# Adapting the Canadian Language Benchmarks for Writing Assessment

*Timothy Stewart, Sally Rehorick, and Bill Perry*

*The purpose of this article is to describe the development of an instrument for assessing the writing development of students in an English-medium university in Japan. We begin with a description of the setting of the college and the unique nature of its program. Next we discuss the process of selecting a language proficiency framework suitable for the four years of the degree. The Canadian Language Benchmarks (Citizenship and Immigration Canada, 1996) were chosen and subsequently formed the basis for the development of the rating scale. The process of developing the scale held a number of challenges, given the target population and the requirement to have an instrument usable by both language development specialists and nonspecialists. Issues such as the institutional context, the framework for evaluating language development, and development and refinement of the assessment scale over the first two years of the project are discussed.*

*Le but de cet article est de décrire le développement d'un instrument pour l'évaluation de la compétence à l'écrit d'étudiants dans une université d'expression anglaise au Japon. L'introduction consiste en une description de l'emplacement du collège et du caractère unique du programme. Suit une discussion du processus qui a mené à la sélection d'un modèle approprié aux quatre années que dure le programme. Les Niveaux de compétence linguistique canadiens (1996) ont constitué la base à partir de laquelle l'échelle d'évaluation a été développée. Compte tenu de la population cible et de la nécessité d'avoir un instrument que pourraient employer tant des spécialistes en acquisition du langage que des non-spécialistes, le développement de l'échelle a présenté quelques défis. L'article se termine par une discussion de questions telles que le contexte institutionnel, le modèle pour l'évaluation du développement langagier et le développement et la mise au point de l'échelle d'évaluation pendant les deux premières années.*

## *Setting*

The professional and institutional context where this project was undertaken is unique. The institution, Miyazaki International College (MIC) in Kyushu, Japan, is a small four-year liberal arts college. It is an English-medium institution accredited by the Japan Ministry of Education (*Mombusho*). The university's mission is to develop students who are fluent in English and Japanese and who can employ critical thinking skills; the students acquire

Japanese and who can employ critical thinking skills; the students acquire these skills through pursuing a liberal arts degree. To meet this challenge the college has one of the greatest concentrations of English-speaking faculty of any postsecondary institution in Japan (Otsubo, 1995). All of the faculty speak English, and 80% are non-Japanese (although not necessarily native speakers of English) compared with fewer than 2% in the entire Japanese system of higher education. All courses, except those in Japanese expression and teacher education, use English as the language of instruction. In addition, in order to facilitate the use of active and cooperative learning techniques (Bonwell & Eison, 1991; Johnson, Johnson, & Smith, 1991; Kessler, 1992; Sutherland & Bonwell, 1996), the college administration makes a concerted effort to keep class sizes small (10-20 students).

First- and second-year credit-bearing courses in content areas such as sociology, economics, and art history are team taught by EFL faculty and discipline specialists. In collaborative team teaching, pairs of language and content-area faculty work together extensively; they co-plan syllabi and lessons, and teach together in the same classroom. Students concurrently learn content concepts, language, and critical thinking skills.

Wesche (1993) describes three prototype curriculum models for discipline-based instructional formats: (a) theme-based, which are principally language courses organized around a series of topics or themes, but which are not regular academic discipline courses (although they might usefully be an introduction to these); (b) sheltered courses, which are regular academic courses in disciplines such as economics or political science and limited to students whose native language is other than the language of the course; and (c) adjunct courses, in which "a specially tailored second/foreign language course for advanced L2 speakers is organized around the content and language needs arising from a selected discipline course" (p. 61).

The model used at MIC shares some characteristics of all three of these, although it is not strictly speaking an example of any. In the first two years of study students take a limited number of courses that focus on the development of English proficiency skills. They also take regular academic courses (e.g., economics, sociology) team taught by a discipline specialist and a language specialist. In some ways the structure resembles the sheltered model as all the students in the classes are non-native speakers of English. In other ways the program more closely resembles the adjunct model inasmuch as instruction in the linked but separate language course is "organized around the content and language needs arising from a selected discipline course" (Wesche, 1993, p. 61). However, at MIC language is taught in subject courses alongside discipline content during a single class period facilitated by both instructors. In other words, the assumption is not language before content or vice versa, but rather simultaneous learning of both. In this way the instructional delivery model is closest to the "four-handed" approach

described by Corin (1997) and Stryker (1997) in which two instructors co-teach in a classroom. In four-handed teaching two instructors interact in a complementary fashion to facilitate learning and to maximize the exposure of learners to native speakers engaged in natural communication.

In the third and fourth years the courses are taught only by the discipline specialists. By this point students can generally cope quite well with the language of the specific content areas, although they still need help with writing essays. This is particularly so for the senior thesis, and the discipline specialists often call on the language specialists to sit on thesis supervisory committees to help the students with their writing.

