PAC3
at
JALT
2001
Conference
Proceedings

MENU
Text Version
Help & FAQ

International
Conference
Centre
Kitakyushu
JAPAN
November
22-25, 2001

Enhancing Performance of Communicative Testing Instruments

Richard Blight
Ehime University

Communicative principles have greatly changed the form of assessment procedures over the last three decades. Contemporary theory emphasizes the importance of incorporating authenticity and practicing meaningful negotiations. In addition, test design principles are concerned with establishing specific performance criteria against which students can be assessed. Yet despite careful planning and preparation, teachers frequently encounter difficulties when attempting to implement communicative assessment procedures. Discrepancies occur between design principles and the requirements of practical implementation. In this paper, typical problems are outlined and a framework is developed whereby tests are evaluated against design principles and specific communicative goals. Teachers make professional judgements concerning the relative value of various aspects of a test, with a view to directly enhancing test performance. As a result, a practical communicative testing system is developed according to the contextual requirements of each situation of use.

ここ３０年の間に、様々なコミュニケーション原理により、評価手順の形態が大いに変わった。現代の理論は、評価手順に正当性を取り入れるとともに適度な修正を行うことが重要だと強調している。また、テスト作成原理の関心は、生徒の言語運用能力を測るための特別な言語運用基準を制定することに向けられている。しかし、いかに入念な計画と準備を持ってしても、コミュニケーション評価手順を実行しよう

とすると、教師たちは様々な困難に直面する。テスト作成原理と実際の実行に伴う必要条件との間に様々な矛盾が当然のことのように生じる。本稿では、いくつかの典型的な問題を概説し、しかる後に、様々なテストを、その作成原理とそれぞれ独自なコミュニケーションの目的に照らして評価する際の一つの枠組みを模索する。テスト実行者たちは、直接的にテスト性能を高めることをもくろんで、テストの様々な側面に対する比較的価値に専門的な評価を下す。結果として、実際のコミュニケーションテストシステムは、それぞれの使用状況を取り巻く様々な必要条件に応じて開発されることになる。

In a plenary speech at the 27th Annual International Conference on Language Teaching and Learning, David Nunan discussed recent research into the principles underpinning English language curricula throughout the Asian region. The preliminary outcomes found "universal adherence to principles of communicative language teaching," but qualified the "universal adherence" as occurring at a level of ministerial rhetoric: "All of the countries surveyed paid lip service to communicative language teaching (CLT), and the principles of CLT are enshrined in all of the documents ... However, all of my informants reported a huge gap between ministerial rhetoric and classroom reality" (Nunan, 2001). Chris Candlin, at the same plenary presentation, spoke of the need to establish communicative teaching directly in classrooms at the "grass roots level" in order to realize the formal acceptance of CLT principles, and to effectively develop language education systems throughout the region.

These viewpoints are however not particularly new, and it appears that a number of complex issues have impeded progress in this area over recent years. There are ongoing practical difficulties involved in implementing communicative teaching principles, which have been encountered in many contexts throughout the region. One crucial concern is the problem of developing useful assessment instruments which can be employed accurately and efficiently in a range of typical classroom situations. This paper considers the types of problems which may be encountered with communicative assessment procedures, and develops a framework for enhancing performance of communicative testing instruments. Language tests are evaluated by measuring student performances against established criteria. Teachers make professional judgements concerning the significance and relevance of the performance criteria within their specific teaching program. The framework provides a practical system for developing effective assessment systems in local contexts, and one which is based on attending to the theoretical requirements of both communicative teaching goals and test design principles.

Communicative assessment: discrepancies between theory and practice

The importance of an effective means of *communicative assessment*, when attempting to implement CLT

principles directly in classroom situations, cannot be understated. Well-developed assessment instruments can be utilized to effectively underpin the communicative goals of a language program. The complex challenge frequently handed to teachers in many institutions begins with the requirement to establish practical and effective testing instruments. The range of issues that need to be confronted in this process is often daunting, and causes significant concern even to highly capable teachers with many years of professional experience. Communicative theory emphasizes the importance of using authentic materials and of practicing meaningful communication in realistic social situations (Hedge, 2000; Larsen-Freeman, 2000; Lynch, 1996), but these goals are almost impossible to achieve in English as a Foreign Language (EFL) settings. The social situations commonly found in EFL classrooms are at best derived from a foreign cultural context, and simulated in learning activities. There is also negligible opportunity for students to consolidate their classroom practice in real world situations, although it is generally agreed that this is of fundamental importance. Furthermore, the extent to which it is possible to achieve the goal of meaningful communication remains entirely unclear, given the contextual requirement for simulating foreign interactions in EFL classrooms.

