliability* of assessment, though since we are also using the *ESU* scales in conjunction with our other tests at the Institute, we feel we have come some way towards this elusive goal.

References

Davies, A. 1978. Language testing. Survey article Parts I and II. *Language Teaching and Linguistics Abstracts*, II 3/4.

Hymes, D. H. 1972. On communicative competence. In Pride, B. J. and Holmes (Eds.), *Sociolinguistics*. Penguin.

Morrow, K. E. 1979. Communicative language testing: revolution or evolution. In Brumfit, C. J. and Morrow, K. E. (Eds.), *The Communicative Approach to Language Teaching* (pp. 143-158). Oxford: Oxford University Press.

Oller, W. 1979. *Language Tests at Schools*. London: Longman.

di Pietro, R. J. 1987. *Strategic Interaction: Learning Language through Scenarios*. Cambridge: Cambridge University Press.

Savignon, S. J. 1972. Teaching for communicative competence: a research report. *Audio-Visual Language Journal*, 10/3: 153-62.

Skehan, P. 1988. State of the art article. Language testing Parts I and II. *Language Teaching*, 21/4.

Spolsky, B. Ed. 1968. Language testing: the problem of validation. *TESOL Quarterly*, 2: 88-94.

Weir, C. J. 1990. *Communicative Language Testing*. Cambridge: Cambridge University Press.

Widdowson, H. G. 1978. *Teaching Language as Communication*. Oxford: Oxford University Press.