Our students' educational backgrounds do, however, present some obstacles to achieving success in this model. First, the incoming students have widely varying levels of English language proficiency, yet they generally take the same classes. Second, in high school Japanese students have limited experience communicating in English. Their secondary education has been predominantly passive in nature and their exposure to academic class work in English has been virtually nonexistent (Nozaki, 1993; Rohlen, 1983). For these reasons our challenge at MIC has been to meet the needs of individual students while delivering college-level courses. With such a heterogeneity of language levels among the students, we had to devise teaching methods to activate their essentially passive knowledge of English so that they could discuss ideas critically in the classroom. One of the goals of MIC is to produce graduates who are fluent in English and Japanese and who have an understanding of international issues.

### *Selecting a Framework for Evaluating Language Development*

Since the beginning of the instructional program at MIC, assessments of progress in students' language acquisition have been conducted mainly through the measures of the TOEFL and TOEIC tests. Analysis of these scores indicates considerable improvement of students' English proficiency over the course of the four-year program. However, these standardized tests are of limited use in the college itself. In short, the TOEFL and TOEIC are measures quite removed from the ongoing work of instructors in their classes. The scores do not relate to the particular model of language competence around which the curriculum is structured, nor do they provide sufficient diagnostic information to allow for specific analysis of students' progress in writing. Indeed, the sole use of tests such as the TOEFL and TOEIC may be misleading according to Savignon (1992) who maintains that "a desire to quantify, to objectify, to render absolute, coupled with incomplete understanding of basic concepts in language and its measurement, leads to assertions of truth rather than acceptance of half-truths" (p. 52). Nunan (1992) reviewed studies that evaluated collaborative classrooms like those found at

MIC and concluded by questioning the use of standardized tests as a principal measure of student achievement in such situations.

MIC faculty indicated their need to understand language development in our students and to have a common discourse, accessible by both language and content faculty, for discussing this development.

In January 1996 the Dean of Faculty asked the Language Area Coordinator to convene a committee to research various measures of language proficiency and to make recommendations to the faculty. After some initial explorations, the committee agreed that the unique institutional context called for institutionally developed measures. Therefore, the committee began with a project to create a set of language benchmarks specific to MIC. The process was an iterative one: the committee drew up an initial document, incorporated suggestions from the faculty as a whole, revised the document, obtained more suggestions, revised it again, and so forth. Before the third round of input from faculty was solicited for this project, committee members realized that it would not be possible to describe the L2 development of our students without a longitudinal, empirically based study. Committee members agreed that such a study would have to track a cohort of students through their four years at the institution. Unfortunately, we did not have that luxury of time. The faculty committee, which had originally rejected the notion of adopting an existing assessment framework, decided to reconsider and consequently examined several scales of proficiency and models of language competence. The committee eventually recommended the Canadian Language Benchmarks (CLB, Citizenship and Immigration Canada [CIC], 1996) as the instrument most relevant to the needs of the institution.

## *Choosing the Canadian Language Benchmarks*[1]

The publication of the *Canadian Language Benchmarks* (CIC, 1996) represented a major initiative in Canada to provide non-English-speaking newcomers to the country with effective English-language learning opportunities. The purpose of the CLB is stated as follows: the benchmarks "provide consistency of outcomes for learners across the country, a common basis for both learner and program assessment and a concrete statement of language competences to all stakeholdes including learners, educators, emplyers and community and settlement agencies" (p. 1). Unlike other proficiency scales such as those developed by the American Council for the Teaching of Foreign Languages (ACTFL) and Educational Testing Services (see Hughes, 1991, and Liskin-Gasparro, 1987, for descriptions of such scales), the CLB includes not only a set of competences but also a set of specifications that define the conditions under which a particular competence might be best demonstrated. In other words, each of the 12 benchmarks in the three skill areas of listening/speaking, reading, and writing describes the competences in the contexts of perfor-

mance conditions, situational conditions, sample tasks, and background knowledge, which are key to the performance of a particular task. Although the benchmarks are hierarchical in structure, the developers of the CLB caution that this structure "does not suggest, however, that certain functions, or 'competencies' are more difficult than others" (CIC, 1996, p. 3). In fact they specifically differ from the ACTFL scale in that language functions are not seen to be more or less difficult in themselves, but rather only in relation to how the particular function is handled by the individual: "functions that one may be tempted to consider as 'higher' or more complex, like suasion ('getting things done'), can indeed be realized through very simple linguistic means, if one chooses to do so. Ultimately, it is the linguistic forms chosen by the speaker to perform the function with and the context that determine relative complexity of the communication" (p. 3).