There are also the substantial requirements of contemporary theory concerning test design principles. Effective testing instruments are generally understood as being required to balance three complex and often conflicting goals -- *validity*, *reliability*, and *practicality* (Brindley, 1995; Hughes, 1989; Weir, 1993). Validity is concerned with how well a test measures what it is intended to measure, and is often considered in terms of a number of basic components, including: *content validity* (how well performances demonstrate the specified learning domain), *construct validity* (the extent to which theoretical principles are reflected in the test design), and *face validity* (whether a test appears to measure what is intended). Reliability is concerned with the dependability of results, or the extent to which performance is consistently measured (i.e. would identical results be obtained if the same test was administered to the same students on another occasion?). Two aspects of reliability are generally associated with the assessor, *inter-rater reliability* (agreement between different assessors of the same performance), and *intra-rater reliability* (the same rater assessing the same performance on different occasions). Practicality is concerned with the "cost effectiveness" of implementation, and considers the value of benefits achieved as compared to the effort required in administration and the practical resources available to the teacher within the local context.

Mismatches often occur between the theoretical requirements and practical limitations associated

with each context of use, and these may substantially impede the development of effective assessment procedures. Bachman & Palmer (1996) discuss the tension which commonly exists between different test qualities, and argue that "test developers need to find an appropriate balance among these qualities, and that this will vary from one testing situation to another" (p. 18). Yet qualitative judgements need to be made concerning how such a balance can be determined in each local context of administration. The lack of direct correlation between theory and practice often occurs as specific discrepancies between test-design principles and efficiency requirements, which routinely act to restrict development of useful assessment tools. Workloads imposed on teachers typically include a full schedule of classroom hours, materials preparation, and administrative duties, and allow for limited effort to be directed towards analysing test performances. From a viewpoint of human resource management, it is necessary to provide teachers with substantial support in order to ensure the effective development of quality assessment procedures. The research reported in this paper establishes a framework which could be implemented with a reasonable amount of effort, and which produces effective results in specific classroom contexts.

Developing effective communicative testing instruments

The first stage in the current research project was selection of a communicative language test appropriate for a group of adult intermediate level learners. It would have been possible to develop an original testing instrument, and to establish an associated set of performance criteria to represent specific communicative goals and assessment principles of the language curriculum. However, since this project was particularly concerned with establishing a beneficial process for evaluating and improving tests, a current assessment instrument was instead chosen. The test had a previously developed set of associated performance specifications (Adult Migrant Education Service, 1995). These included *Elements* (essential linguistic features, knowledge relevant to the content, context requirements), *Performance Criteria* (statements about the learner's performance in the language interaction, specified minimum performance for achievement of competency), *Range Statements* (conditions or parameters to be associated with the assessment task), *Evidence Guide* (suggestions for tasks which could be used to assess the competency), *Benchmark Performances* of learners' assessments (accompanied by specific grading information at various levels), and a continuing *moderation* process (assessors participate in routine moderation sessions which provide the

opportunity to compare and discuss complex assessment determinations). Stages of pre-training were first undertaken with the class, whereby learners were introduced to the test format and the expected standards of performance were explained and demonstrated. Students then attempted some trial tests (working according to assessment conditions), and the test itself was subsequently administered.

Evaluating tests to enhance performance in local contexts

Students' performances on the test were then subjected to close analysis. This comprised a type of informal evaluation process, which aimed to gauge the test's performance (Alderson, Clapham, & Wall, 1995) in terms of what were regarded as desirable communicative goals for the language program. Since the test was a communicative writing test, the analysis firstly involved close consideration of aspects pertaining to performance of writing tasks generally: level of difficulty, task clarity, timing, layout, marking system, purpose, degree of authenticity, amount of information provided, familiarity with test format, and uniformity of administration (Weir, 1993). The analysis next considered the specific test criteria: follows convention of layout for formal letter; stages text appropriately; writes paragraphs which clearly express objective information about situations / events; provides information / supporting evidence; substantiates claim, requests action as required; uses appropriate conjunctive links e.g. causal, additive, temporal, conditional as required; uses appropriate vocabulary to reflect topic and politeness / level of formality; and uses grammatical structures appropriately (Adult Migrant Education Service, 1995). Criteria that were either too challenging or too simple for the student population were reviewed. Criteria were also carefully considered in terms of how accurately it was possible to measure student performances, and whether the rater was sufficiently trained to determine what were "appropriate" performance standards, since this description was used repeatedly throughout the test criteria. The situation required simple "yes/no" determinations to be made based upon what in fact amounted to a complex greyscale of performance variations. Areas where student performances were consistently low were considered with a view to providing more substantial pre-teaching prior to subsequent test administrations. The Range Statements were also carefully reviewed in the same light: topic relevant to learner, recourse to a dictionary, approximately one hundred words in length, time limit -- one hour, learner may draft and self-correct before final presentation, may include a few grammatical errors but errors do not interfere with meaning (Adult Migrant Education Service, 1995).

