Review Process for Improving the Evaluation of Communicative Activities

Richard Blight
Ehime University

Reference Data:
Blight, R. (2005). Review Process for Improving the Evaluation of Communicative Activities. In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.) *JALT2004 Conference Proceedings*. Tokyo: JALT.

A critical stage in communicative assessment involves the development of appropriate evaluation criteria which function to relate the teaching objectives and curriculum goals to the classroom learning being experienced by students and the instruction being delivered by teachers. This paper describes the process of reviewing a set of evaluation criteria developed for the purpose of evaluating communicative activities in a first-year university English course. The evaluation system was initially produced as a materials development instrument to measure the effectiveness of a range of activities against the curriculum objectives of the first-year program. The next stage in the development of the evaluation system involves examining the effectiveness of the evaluation criteria on a series of trial activities used in the course. The results of a survey providing feedback on the evaluation criteria are also presented and discussed. Suggestions are made for developing the evaluation system based on the combined input from the different sources of information.

コミュニケーション能力評価の正念場には、学生が経験する教室学習と教師が施す指導の教育目的とカリキュラム目標に関係する適切な評価基準の開発が含まれている。本稿では大学一回生英語クラスにおけるコミュニケーション能力の授業活動を評価する目的で開発された一つの評価基準の見直し過程を述べる。当評価システムはまず初めに第一学年プログラムのカリキュラム目標に対する一連の活動の効果を測る教材開発装置として生み出された。評価システム開発の次の段階は、コースにおいて用いられた一連の試験的活動に関する評価基準の効果の検討を含む。評価基準に関するフィードバックを提供する調査結果も同様に示し、検討する。様々な情報源からの複合的投入を基に、評価システム開発のための提案を行なうものである。

Language practice activities play a fundamental role in communicative teaching programs, and the success of a language course is often closely related to the effectiveness of the activities used in the course. The author of a current textbook series, Marc Helgesen, explains the close relationship between activities and the curriculum: "To a large degree, activities—organized around a sequence of functions / grammar points / vocabulary areas, etc.—ARE the curriculum. Activities are what the learners do and that is how they learn" (Bradford-Watts & O'Brien, 2004, p. 13). Teachers consequently need to understand the nature of the activities they choose in their courses and the effectiveness of the activities with different groups of learners. What are the strengths and weaknesses of each activity for specific ability levels? What are the learning objectives and how successfully are those objectives achieved? Does each activity serve to effectively achieve the course goals (Brown, 2001)? What qualities can be regarded as consistently contributing towards the success of activities in similar learning contexts? Insight can be gained into these areas by evaluating the effectiveness of activities in specific classroom environments. However, evaluation results are

themselves open to a range of possible interpretations, and as a consequence it is also necessary to investigate the effectiveness of the evaluation system in order to improve the accuracy of the results in future evaluations.

The present study relates to a research project that aims to develop an evaluation system to accurately measure the effectiveness of the communicative activities used in a first-year university course. The first stage of development involved producing a set of evaluation criteria to measure the performance of the classroom activities (Blight, 2005). The effectiveness of each activity can be determined in terms of how successfully the pedagogical objectives of the course are mapped onto the learning outcomes being achieved by students (Rea-Dickens & Germaine, 1992). In the next stage of development, which is discussed in the present paper, the validity and reliability of the evaluation results are improved by examining the effectiveness of the evaluation criteria (McNamara, 2000). Since no single measure of the effectiveness of the criteria is apparent, the criteria are examined from three complementary perspectives. First, they are considered in terms of how well they are achieved in a series of trial activities from the first-year course. The results of a teacher survey on the importance of the criteria and the difficulty of the rating process are subsequently considered and discussed. The three sources of input provide beneficial directions for developing the evaluation system to produce improved measures of the effectiveness of the communicative activities used in the first-year course.

Evaluation System for Communicative Activities

The evaluation system was developed as part of a materials development project to measure the effectiveness of communicative activities used in a first-year university course (e.g. Brown, 1995; Tomlinson, 1998). Since an internally produced textbook was being used across the first-year program, feedback from a classroom-based evaluation system (Genesee & Upshur, 1996) could assist to determine which activities should be revised or dropped from future editions of the textbook (Nunan, 1989; Richards, 2001). The classroom materials followed a communicative approach designed to improve the components of communicative competence (Brown, 2001; Jacobs & Farrell, 2003; Larsen-Freeman, 2000). An evaluation process was first conducted on the textbook, but this focussed on the lessons in the course and did not provide specific feedback about the activities. The present system was consequently developed as a form of micro-evaluation process (Ellis, 1998) to provide information concerning the relative performance of the activities used in each lesson (Graves, 2000). The evaluation criteria were devised by considering the qualities that tended to result in activities being successful or unsuccessful in the first-year classes. They covered a range of areas including the objective or learning purpose of an activity, aspects of communicative learning that should be incorporated in activities, and other properties that were associated with the successful activities (see Table 1).