MIC was attracted to this underlying principle of the CLB. In addition, the CLB was adopted by the institution for the following reasons: First, the CLB reflects language in real use. The generic nature of the benchmarks means that they describe language development generally and are not tied to a specific context. Second, meeting the individual needs of our learners is important given the heterogeneity of the language levels of the student population. The faculty felt that the detailed descriptions of usage for each benchmark would enable individual progress to be tracked. Third, faculty members strongly wished to retain their individuality in teaching and noted that the selection of sample tasks provided in the document, along with the generic nature of the benchmark descriptions, make it a descriptive rather than a prescriptive guide. Fourth, the format of 12 benchmarks within three broad-level bands in three language skill areas is easily understandable by both language and content faculty as well as by students. Fifth, the CLB was seen to be extremely useful in curriculum planning, an ongoing challenge for the experimental college. Finally, and probably most important, the CLB document could provide the MIC faculty with a common discourse to discuss student growth and ultimately to have a positive washback on the MIC program as a whole. Adopting the CLB would "establish a frame of reference that can describe achievement in a complex system in terms meaningful to all the different partners in or users of that system" (North, 1993, p. 6). This frame of reference provided by the CLB eliminated, therefore, any further consideration of adopting other testing scales such as the TOEFL Test of Written English (Sharpe, 1999).

The use of testing systems designed specifically for washback, or feedback, on the classroom is not a new concept. Indeed, as Shohamy (1993) points out, "Few devices are as powerful, or are capable of dictating as many decisions, as tests" (p. 1). Researchers such as Hamp-Lyons (1991) note the rich diagnostic information that can emerge from well-designed testing programs. Rehorick and Dick (1996) extend the notion of washback to include

decision-making processes in and among organizations. In other words, the impact of tests is not only on the learners themselves, but can include curriculum design, instructional methods, program structure, and organizational development. That MIC recognized the potential for positive washback on its program is reflected in the following quote from an internal research memo: "It is hoped that the writing sample, obtained under similar conditions for all students, will be the first step in a variety of studies that will help develop an institutional profile of MIC students and will serve to provide a baseline for curriculum revision and further research" (Miyazaki International College, 1997).

### *Using the CLB as the Basis for an Assessment Scale*

The committee recognized the advantages in using the Canadian Language Benchmarks as the basis for an assessment scale because of the potential for establishing a solid scaffold to support other areas of the MIC system. North (1993) has noted that this kind of proficiency scale can "provide coherent internal links in a system among pre-course or entry testing, syllabus planning, materials organization, progress and exit assessment, and certification" (p. 6). The committee felt that faculty members would be much more likely to embrace the assessment scale if they first had an opportunity to use it for course-planning, task design, and student advising. Thus faculty were encouraged to use the document for gathering exemplars of student work that represented the various benchmarks. Within a few months it was recommended that the CLB writing descriptors be adapted to form the measurement scale. Writing was chosen for this first college-wide assessment effort because it is arguably the easiest skill to test and the most representative of the academic skills being emphasized in MIC's academic curriculum.

The challenge of using the CLB as an assessment tool was that it was designed initially to describe language development, not to evaluate it. "By itself," the developers caution, "it does not measure the learner's proficiency in communicating in English" (CIC, 1996, p. 4). In order to use the CLB as an assessment tool, we would need to develop "a well-calibrated testing instrument with specific test tasks, procedure and the rating scale [to] be used in conjunction with the CLB descriptions to assess the learner's proficiency and place her/him on the continuum of increasing competencies" (p. 4). The design of the rating scale and the training of the assessors were informed partly by research in holistic scoring methods (White, 1985). Hamp-Lyons (1991) notes that "in holistic scoring, each reader of a piece of writing reads the test rather quickly ... and assigns the test a single score for its writing quality [usually] ... by reference to a scoring guide or rubric" (pp. 243-244).

Hamp-Lyons (1991) is critical of holistic scoring because of its limitations in providing in-depth diagnostic information for feedback onto the system. She says that she is "increasingly coming to view this as a severely limiting

feature of holistic scoring, and to demand a richer definition of a 'valid' writing assessment" (p. 244). Her research into primary-trait assessment and multiple-trait assessment provided the basis for the design of the MIC scale. In primary-trait scoring, "test developers decide on a narrowly specified task to assess whatever facet of writing competence has been identified in the context as the most salient and develop a scoring guide specifically for it, focusing only on, for example, whether the writer can take and support a position on an issue" (p. 246). Thus a writing sample can only be rated based on the context in which it is written. In multiple-trait scoring, samples are rated on more than one specified trait according to the criteria of language development.

In MIC's case these criteria are based on the 12 benchmarks of the CLB document. The advantages of the multi-trait approach are that it "implies a view of writing as a complex and multifaceted activity, and of the response of each reader to text as similarly complex and multifaceted" (Hamp-Lyons, 1991, p. 248). Hamp-Lyons stresses the gains to be made through discussions among raters about scores that do not focus solely on the score itself (as in holistic scoring), but rather on the sharing of ideas about a particular writing sample through the language of the assessment instrument. The committee felt that using the benchmarks would be an important step toward opening a dialogue among faculty regarding our specific expectations for students' writing. It was hoped that this dialogue based on the benchmarks would move into other areas of assessment such as listening, speaking, and reading, as well as into our work in curriculum development. The CLB provides a rich scaffold for matching course content and activities to the benchmark descriptions.

### *The MIC College-Wide Writing Assessment*

In the spring of 1997 the language faculty at MIC unanimously voted to support a project to undertake immediately a college-wide writing assessment. The language faculty also requested that an assessment team be assembled following the administration of the college-wide writing sample to develop a rating scale, rate the students' writing samples, and outline future directions for institutional assessment.