During the test evaluation process, a number of

complex issues were considered. Did the pre-teaching stage achieve its purposes? Was the testing instrument appropriate for the learners, the program, and the context? How representative were the learners of a typical class in the same course? While such questions could not be resolved from the data, it was reasoned that the teacher's insight and professional experience would provide a valuable first step towards resolution. The last question, for example, concerning the *representativeness* of the selected sample, can ultimately only be verified through statistical analysis of large amounts of performance data (Burns, 1997), so as to ensure "the sample is representative of the population and as far as possible not biased in any way" (p. 76). However, this does not preclude informal progress being made based on the teacher's viewpoints. In the current project, the communicative test was found to have performed adequately, but to also have been limited in a number of areas that were targeted for subsequent improvement. Content validity was compromised through presumption of topical knowledge associated with using a word processor. The students' capacity to achieve some of the stated performance criteria (e.g. express objective information, provide supporting evidence) was also reduced by lack of specific topical knowledge (including associated lexical items), which would have greatly improved the quality of responses. It was concluded that vocabulary extension in the given content area should be undertaken prior to subsequent administrations of the test, either as part of the pre-teaching stage, or earlier during the curriculum. Validity was also compromised by unclear question wording, since a technical term used in the task could be puzzling even to native speakers. Face validity was diminished through choice of subject, since some learners may dislike using a word processor, and consequently react negatively to the task. Indeed, informal feedback provided to the teacher directly after the test administration confirmed that some learners were disconcerted by this topic, although it remains unclear as to what extent this affected their test performance. Reliability was also compromised since the task did not clearly specify distinct stages which were expected to be incorporated in the letter, so that the test would be improved by rewording the question to state the task requirements more clearly and in more detail. The test was however determined as being efficient, practical, and straightforward to administer in the learning context.

Enhancing test performance as a perpetual cycle

The testing instrument was modified according to the conclusions of the first evaluation process, and subsequently administered to another group of learners on the same course. Preliminary results are encouraging, and it appears that a number of improvements have been

achieved early in this process. A system for enhancing test performance within local classroom contexts was consequently established as a form of perpetual cycle, whereby modified versions of the test would be evaluated for each subsequent group of learners, and further refinements devised and implemented. Subsequent test administrations will also be further enhanced through incorporation of a formal system for collecting feedback from test takers (Bachman & Palmer, 1996), particularly given the value of the informal feedback received subsequent to the initial administration. While it is not necessary to devote substantial resources to this task, feedback data can be collected easily and could be of significant value: "low-stakes tests can be improved by planning to use them over an extended period of time and collecting feedback on usefulness during each operational administration" (p. 246). A questionnaire is being designed for this purpose to include a rating scale and open-ended questions which will complement the current evaluation process. In the current project, it was found that a number of complex and somewhat subjective judgements were required during early stages of the enhancement process. The framework however appears to be successfully developing effective assessment instruments to serve the communicative goals of the language program, and substantial improvements have now been incorporated into prerequisite instruction and the testing instrument itself.

References

Adult Migrant Education Service. (1995). *Certificate in spoken and written English III*. Sydney: Author.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brindley, G. (1995). Competency-based assessment in second language programs: Some issues and questions. In G. Brindley (Ed.), *Language assessment in action*. Sydney: NCELTR, Macquarie University.

Burns, R. B. (1997). *Introduction to research methods*. Melbourne: Longman.

Hedge, T. (2000). *Teaching and learning in the language classroom*. Oxford: Oxford University Press.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Larsen-Freeman, D. (2000). *Techniques and principles in language teaching*. Oxford: Oxford University Press.

Lynch, T. (1996). *Communication in the language classroom*. Oxford: Oxford University Press.

Nunan, D. (2001). English as a global language. Plenary speech notes presented at the 27th Annual International Conference on Language Teaching and Learning, Kokura, Kitakyushu.

Weir, C. J. (1993). *Understanding and developing language tests*. London: Prentice Hall.