Table 1. Communicative evaluation criteria

No.	Evaluation Criteria
C1	Clear learning objective
C2	Learning purpose is useful / beneficial
C3	Involves meaningful communication
C4	Provides practise / repetition of target language forms
C5	Level of learner activation / active participation
C6	Motivation factor / interesting, enjoyable
C7	Personalization / personal experiences, opinions, feelings
C8	Learning challenge / tension
C9	Volume of language production
C10	Appropriate difficulty level
C11	Appropriate pace / rate of progression

The evaluation system was first run on a series of trial activities from the first-year university course. Fifteen teachers rated five classroom activities against the evaluation criteria using a five point Likert Scale (5 = Very Good; 4 = Good; 3 = Satisfactory; 2 = Poor; 1 = Very Poor). The evaluations were analysed from a perspective of identifying the strengths and weaknesses in the activities (see Blight, 2005), so that the materials could be improved for future editions of the textbook (Rea-Dickens & Germaine, 1992). However, a number of problem areas and limitations were recognized in the evaluation system. For example, teachers were likely to rate highly the types of activities that corresponded with their own teaching styles, and to be more critical of the other activities. The teachers could also apply inconsistent standards of measurement against the evaluation criteria. So how could consistency between different teachers, as well as from individual teachers on different activities, be ensured in the rating process (McNamara, 2000)? Furthermore, how effectively did the evaluation criteria respond to the range of learning styles and ability levels apparent in the first-year classes? And how closely did the criteria map the curriculum goals onto the learning being achieved by the students (Brown, 1995)? In order to address issues concerning the accuracy of the evaluation results, it was clearly necessary to examine the effectiveness of the evaluation system and the degree to which it can be relied upon to achieve consistent results.

Criteria Performance on the Trial Activities

The evaluation criteria are first considered in terms of the results produced on the trial activities. Mean scores for the criteria were determined by averaging the scores from the fifteen teachers on the five activities. While this type of analysis cannot provide conclusive results about the evaluation criteria, it can serve to indicate whether problems are evident in relation to specific criteria. However, the results would need to be considered from two perspectives before this type of interpretation could be derived. First, low results against specific criteria are likely to occur with an effective evaluation system, in which case they would indicate weak areas of performance that were common in the activities. This interpretation would lead to considering how the activities can all be improved to perform better against those specific evaluation criteria. Alternatively, and relevant to the current investigation, low results could also suggest problems in rating the criteria, which may have caused the teachers to grade down the results. For example, teachers may have perceived a degree of content

overlap between criteria, or perhaps the definitions of some criteria were not sufficiently clear. These types of problems could contribute to low evaluation results, since they fail to distinguish the specific purposes of the criteria and also cannot serve as validity checks, which would require close correspondence between the two content areas (e.g., with a statement represented in negated form). However, since we cannot attribute causes for the low results on the basis of the present analysis, our investigation is limited to considering low scores in terms of whether they are likely to indicate potential problem areas in the evaluation criteria.