One of the challenges of this kind of assessment is to construct a prompt that would elicit a valid sample of the students' writing ability. For the first MIC college-wide writing sample, a reading on the topic of male and female roles in the Japanese home was written and adapted to an appropriate level for the students. Students were asked to write their ideas about the roles of men and women in the Japanese home 10 years from now. This topic was developed because the personal experiences of all of the students would include direct knowledge of the well-defined gender differences in most Japanese homes.They were instructed to use ideas from the reading and from

their own experience to formulate their responses to the question (see Appendix A for the full text of the prompt).

Two hundred, twenty writing samples were obtained from students across the four years. This number (the "writing sample" group) was 83% of the entire student population at MIC, and the distributional characteristics of this group were nearly identical to those of the entire population (see Appendix B for the details).

### *First Steps in Developing an Assessment Scale Based on the CLB*

In the fall of 1997 an assessment team[2] composed of six experienced language specialists was created. The team's first task was to develop a writing scale based on the CLB. This seemingly straightforward task marked the beginning of weeks of debate regarding both the content and format of the proposed scale. It became immediately clear that there was a wide range of beliefs about the specific elements that should be selected from the CLB document for the scale. There were also divergent philosophical views among team members about the best way to approach the development of a scale.

In order to get the development task underway, the team members agreed to choose descriptors that seemed appropriate to the writing task from the CLB at each of the writing benchmarks. The evolution of one benchmark (Benchmark 5) is used to illustrate the lengthy and often contentious process of developing our writing assessment scale. In Table 1 two initial renderings of Benchmark 5 are presented. Both were based on descriptions from the CLB working document (CIC, 1996), but each introduced ideas about writing at Benchmark 5 that were not specifically mentioned in the document (see the underlined phrases). Draft A in Table 1 has added the idea of "introduction, body and conclusion," probably a substitution for "beginning, middle and end" in the original document. In the same table Draft B sees the idea of "topic" introduced, along with the concept of "paragraph form" defined as "topic sentence" with "support and conclusion." The length of the text in the CLB description of Benchmark 5 is given as "100 words" (CIC, 1996, p. 61), whereas Draft B specifies "one to two paragraphs." These seemingly minor differences in the description of the benchmark led to extended and often heated discussion of what a writer at Benchmark 5 should be able to do.

The next step in the development process is presented in Table 2. At this stage the ideas from Table 1, Draft B above concerning topic, paragraph form, and length of text were accepted by the group, and the general categories shown in the left column, "response to prompt," were added to the scale.

After general agreement on the descriptions for all of the benchmarks in Draft #2 was achieved, the assessment team began using this scale to rate actual writing samples. Extended group discussion continued on each of the

Table 1
Two Initial Drafts of Benchmark 5

| *Draft A* | *Draft B* |
|---|---|
| • Conveys ideas clearly. Organizes text with introduction, body and conclusion with proper sequencing of events. | • Conveys ideas about general topic clearly in conventional paragraph form: topic sentence, support and conclusions are evident. Length one-two paragraphs. |
| • Linguistic means of expression remain simple: compound sentences, present tense. | • Linguistic means of expression remain simple: compound sentences, present tense. |
| • Frequent errors in accuracy and awkward sounding phrases. | • Frequent errors in accuracy and awkward sounding phrases. |

samples rated, and changes in the scale were made as necessary. In Table 3 we can see the continued evolution of Benchmark 5 in its sixth draft.

In this draft several key descriptions from the original CLB document were included. The idea of "beginning, middle and end" was reintroduced, and "present perfect" was added. The heading "initial competence—medium complexity" from the Benchmark 5 description was added as well. The group also chose to add "L1 borrowings" from the original document. These changes, based on the building of group consensus, marked an important shift from a "paragraph" and "topic sentence" oriented description to one that more closely approximated the concept of developing text described throughout the benchmark sequence.

By the time the assessment team had completed the seventh revision of the scale based on group rating, time pressure to complete the rating of the writing samples curtailed the extended discussion. We needed to produce a writing assessment scale and begin scoring the samples. The team decided that the entire assessment scale should fit onto one page in order to make it easier to use. This decision necessarily limited the amount of descriptive information that could be listed for each benchmark because all 12 were to be presented on a single page. It was also decided that shading should be used in the one-page format to indicate the groupings of the four benchmarks at

Table 2
Benchmark 5: Draft #2

| | |
|---|---|
| **Response to the Prompt; Content; Organization and Development** | Conveys ideas about the topic, in general, within conventional paragraph form: topic sentence, support and conclusion are evident. Sample length one-two paragraphs. |
| **Vocabulary and Structure** | Linguistic means of expression remain simple: compound sentences, present tense.<br>Frequent errors in accuracy and awkward sounding phrases. |

Table 3
Benchmark 5: Draft #6

| | *Initial competence—medium complexity* |
|---|---|
| **Response to the Prompt; Content; Organization and Development** | Produces coherent text, beginning, middle and end with appropriate sequencing of events.<br>Does better with controlled writing contexts—not creating text freely. |
| **Vocabulary and Structure** | Linguistic means of expression remain simple: compound sentences, present tense, some present perfect.<br>Frequent errors in accuracy and awkward sounding phrases, and L1 borrowings. |

each of the three proficiency stages: *basic, intermediate,* and *advanced.* The levels in each of the stages (*initial, developing, adequate,* and *fluent* competence) were labeled as well (see Appendix C for the scale used in the rating).

Revision of the descriptors in each of the categories continued throughout the rating period. As writing samples were discussed and rated, changes in the descriptors were suggested in order to describe the bases of our judgments more accurately. The only change made in the two basic categories for each benchmark was removing "response to the prompt" from the categories and integrating it into the individual benchmark descriptors.

## *Applying the Writing Assessment Scale: Norming and Scoring*

After the format and content of the assessment scale had been determined, the assessment team began using the scale to rate the writing samples. Each of the several scoring sessions began with calibration among the scorers. Therefore, every scorer read the 12 or so samples selected at random by the facilitator. All the scores were then recorded on a whiteboard, and each paper was discussed in detail before a final score was agreed on by the team. Once the agreement among scorers was consistent, scoring of samples began in earnest. However, using the scale in the first few rating sessions was awkward for some team members, and so additional norming was done as necessary. Twenty percent of the writing samples were rated by all six scorers for the purpose of calibration. Samples were scored double-blind with exact score agreement between readers required. As Table 4 below shows, for over 75% of the ratable papers, not including those rated by the entire group, no more than three readers were required to assign a benchmark based on the scale.

## *Refining the Assessment Scale: The Second Year*

Before the second year of use, minor changes in wording were made to the writing assessment scale. Once again the scoring team was composed of six

Table 4
Rating Statistics: 1997

| *Readers* | *N* | *%* | *Cumulative %* |
|---|---|---|---|
| 2 | 73 | 42% | 42% |
| 3 | 60 | 35% | 77% |
| 4 | 32 | 18.5% | 95.5% |
| 5 | 7 | 4% | 99.5% |
| 6 | 1 | 0.5% | 100% |
| Total 2+ method | 173 | 100% | |
| Group-Rated | 44 (20%) | | |
| Unratable | 3 | | |
| Total Sample | 220 | | |

members: four language specialists and two subject specialists. To familiarize scorers with the instrument at the outset, the group facilitator selected a range of writing samples from the 1997 papers. After considering the scale, team members scored these papers and discussed how they determined the benchmark they assigned for each paper. The facilitator directed team members to consider the general characteristics of these papers as they related to the benchmark descriptors used in the scale. The scoring procedure itself was the same as described above. The key discussions in the norming sessions were always on how to distinguish between, for example, a Benchmark 3 sample and a Benchmark 4 sample.

The scale was not changed during the assessments for the second year of the project. A new writing task prompt was drafted and approved by the college committee that oversees testing. The procedures followed the status quo set in the first year. Each rating session began with extensive calibration, and the reasoning behind scores assigned to these group-assessed samples was discussed, often at length. The results of this writing assessment were comparable to those produced the year before. In the end, 18% of the samples were group rated for the calibration discussions.

## *What We Have Learned*

The faculty members who have been involved in this development process have learned first hand that the development of assessment scales is a complex undertaking. Individual teachers each have their own set of beliefs about language development and assessment. Furthermore, language instructors have experience in writing assessment, and each teacher seems to have his or her own view of how writing should be assessed. By organizing

a team of experienced language specialists to develop a new writing assessment scale and use it, we learned just how difficult such a task could be.

A major challenge proved to be achieving consensus among members of the assessment team concerning the contents of the scale and its application. Every member of the team had the same goal in mind. Nevertheless, in applying his or her own experience to the development process, a great divergence of opinions emerged. Our experience in developing this scale has reinforced the importance of open discussion in the process of rating writing samples. Scoring must be calibrated, and this procedure must involve the group in an exploration of both how they are scoring and which traits of a rating description a sample contains to justify its score. Through these discussions we are able to clarify the language used in the scale for individual benchmarks as we build a group understanding of its use.

We also learned that it is essential to have consistency in the composition of assessment teams. For a variety of reasons, only two members of the original assessment team participated in the group the following year. Although the results were similar for the two years, this lack of consistency was clearly problematic early on in terms of the dynamics of the group.

A related learning is the need to ensure consistency with the prompt used for the writing task. The second assessment team was careful to write a prompt about a social concern at an appropriate linguistic level, but some faculty members later criticized the choice of prompt. Part of the concern was the length of the prompt. Members of the faculty expressed the view that the prompts used for the first two college-wide writing assessments were too long and, therefore, required too much time for some students to read and comprehend.

Finally, we learned that the development of an assessment scale, including its use and revision, must be from the outset viewed as a long-term process. Faculty and administrators must have the patience to examine the results of assessments over time. There also needs to be a willingness to revise assessment scales to make them accessible to all potential users.