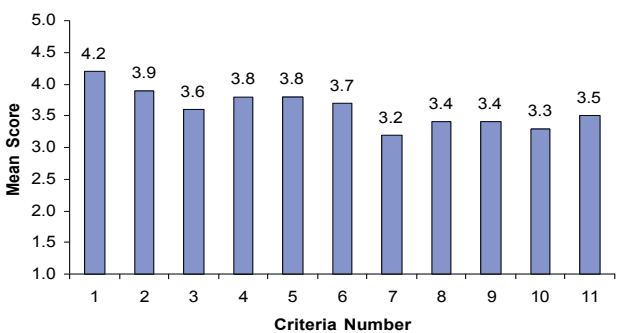

Figure 1. Mean scores for the evaluation criteria

The mean scores from the teachers' ratings of the trial activities are shown in Figure 1. Four criteria (C1, C2, C4, C5) were rated with a score close to 4 points, equivalent to a *good* rating on the Likert scale. Six criteria (C3, C6, C8, C9, C10, C11) were rated around the mid-point between 3 and 4 points (i.e., between *good* and *satisfactory*). One criterion (C7) was rated close to 3 points (i.e., as *satisfactory*). To identify potential problem areas, we should consider why teachers have provided consistently low ratings for C7, C10, C8, and C9. These results could indicate that we should revise all the activities to include more opportunities for students to discuss personal experiences, and check the difficulty level, learning challenge, and volume of language production in the activities. However, it is likely that the low result for personalization could also be a consequence of some activities not being designed to meet this purpose, since at least one activity was rated very low in this area (see Blight, 2005). While personalization is an important communicative objective (Bradford-Watts & O'Brien, 2004), we should consider whether it is a design goal of the first-year course to include this quality in every activity, particularly since this type of goal would exclude other types of potentially valuable activity (e.g. grammar substitution and vocabulary practise activities; see Ur, 1988). If personalization is not necessary in all the activities, this criterion should be omitted from this type of general evaluation framework.

The results on the other three criteria (difficulty level, learning challenge, volume of language output) are also of significant concern. It is not apparent, for example, why the difficulty rating was consistently low, since the trial activities did not appear to be consistently difficult for first-year students (Blight, 2005). This response pattern could instead suggest that teachers found it difficult to rate a *good* difficulty level, and this would appear to be a major problem

in classes with mixed levels of abilities or skills. In the next stage of the study, we will examine the results from a teacher survey to consider how we can gain further insight into these issues.

Importance of the Evaluation Criteria

The teachers who rated the trial activities were also given a short survey asking them to rate the effectiveness of the evaluation criteria in two areas (see Appendix). The first question asked the teachers to rate how important the criteria were in relation to the curriculum objectives for first-year university students. The responses from fourteen teachers were subsequently averaged to produce mean scores for the evaluation criteria (see Figure 2). The teachers rated seven criteria (C1, C2, C3, C4, C5, C6, C10) between 4 and 5 points (i.e., between *important* and *very important*) and the other four criteria (C7, C8, C9, C11) at 4 points (i.e., as *important*). C1 was the highest rated criterion, so a clear learning objective appears to be highly important in all activities. The next highest ratings were for a beneficial learning purpose (C2) and active participation (C5), which are clearly also important in all activities. Perhaps the most interesting observation is that the teachers rated the criteria at two distinct levels, with the first group of seven criteria being rated half a point higher than the group of four criteria. This pattern is useful to the review process since it identifies the criteria (C7: personalization; C8: learning challenge; C9: volume of language production; C11: rate of progression) that should be further considered in terms of their relevance to the language curriculum.

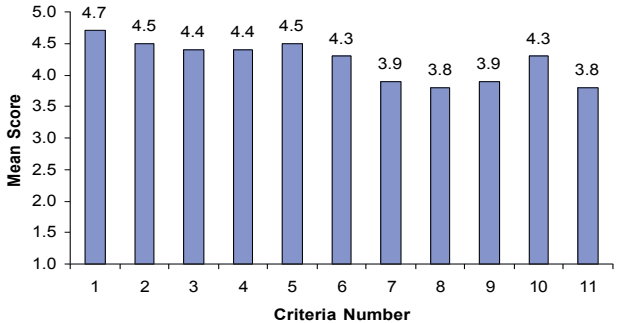

Figure 2. Importance of the evaluation criteria

The next stage of the investigation involves identifying likely causes to explain the survey results (Genesee, 2001). This once again involves a process of interpretation that cannot be considered conclusive (Ellis, 1998), but which can serve to identify general problem areas and hence to provide directions for improving the evaluation system (Hughes, 2003). First, the lower rating for personalization (C7) appears to be consistent with our previous observation that it may not be necessary to include this in all activities, since different types of activity may be relevant to the curriculum goals. There also appears to be a degree of content overlap between learning challenge (C8) and difficulty level (C10), so it may not be clear how teachers should differentiate between these criteria. The third criterion (C9: volume of language production) relates to a valid type of fluency objective, but it is not clear whether producing a high volume of language is sufficient with no

reference to the quality or type of language. While volume of language is clearly important, language content has also been described as an important type of fluency objective (Harmer, 2001; Jacobs & Farrell, 2003). However, if we change the definition to include the type of language (i.e., volume of *target language* production), we produce an overlap with another criterion (C4: provides repetition of target forms).

Different types of factors could be contributing to the lower result for the rate of progression (C11). First, how can teachers rate the pace of an activity in relation to a class of learners when the students each have different ability levels and degrees of motivation? The weaker students would require a slower rate of progression and more response time, while the stronger students could become frustrated by slow responses. Another problem relates to how this criterion should be applied, since the pace of an activity is not directly related to the quality of the activity. Some activities involve a degree of reflection or consideration that is valuable to the students' learning progress and should not be considered detrimental to the activity (Bradford-Watts & O'Brien, 2004). If teachers rate fast activities as *good*, they are placing an emphasis on frequent interactions between partners, which is another type of fluency objective. While it is appropriate for fluency goals to encourage communication between students (Comeau, 1987; Jacobs & Farrell, 2003), it is not clear whether these types of objectives can be applied generally to all activities in the course.

Difficulties with the Evaluation Criteria

The second question on the teacher survey investigated the degree of difficulty in applying the evaluation criteria (see Appendix). Problems with the rating procedure could produce inconsistent results and hence impact negatively on the accuracy of the evaluation process. There may have been a range of difficulties in the rating process, including unclear criteria definitions or degrees of content overlap between criteria, which may have led to subjective interpretations being necessary from the teachers. The responses to this survey item from fourteen teachers reveal a substantially different response pattern to the previous survey question (see Figure 3). The criteria are generally rated much lower, and there is also significant variation in the ratings between different criteria. Three criteria (C1, C4, C9) are rated at the 4 point level (i.e., as *easy*), five criteria (C2, C3, C5, C6, C7) are rated between 4 points and 3 points (between *easy* and *average*), two criteria (C10, C11) are rated at 3 points (*average*), and one criterion (C8) is rated between 3 points and 2 points (between *average* and *difficult*). The lower general level of responses from the teachers is clearly of concern, and we should also consider causes for the teachers rating eight criteria less than *easy* to apply.

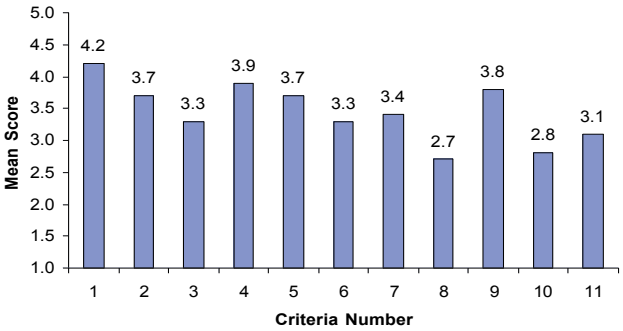

Figure 3. Rating the evaluation criteria

It is likely that several factors are negatively impacting the performance of the evaluation criteria in relation to the rating process. First, several criteria (e.g. C8, C10, C11) may have been difficult to rate on account of classes having students with different levels of language ability and different skill competencies. The other five criteria (C2, C3, C5, C6, C7) appear to involve similar degrees of subjective interpretation in the rating process. How can teachers determine whether the learning purpose is beneficial (C2) to a class configuration, when this varies substantially according to individual students? And how can teachers know what constitutes meaningful communication (C3) between different students? Similar types of problematic subjective measures are also required for teachers to determine the level of activation (C5), the motivation factor (C6), and the degree of personalization (C7) in an activity, which are also likely to produce inconsistency in the rating process.

Combining Three Perspectives of Analysis

The next stage in the review process involves compiling the information from the three perspectives of analysis into a common framework. This stage serves to identify the criteria that have been consistently rated as problematic and which should be revised prior to the next evaluation. For this purpose, a necessary performance standard was determined for each analysis, with those criteria rated below the standard being indicated for review. The standard was set at 3.5 points (i.e., between *good* and *satisfactory*) for the results with the trial activities and the difficulty of applying the criteria. The ratings for the importance survey were expected to be generally higher for all the criteria, so the standard was set correspondingly higher (at 4.0 points or *good*) for this analysis. Based on these standards, a summary of the results for the evaluation criteria against the three systems of analysis is next derived (see Table 2). The combined data analysis shows that four criteria (C1, C2, C4, C5) have achieved the performance standards in all three perspectives of analysis. Two criteria (C3, C6) have achieved the standards in two perspectives, and three criteria (C9, C10, C11) have achieved the standards in a single perspective. The remaining criteria (C7, C8) have missed the performance standards in all three categories.