## *Next Steps*

Immediately following the 1998 scoring, the assessment team reviewed the scale with insights gained from the experience of using it. Criticism of the sheer number of benchmarks (12) had been raised the previous year and was heard again during the 1998 scoring sessions. Directly related to this were the concerns of scorers about the similarity of the descriptors used for numerically close benchmarks. Out of discussions on these matters, the team revised the original scale from 12 to six benchmarks. The new six-point scale[3] (Appendix D) contains two benchmarks in each of the three stages (basic, intermediate, advanced). Thus the key descriptors used in the CLB for Benchmarks 1 and 2 were combined to form the rating of 1 on the new scale.

The descriptors remain true to those found in the CLB document with the minor addition of wording related to the severity of errors and their impact on understanding for the reader.

The streamlining of the scale was done in the interest of promoting wider use of the benchmarks at the college. The 12-point scale, like the CLB document, was widely seen by language and discipline specialists alike as being too cumbersome to use as a teacher's aid. The development of a six-point scale was a reaction to this concern. In July 2000 the revised scale in Appendix D was used by 16 raters to score writing samples. To our amazement and delight, no complaints were voiced about the scale being at all awkward or cumbersome. Of course, use of the scale has been limited to date, but over time we will see if this revised scale meets the needs of the college and the faculty.

## *Notes*

[1]The Canadian Language Benchmarks is a document widely available for consultation by readers.

[2]The other members of the assessment team have made invaluable contributions to this project. We would like to acknowledge the hard work of Judy Gallian, Michael Sagliano, and Margaret Simmons.

[3]The authors are indebted to Liz Hamp-Lyons for her comments on problems with the wording used in the original scale.

## *The Authors*

Sally Rehorick and Tim Stewart are founding faculty of Miyazaki International College (MIC) in southern Japan. Both Bill Perry and Tim Stewart are faculty members at MIC. Sally Rehorick is currently a professor in the Faculty of Education at the University of New Brunswick and is Director of the Second Language Education Centre.

## *References*

Bonwell, C., & Eison, J. (1991). *Active learning: Creating excitement in the classroom* (ASHE-ERIC Higher Education Report No. 1). Washington, DC: George Washington University.

Citizenship and Immigration Canada. (1996). *Canadian language benchmarks: English as a second language for adults/English as a second language for literacy learners*. Working Document. Ottawa, ON: Minister of Supply and Services Canada.

Corin, A. (1997). A course to convert Czech proficiency to proficiency in Croatian and Serbian. In S.B. Stryker & B.L. Leaver (Eds.), *Content-based instruction in foreign language education: Models and methods* (pp. 78-104). Washington, DC: Georgetown University Press.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-278). Norwood, NJ: Ablex.

Hughes, A. (1991). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.

Johnson, D.W., Johnson, R.T., & Smith, K.A. (1991). *Cooperative learning: Increasing college faculty instructional productivity* (ASHE-ERIC Higher Education Report No. 4). Washington, DC: George Washington University.

Kessler, C. (Ed.). (1992). *Cooperative language learning: A teacher's resource book*. Englewood Cliffs, NJ: Prentice-Hall Regents.

Liskin-Gasparro, J. (1987). *Testing and teaching for oral proficiency.* Boston, MA: Heinle and Heinle.

Miyazaki International College. (1997). *Internal Memo, June 9, 1997, College-wide in-class writing sample.* Miyazaki, Japan: Author.

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

Nozaki, K.N. (1993). The Japanese student and the foreign teacher. In P. Wadden (Ed.), *A handbook for teaching English at Japanese colleges and universities* (pp. 27-33). Oxford, UK: Oxford University Press.

Nunan, D. (Ed.). (1992). *Collaborative language learning and teaching.* New York: Cambridge University Press.

Otusbo, H. (1995) Japan's higher education and Miyazaki International College: Problems and solutions. *Comparative Culture: The Journal of Miyazaki International College, 1,* 1-11.

Rehorick, S., & Dick, J. (1996). Test development as transformational process: The case of the Maritime Oral Communication Assessment Portfolio. *Comparative Culture: The Journal of Miyazaki International College, 2,* 42-56.

Rohlen, T.P. (1983). *Japan's high schools.* Berkeley, CA: University of California Press.

Savignon, S. (1992). This is only a test. In E.W. Shohamy & A. Ronald (Eds.), *Language assessment for feedback: Testing and other strategies* (pp. 53-71). Dubuque, IA: Kendall/Hunt.

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

Sharpe, P.J. (1999). *How to prepare for the TOEFL test* (9th ed.). Princeton, NJ: Educational Testing Services.

Stryker, S.B. (1997). The Mexico experiment at the Foreign Service Institute. In S.B. Stryker & B.L. Leaver (Eds.), *Content-based instruction in foreign language education: Models and methods* (pp. 176-202). Washington, DC: Georgetown University Press.

Sutherland, T.E., & Bonwell, C.C. (Eds.). (1996). *Using active learning in college classes: A range of options for faculty.* San Francisco, CA: Jossey-Bass.