Table 2. Combined data analysis

No.	Evaluation Criteria	Trials	Value	Rating
C1	Clear learning objective	O	O	O
C2	Learning purpose is useful / beneficial	O	O	O
C3	Involves meaningful communication	O	O	X
C4	Provides practise / repetition of target forms	O	O	O
C5	Level of learner activation / active participation	O	O	O
C6	Motivation factor / interesting, enjoyable	O	O	X
C7	Personalization / experiences, opinions, feelings	X	X	X
C8	Learning challenge / tension	X	X	X
C9	Volume of language production	X	X	O
C10	Appropriate difficulty level	X	O	X
C11	Appropriate pace / rate of progression	O	X	X

Conclusions

The review process conducted in this study has examined the evaluation criteria according to three complementary perspectives: their performance in the trial activities, the importance of the criteria, and difficulties in the rating process. Since four evaluation criteria (C1, C2, C4, C5) achieved the performance standards in all three forms of analysis, it is recommended that they should progress directly to the next stage of the evaluation system. Conversely, the two criteria (C7, C8) that missed the performance standards in all three investigations should be dropped from the evaluation system. The other criteria received mixed results according to the different investigations and should be revised for future versions of the evaluation system. Specific revisions to each criterion should be considered based on the interpretations provided for likely causes of the low results. Suggestions can also be made for prioritising the revisions based on the combined data analysis (see Table 3).

Table 3. Revisions to evaluation criteria

No.	Priority	Action	Suggested Revision
C3	*	Revise	Improve measurement
C6	*	Revise	Improve measurement
C7	***	Drop / omit	--
C8	***	Drop / omit	--
C9	**	Revise	Improve relevance
C10	**	Revise	Improve measurement
C11	**	Revise	Improve measurement Improve relevance

References

Blight, R. (2005). Evaluation system for communicative learning activities. In T. Newfields & Y. Ishida (Eds.), *2004 JALT Pan-SIG conference proceedings*. Tokyo: JALT Publications.

Bradford-Watts, K., & O'Brien, A. (2004). Interview with Marc Helgesen. *The Language Teacher, 28* (9), 13-16.

Brown, H. D. (1995). *The elements of language curriculum: A systematic approach to program development*. Boston: Heinle & Heinle.

Brown, H. D. (2001). *Teaching by principles: An interactive approach to language pedagogy*. New York: Pearson Education.

Comeau, R. F. (1987). Interactive oral grammar exercises. In W. M. Rivers (Ed.), *Interactive language teaching* (pp. 57-69). Cambridge: Cambridge University Press.

Ellis, R. (1998). The evaluation of communicative tasks. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 217-238). Cambridge: Cambridge University Press.

Genesee, F. (2001). Evaluation. In R. Carter & D. Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages* (pp. 144-150). Cambridge: Cambridge University Press.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Graves, K. (2000). *Designing language courses: A guide for teachers*. Boston: Heinle and Heinle.

Harmer, J. (2001). *The practice of English language teaching*. Harlow: Pearson Education.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jacobs, G. M., & Farrell, T. S. (2003). Understanding and implementing the CLT (communicative language teaching) paradigm. *RELC Journal: A Journal of Language Teaching and Research in South East Asia, 34* (1), 5-30.

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

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

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

Rea-Dickens, P., & Germaine, K. (1992). *Evaluation*. Oxford: Oxford University Press.

Richards, J. C. (2001). *Curriculum development in language teaching*. Cambridge: Cambridge University Press.

Tomlinson, B. (1998). *Materials development in language teaching*. Cambridge: Cambridge University Press.

Ur, P. (1988). *Grammar practice activities: A practical guide for teachers*. Cambridge: Cambridge University Press.

Appendix

Teacher Survey on the Evaluation Criteria

Please rate the following two questions on the same 5-point scale used for rating the activities

(5 = Very Good; 4 = Good; 3 = Satisfactory; 2 = Poor; 1 = Very Poor):

a) How important are the evaluation criteria as curriculum objectives for first-year university students?

b) How easy was it to use the evaluation criteria to rate the classroom activities?

No.	Evaluation Criteria	Value	Usage
C1	Clear learning objective		
C2	Learning purpose is useful / beneficial		
C3	Involves meaningful communication		
C4	Provides practise / repetition of target language forms		
C5	Level of learner activation / active participation		
C6	Motivation factor / interesting, enjoyable		
C7	Personalization / experiences, opinions, feelings		
C8	Learning challenge / tension		
C9	Volume of language production		
C10	Appropriate difficulty level		
C11	Appropriate pace / rate of progression		