Wesche, M.B. (1993). Discipline-based approaches to language study: Research issues and outcomes. In M.R. Krueger & F. Ryan (Eds.), *Language and content. Discipline- and content-based approaches to language study* (pp. 57-79). Toronto, ON: D.C. Heath.

White, E. (1985). *Teaching and assessing writing.* San Francisco, CA: Jossey-Bass.

## *Appendix A: The Writing Prompt:*

*The Role of Japanese Men in the Home*

> What will the roles of Japanese men and women in the home be ten years from now? Write an essay using the following reading and your own ideas. Be sure to give reasons for your answer.

Japanese husbands do housework only eight minutes per day, while Japanese working women do housework about three hours each day. In fact, Japanese men do less housework than men do in most other countries.

Why do Japanese men do less housework than men in other countries? First, Japanese men have less time to help with housework. They work longer hours than European or North American men. Second, Japanese men do not spend much leisure time at home. They relax at their favorite bars or restaurants. Third, Japanese men did not learn from their mothers how to do housework. Japanese men want their wives to do the housework like their mothers did. Therefore, for many Japanese men, the home is the wife's responsibility. In short, Japanese men feel that they can't do housework; they feel that they shouldn't do housework.

What do young Japanese women think about marriage nowadays? Many believe that there is no freedom or advantage to marriage. Marriage means a lot more work for them, especially if they have children. Today nearly 40% of Japanese women are unmarried at the age of 29, and the divorce rate is four times what it was in the 1950s. In addition, fewer Japanese children are being born, and the population of Japan is actually decreasing. Some men have responded by taking classes in how to be "better" husbands, but obviously many Japanese women don't see much change. What do you think the Japanese household will be like in the future?
(approximately grade 8 level based on the Flesch-Kincaid readability index)

## *Appendix B*

1997 Writing Sample (WS) and College-Wide (C-W) Groups by Year

| | *WS Totals* | *C-W Totals* | *WS % of C-W* |
|---|---|---|---|
| First Year | 75 (34%) | 82 (31%) | 91% |
| Second Year | 71 (32%) | 86 (33%) | 83% |
| Third Year | 44 (20%) | 52 (20%) | 85% |
| Fourth Year | 30 (14%) | 44 (16%) | 68% |
| | 220 | 264 | 83% |

## The Benchmarks Rating Scale for Writing

| | Benchmark 1 | Benchmark 2 | Benchmark 3 | Benchmark 4 | Benchmark 5 | Benchmark 6 |
|---|---|---|---|---|---|---|
| **Content, Organization and Development** | • *Initial competence - simple texts*<br>• Text is very short<br>• A few simple sentences or phrases about self | • *Developing competence - simple texts*<br>• Few sentences or phrases in form of a simple description<br>• Relies heavily on simple restatement -- context immediate and personal | • *Adequate competence - simple texts*<br>• A number of one clause sentences about topic in form of simple description or narration<br>• Mostly memorized words and phrases and some additional material | • *Fluent competence - simple texts*<br>• Content of text controlled<br>• Clear statement of topic, actors and circumstances<br>• Conveys ideas within predictable contexts | • *Initial competence - medium complexity*<br>• Produces coherent text, beginning, middle and end with appropriate sequencing of events<br>• Does better with controlled writing contexts -- not creating text freely | • *Developing competence - medium complexity*<br>• Coherent text with beginning, middle and end<br>• Proper development and linking of ideas |
| **Vocabulary and Structure** | | • Accuracy in structure and spelling inconsistent | • Basic tenses and structures with correct capitalization and punctuation | • Mostly one-clause sentences or coordinated clauses, "enriched" with modifiers (adjectives and adverbs), some linked with "and" and "but"<br>• Basic simple tenses, simple time signals (now, after, before) | • Linguistic means of expression remain simple: compound sentences, present tense, some present perfect<br>• Frequent errors in accuracy, awkward sounding phrases, and L1 borrowings | • Linguistic means of expression relatively simple: simple compound and few complex sentences (time clauses, because), present perfect, modals<br>• Sufficient vocabulary within topic |
| | **Benchmark 7** | **Benchmark 8** | **Benchmark 9** | **Benchmark 10** | **Benchmark 11** | **Benchmark 12** |
| **Content, Organization and Development** | • *Adequate competence - medium complexity*<br>• Coherent, well-developed paragraph on familiar and relevant topic<br>• Conveys main ideas clearly and coherently within one paragraph form but overall discourse may be awkward in textual cohesion<br>• Expresses sequence of events with detail and precision | • *Fluent competence - medium complexity*<br>• Coherent, well-developed paragraph on familiar and relevant topic: can join 3 or 4 paragraphs into a larger text<br>• Clearly and coherently conveys main ideas, but discourse structure may sometimes seem awkward<br>• Presents causal relationships; expresses personal opinion | • *Initial competence - complex texts*<br>• Descriptive writing is coherent, detailed, comprehensive; writer also familiar with other discourse types (narrative, expository, argumentative)<br>• Conveys information clearly and precisely<br>• Communicates effectively and independently in unpredictable contexts<br>• Follows conventions in organization of ideas | • *Developing competence - complex texts*<br>• Text length is appropriate to purpose and format<br>• Effective, coherent, stylistically fluent/ complex and interesting text<br>• Incorporates a variety of rhetorical structures<br>• Purpose, main points, coherence of text is clear to the reader | • *Adequate competence - complex texts*<br>• Uses full range of complex sentences and discourse structures and appropriate register Effective, coherent, stylistically complex and sizable expository text<br>• Writes effectively and independently in various contexts<br>• Can present ideas synthesized from a variety of sources<br>• Uses appropriate expository devices with precision and flexibility | • *Fluent competence - complex texts*<br>• Effective, coherent, stylistically complex and sizable argumentative text<br>• Uses argumentative patterns, including thesis statement, with precision and flexibility<br>• Fluently and effectively writes texts for various purposes<br>• Clearly and effectively conveys ideas, intent and tone<br>• Content and audience determine complexity, formality and length of text |
| **Vocabulary and Structure** | • Complex structures: comparison/contrast, adjectival clauses, agents-event relations<br>• Contains errors and stylistically rigid use of structures | • Structures are complex reflecting agent-event relations and logical relations of contrast, time, cause and reason<br>• Writing, although clear in meaning to reader, contains accuracy and lexical collocation errors, and lacks flexibility in style | • Wide range of complex grammatical structures and vocabulary<br>• Reasonably good fluency, accuracy or both -- often able to self-correct | • Uses appropriate structures with precision and flexibility.<br>• Wide range of complex grammatical structures and vocabulary<br>• Grammatical accuracy high -- occasional grammatical/ lexical errors, but high degree of self-correction | • Uses full range of complex sentences and discourse structures and appropriate register<br>• Accuracy is consistent with correct tenses, modals, appropriate vocabulary<br>• Errors are occasional and minimal | • Excellent fluency, accuracy and style<br>• Fluent command of language structures, vocabulary, idiom and discourse structures<br>• Minimal grammatical and lexical errors may occur<br>• Edits texts effectively |

Adapted from *Canadian Language Benchmarks*, Working Document, 1996

December 1998

## The Benchmarks Rating Scale for Writing

| | *BASIC* | | *INTERMEDIATE* | | *ADVANCED* | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** |
| | **Benchmark 1 ~ 2** | **Benchmark 3 ~ 4** | **Benchmark 5 ~ 6** | **Benchmark 7 ~ 8** | **Benchmark 9 ~ 10** | **Benchmark 11 ~ 12** |
| **Content, Organization and Development**<br>**The text is/ has/ shows ...** → | • Incoherence, consists of a few simple sentences or phrases<br>• Mostly memorized words and phrases copied from prompt<br>• Unclear meaning | • A number of simple sentences (mostly one clause) about prompt<br>• Evidence of control of text but limited; coherent paragraph development<br>• Identification of topic, actors and circumstances<br>• Some expression of personal opinion | • Coherence with well developed paragraph form to respond to prompt<br>• Clear main ideas, but overall discourse sometimes awkward in cohesion<br>• Logical development<br>• Expression of personal opinion | • Well-developed paragraphs responding to prompt: paragraphs joined coherently<br>• Clear and coherent main ideas, but discourse structure sometimes awkward<br>• Causal relationships<br>• Fluent expression of personal opinion | • Effective response to prompt – coherent, stylistically fluent/ complex and interesting<br>• Clear and precise information within appropriate structure<br>• Clear purpose, main points, coherence<br>• Variety in rhetorical structures | • Effective, coherent, stylistically complex and sizable expository/ argumentative discourse<br>• Clear and effective expression of ideas, intent and tone<br>• Argumentative patterns, including thesis statement, used with precision and flexibility<br>• Appropriate expository devices |
| **Vocabulary and Mechanics** | • Errors that severely restrict understanding | • Frequent errors that restrict understanding<br>• Mostly one-clause sentences or coordinated clauses, "enriched" with modifiers (adjectives and adverbs), some linked with "and" and "but"<br>• Basic simple tenses, simple time signals (now, after, before) | • Errors that interfere with understanding; awkward sounding phrases, and first-language borrowings<br>• Relatively simple linguistic means of expression: simple compound and few complex sentences, present perfect, modals<br>• Sufficient vocabulary to address prompt | • Some errors but they do not interfere with understanding<br>• Complex structures reflecting agent-event relations and logical relations of purpose, contrast, time, cause and reason<br>• Stylistically rigid use of structures | • High grammatical accuracy -- occasional grammatical/ lexical errors, but often self-corrected<br>• Wide range of complex grammatical structures and vocabulary<br>• Appropriate structures used | • Minimal grammatical and lexical errors<br>• Consistent accuracy with correct tenses, modals, appropriate vocabulary<br>• Effective editing<br>• Full range of complex sentences and discourse structures and appropriate register<br>• Fluent command of language structures, vocabulary, idiom and discourse structures |

Adapted from *Canadian Language Benchmarks*, Working Document, 1996

June 2000